LUMILENS: CHOREOGRAPHING SPATIALLY-VARYING LIGHT AND LENS FOR GENERATIVE MONOCULAR VIDEO RE-RENDERING

Anonymous authors
Paper under double-blind review

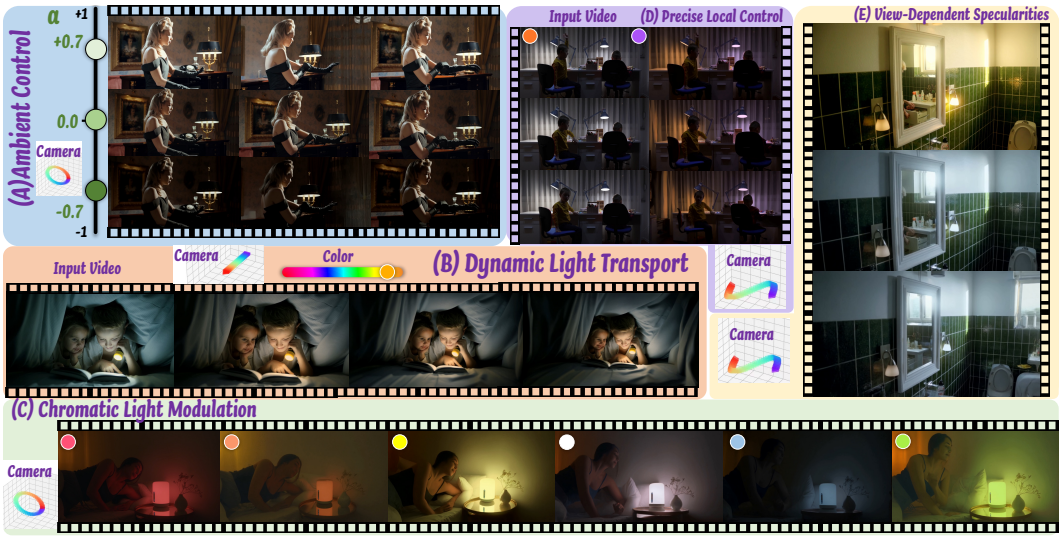


Figure 1: ***LumiLens*** enables physically-plausible monocular video re-rendering with precise, disentangled illumination and camera control: (A) **Ambient Control**, fine-grained global lighting bias; (B) **Dynamic Light Transport**, deformable moving-light shadows faithfully conforming to scene geometry; (C) **Chromatic Light Modulation**, modifying emissive-source spectra; (D) **Spatially-Varying Local Control**, independently manipulating near-field lights with geometric awareness; and (E) **View-Dependent Specularities**, dynamic reflections under varying camera poses.

ABSTRACT

We present ***LumiLens***, the first generative framework to internalize the dynamic interplay of geometry, intrinsic attributes, and lighting for monocular video re-rendering with simultaneous control of viewpoint and fine-grained, parametric spatially-varying local illumination. Two observations motivate this framework: (1) unlike global relighting, which essentially acts as a uniform 2D filter, local relighting is a high-frequency near-field 3D rendering problem with pixel-specific light vectors, requiring geometry-aware spatial reasoning; (2) radiance transport inherently couples surface appearance to 3D geometry, lighting, and viewing direction, making precise camera control necessary for reflections that faithfully respond to perspective shifts. Existing methods fail to achieve this joint control or exhibit static radiance artifacts, including structural stretching and original highlights unresponsive to viewpoint changes. Our Physically-Motivated Dual-Stream Conditioning Mechanism decouples spatial illumination modeling from intrinsic scene semantics. The Spatial Stream explicitly encodes the spatial illumination distribution of light-geometry interactions through warped depth and parametric light proxies; the Global Stream injects global lighting conditions and anchors inherent scene structure and dynamic semantics through the source video. To address scarce paired training data, our scalable Unreal Engine 5-based Omni-View Illumination (OVI) pipeline synthesizes over 134K video pairs with physically decoupled illumination streams. Extensive experiments demonstrate LumiLens's physically plausible, spatially-varying local illumination under novel views with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

preserved intrinsic scene integrity, and its precise joint and flexible independent lighting and viewpoint control across both synthetic and real-world scenarios.

1 INTRODUCTION

Video Diffusion Models (VDMs) excel at high-quality video synthesis (Guo et al., 2024; Yang et al., 2025; Wan et al., 2025) but lack unified control over camera movement and lighting design, two fundamental elements of cinematic storytelling. True visual immersion requires revisiting scenes from novel viewpoints while dynamically adjusting spatially-varying local illumination, yet such physically-grounded joint control from monocular video remains an outstanding challenge (Liu et al., 2025a; He et al., 2025). Existing research treats these controls separately: video relighting methods (Zhou et al., 2025; Fang et al., 2025a; Liu et al., 2025c) modify ambient mood through global, 2D pixel-wise modulation rather than local illumination control, failing to capture high-frequency 3D interactions such as cast shadows or distance falloff; camera-controlled generation (Hu et al., 2025a; Ren et al., 2025; YU et al., 2025; Bai et al., 2025) typically learns camera-RGB distributions without explicitly decoupling lighting, leading to the **"Baked-In Reflection Trap"** (Fig. 2): original lighting stays fixed to surface textures, yielding unrealistic novel-view appearances.

We argue that joint illumination and viewpoint modeling and control is a physical prerequisite in a dynamic world shaped by geometry, motion, and illumination. According to the principles of radiance transport, the light emitted from a surface toward an observer is a coupled function of incident light, geometry, and viewing direction. Generative frameworks therefore need to internalize these interactions in latent space for physically-plausible re-rendering. Under a fixed camera, a network may bypass physical constraints by "painting" lighting onto textures; only consistency across a moving camera trajectory compels it to capture spatially-varying light transport. This transition to geometry-grounded synthesis is essential for disentangling dynamic radiance from intrinsic scene identity and bypassing view-invariant lighting entanglement and texture memorization. Explicit awareness of lighting–scene–viewpoint interactions enables physically-plausible re-rendering and joint control, naturally eliminating artifacts such as the "Baked-In Reflection Trap".

To achieve this, *LumiLens* grounds generative video re-rendering in explicit geometry and parametric illumination, replacing pixel-space manipulation with structure-guided radiance synthesis. This offers two advantages: 1) Structural Robustness: warped depth provides a noise-resilient structural anchor that maintains geometric coherence across lighting and viewpoint changes, unlike complex, view-dependent RGB textures that distort under projection. 2) Physical Disentanglement: explicit radiometric priors bypass the "Baked-In Reflection Trap", making shadows, dynamic specularities, and multi-bounce global illumination respond to the target physical state rather than residual source textures. Building on this foundation, LumiLens proposes a Physically-Motivated Dual-Stream Conditioning Mechanism. The Spatial Stream combines explicit geometric surrogates, including depth and occlusion masks, with target light position, color, and intensity under new camera trajectories to capture spatially-varying interactions between local illumination and geometry across dynamic viewpoints. The Global Stream uses the source video as a semantic anchor to faithfully preserve scene structure, semantics, dynamic motion, and intrinsic properties. Our scalable Unreal Engine 5-based (Epic Games, 2022) *Omni-View Illumination (OVI) pipeline* automatically generates **134K physically-disentangled video pairs** through fine-grained parametric lighting composition. This supervision enables implicit approximation of complex, non-linear radiometric dependencies, including shadow contrast evolution rather than global hue shifts, supporting high-fidelity, physically-plausible synthesis of intricate view-dependent light transport.

Our contributions are: 1) *LumiLens*, the first generative framework for joint control of spatially-varying local illumination and camera trajectories through physically conditioned re-rendering, achieving seamless light–lens choreography with high-fidelity radiance consistent with complex light transport; 2) a novel Physically-Motivated Dual-Stream Conditioning Architecture combining structural and radiometric priors with semantic anchoring to internalize radiometric interactions and synthesize authentic view-dependent radiance rather than propagate fixed-lighting textures; 3) the scalable OVI pipeline, generating over 134K physically disentangled video pairs to supervise the interplay of dynamic viewpoints and spatially-varying light transport.

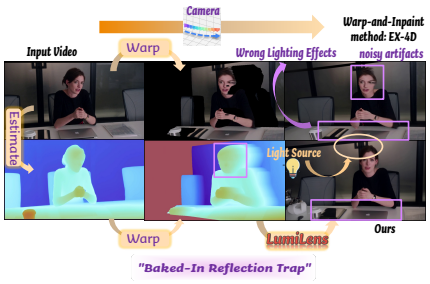


Figure 2: Limitations of warp-and-inpaint methods. RGB warping erroneously fixes lighting to surface textures and causes unrecoverable structural stretching from unstable depth, even with SOTA estimators (Lin et al., 2025a; Hu et al., 2025b).

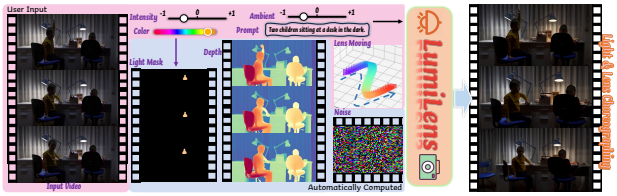


Figure 3: Overview of the *LumiLens* Framework. Given an input video, users provide parameterized controls (in pink backgrounds) for local illumination (intensity, color, ambient) and a target camera trajectory. To spatially ground these controls, we automatically extract (in blue backgrounds) a light mask using a segmentation model guided by user intent (text prompts or first-frame clicks) and estimate scene depth via a monocular depth estimator. These spatial maps and the parametric inputs condition *LumiLens* to synthesize the final video with jointly choreographed light and lens.

2 Related Work

2.1 Learning-Based Illumination Control

Illumination control manipulates lighting in real-world images or videos. Early studies used physical illumination models (Barron & Malik, 2015) or deep networks with explicit lighting representations (Bammey, 2022; Sun et al., 2019), especially for portraits (Shu et al., 2017; Shih et al., 2014; Sengupta et al., 2018). Diffusion models have improved relighting fidelity for global (Cha et al., 2025; Kim et al., 2024; He et al., 2025; Chadebec et al., 2025; Careaga & Aksoy, 2025) and local control (Xing et al., 2025; Choi et al., 2025). IC-Light (Zhang et al., 2025b) uses light-transport consistency loss and large-scale datasets, while LightLab (Magar et al., 2025) enables local image relighting with explicit intensity and color control. Recent works (Zhou et al., 2025; Fang et al., 2025a; Zeng et al., 2025; Lin et al., 2025b) extend relighting to videos: Light-A-Video uses training-free cross-frame light attention and progressive fusion, while RelightVid extends IC-Light's 2D U-Net with temporal attention. These methods still struggle to balance lighting fidelity and temporal consistency. RelightMaster (Bian et al., 2025b) encodes point-light position, color, and intensity through Multi-plane Light Images; LiveLight (Ma et al., 2026) adopts this representation for real-time streaming with depth and normal constraints. Their lighting parameterization does not explicitly encode fixture geometry or fixture-specific emission patterns. In contrast, *LumiLens* binds control to a selected physical emitter and combines scene geometry and appearance cues to model its view-dependent, spatially-varying illumination.

2.2 Video Generation and Camera Control

Video diffusion models evolved from 3D U-Nets (Ho et al., 2022) to latent diffusion (Blattmann et al., 2023; Guo et al., 2024) and Transformers (Peebles & Xie, 2023; Wan et al., 2025; Yang et al., 2025). Camera-controlled methods (Yang et al., 2024; Bahmani et al., 2025; Zheng et al., 2024; 2025; Wang et al., 2024; He et al., 2024; Bai et al., 2025) implicitly condition diffusion models on camera parameters to learn 3D consistency but suffer from substantial camera control errors. RayPE (Yin et al., 2026) further introduces ray-based positional encoding into attention. Multi-view video diffusion (Wu et al., 2025; Bian et al., 2025a; Liu et al., 2025b) targets cross-view consistency. Explicit depth or tracking cues (Zhang et al., 2025a; YU et al., 2025; Ren et al., 2025; Hu et al., 2025a) enable more precise camera control through warp-and-inpaint but introduce unrecoverable distortions, which WorldForge (Song et al., 2025) mitigates without training. DepthDirector (Chen et al., 2026) and Wonder (Xu et al., 2026) use geometric signals solely for camera encoding, and Lyra 2.0 (Shen et al., 2026) uses warped coordinates for spatial memory. Light-X (Liu et al., 2025a) adds only global relighting and remains prone to the "Baked-In Reflection Trap". In contrast, *LumiLens* leverages warped depth as a geometric proxy to supply the surface orientation, layout, and viewing direction that local relighting depends on, while also supporting viewpoint control.

3 METHOD

3.1 PRELIMINARY AND OVERVIEW

Figure 3 shows *LumiLens* mapping source video V_S to target sequence V_T through generative re-rendering conditioned on $\Phi = (\mathcal{P}, \mathcal{L})$, where $\mathcal{P} = \{P_t\}_{t=1}^{T}$ is the target camera trajectory. For fine-grained parametric illumination control, $\mathcal{L}$ combines a user-specific light mask M_{tgt} extracted with SAM 3 (Carion et al., 2025) and user-defined color $\mathbf{C}$, intensity γ, and global ambient bias α. Joint conditioning on Φ faithfully choreographs light and lens movement through view-dependent radiance consistent with the scene's coupled physical evolution.

Radiance Transport and Surface Appearance. Our framework is grounded in light-surface interaction: the Rendering Equation defines the outgoing radiance L_o at a surface point through incident light, surface geometry, and material properties

$$L_o(\mathbf{q}, \omega_o) \approx \sum_{m \in \mathcal{L}} f_r(\mathbf{q}, \omega_m, \omega_o) L_m(\omega_m \cdot \mathbf{n}), \tag{1}$$

where $\mathbf{q}$ denotes the surface point, ω_o is the viewing direction, ω_m and L_m represent the incident direction and radiance of the m-th light source in the set $\mathcal{L}$, $\mathbf{n}$ is the local surface normal, and f_r denotes the Bidirectional Reflectance Distribution Function (BRDF).

This formulation motivates our Physically-Motivated Dual-Stream Conditioning Architecture, which moves beyond simplistic 2D modulation by internalizing radiometric interactions in a spatially-consistent latent distribution. The pipeline (Fig. 4) replaces error-prone RGB warping with explicit geometric proxies (Sec. 3.2) for robust structural grounding. The *Spatial Stream* (Sec. 3.3) combines warped depth and light proxies representing surface orientation $\mathbf{n}$, incident direction ω_m, and dynamic viewing direction ω_o with parametric lighting to synthesize spatially-varying transport effects, including cast shadows and surface shading. The *Global Stream* (Sec. 3.4) anchors material properties (f_r) and high-level semantics to source-video features, faithfully preserving scene structure, motion dynamics, and subject identity. Through DiT attention, these disentangled cues implicitly approximate radiance transport without explicit radiometric calculation, producing high-fidelity, physically-plausible radiance that follows the target trajectory and coupled light–lens evolution.

3.2 EXPLICIT GEOMETRY-BASED SCENE REPRESENTATION

For an explicit, consistent 3D foundation for physically plausible relighting, we estimate temporally consistent relative depths $D_S = \{d_t\}_{t=1}^{S}$ from video $V_S = \{I_t\}_{t=1}^{S}$ using DepthCrafter (Hu et al., 2025b). A 3D foundation model (Lin et al., 2025a) estimates camera extrinsics $P_S = \{p_t\}_{t=1}^{S}$, intrinsics K, and multi-view consistent depths $F_S = \{f_t\}_{t=1}^{S}$ for 3D anchoring.

While D_S provides high-resolution local details essential for capturing fine-grained surface geometry, F_S ensures alignment with the estimated camera poses. The two depth representations are aligned by solving for the optimal scale s and shift b in inverse depth space:

$$s, b = \arg\min_{s,b} \sum_{t=1}^{S} \left\| \frac{1}{f_t} - \left(\frac{s}{d_t} + b \right) \right\|_2^2. \tag{2}$$

Following alignment, we transform the observations into an explicit dynamic 3D mesh. Specifically, each pixel (i, j) in I_t with aligned depth $d_{t,i,j}$ is back-projected into a 3D vertex $\mathbf{X}_{i,j,t} = o_t + d_{i,j,t} \cdot r_{i,j,t}$, where o_t and $r_{i,j,t}$ denote the camera origin and ray direction at frame t, derived from the estimated camera extrinsic p_t and intrinsics K. Adjacent 2×2 pixel blocks are connected into two triangular faces following EX-4D (Hu et al., 2025a), forming a continuous surface. Degenerate face detection and temporally consistent depth priors from D_S keep the mesh topology stable despite object motion. Unlike implicit representations or sparse point clouds, this consistent topology is required to faithfully resolve occlusion boundaries and fill holes in novel views.

For an arbitrary, user-defined target camera trajectory $P_T = \{p_t\}_{t=1}^{T}$, we render view-consistent depths $D^r = \{d_t^r\}_{t=1}^{T}$ and occlusion masks $M^r = \{m_t^r\}_{t=1}^{T}$ from the dynamic mesh. To accommodate the high dynamic range and adapt depth to the video backbone, D^r is normalized to $[0, 1]$ in logarithmic space and mapped to RGB using a predefined colormap (e.g., Turbo). Together, (D^r, M^r) provide rich, spatially-aligned geometric context for the Spatial Stream.

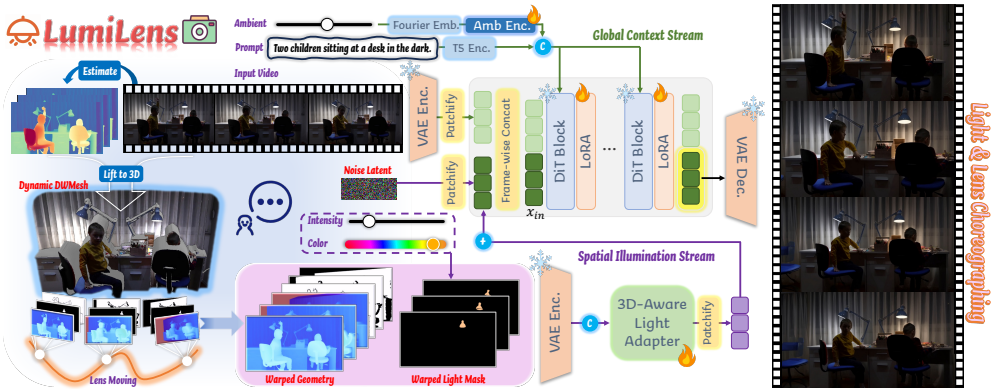


Figure 4: Physically-Aware Dual-Stream Conditioning Mechanism. The **Global Context Stream** anchors semantics and ambient control; the **Spatial Illumination Stream** enforces geometric and physical constraints. Parallel processing and injection into DiT enable joint camera–scene–illumination interaction while maintaining spatiotemporal coherence during re-rendering.

3.3 Spatially-Varying Local Illumination Conditioning

Motivated by the radiance transport variables (e.g. L_m, ω_m, n) introduced in Sec. 3.1, spatial feature proxies are explicitly constructed to guide the generation.

Physically-Motivated Condition Construction. The warped depth D^r (from Sec. 3.2) is utilized as a geometric proxy, which inherently encodes the local surface normal (n), spatial layout, and target viewing direction (ω_o). Simultaneously, the occlusion masks M^r enforce precise occlusion boundaries, ensuring that lighting effects respect the topological depth discontinuities. For target illumination, M_{tgt} is back-projected to define a virtual geometric light primitive at $L_{\text{pos}} \in \mathbb{R}^3$ in the reconstructed scene, rather than a static 2D region.

For each frame t along the target trajectory P_T, this primitive is rendered onto the image plane to generate a dynamic binary mask $M^r_{\text{tgt},t}$, which strictly adheres to motion parallax. Notably, $M^r_{\text{tgt},t}$ strictly delineates the physical light emitter and explicitly excludes secondary effects like mirror reflections, compelling the network to infer rather than simply copy complex light transport. The mask is then modulated by the user-defined color $\mathbf{C} \in \mathbb{R}^3$ and intensity scalar γ:

$$M^r_C = \{M^r_{\text{tgt},t} \odot \mathbf{C}\}_{t=1}^T, \quad M^r_\gamma = \{M^r_{\text{tgt},t} \cdot \gamma\}_{t=1}^T, \tag{3}$$

where $\odot$ denotes element-wise multiplication. By concatenating these components $(D^r, M^r, M^r_C, M^r_\gamma)$, the network is provided with a complete set of spatially-varying illumination contexts. This formulation supports simultaneous injection of multiple light proxies with independently parameterized positions, intensities, and chromaticities, enabling fine-grained control of complex multi-source illumination in a single inference pass.

Without explicit radiometric equations, these dynamically aligned geometric and illumination conditions enable approximation of distance falloff, cast shadows, and view-dependent specularities.

Adapter Architecture and Injection. We use the three-dimensional Variational Auto-Encoder (VAE) $\mathcal{E}$ (Kingma & Welling, 2014) for latent space mapping. A lightweight 3D-Aware Light Adapter $\mathcal{A}$ employing 3D convolutions processes these priors into a latent condition $\mathbf{c}_{\text{spatial}}$:

$$\mathbf{c}_{\text{spatial}} = \mathcal{A}\left(\text{Concat}\left[\mathcal{E}(D^r), \mathcal{E}(M^r), \mathcal{E}(M^r_C), \mathcal{E}(M^r_\gamma)\right]\right). \tag{4}$$

Pixel-wise aligned with the target video, $\mathbf{c}_{\text{spatial}}$ is injected directly into transformer blocks by element-wise addition after a zero-initialized convolution layer applied to the flow state $\boldsymbol{x}_t$:

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t + \mathbf{c}_{\text{spatial}}. \tag{5}$$

This additive injection aligns geometry with target illumination, yielding physically-plausible shadows and shading responsive to light location and intensity rather than image-space painting (Fig. 2).

3.4 GLOBAL CONTEXT CONDITIONING

Source-Guided Semantic and Dynamic Anchor. While the Spatial Stream captures spatially-varying light-geometry interactions, it lacks the semantic context needed to reconstruct scene appearance. The *Global Stream* therefore uses the input video V_S as a holistic semantic and dynamic anchor, preserving source identity, structural details, and temporal dynamics that geometry alone cannot fully provide, without explicitly decomposing albedo or motion vectors. For relighting, the final DiT input (Fig. 4) concatenates source and target states frame-wise:

$$\boldsymbol{x}_{in} = [\texttt{patchify}(\boldsymbol{x}_s), \texttt{patchify}(\hat{\boldsymbol{x}}_t)]_{\text{frame-dim}}. \tag{6}$$

Self-attention uses the input-video encoding $\boldsymbol{x}_s$ as a conditional reference rather than directly propagating source pixels. The target state $\hat{\boldsymbol{x}}_t$, carrying the Spatial Stream's 3D-aware illumination constraints (Eq. 5), queries the unperturbed source features, naturally aligning retrieved dynamic semantic priors and identity cues with the target lighting geometry. This enables new, view-dependent radiance while preserving source identity and dynamic fidelity, unconstrained by source lighting.

Global Ambient Intensity Control. A user-defined scalar α regulates overall ambient radiance. Fourier encoding (Tancik et al., 2020) maps α to high-frequency embeddings, which an Ambient Injector MLP processes with zero-initialized weights to preserve pretrained generative priors. Its output is channel-wise concatenated with text embeddings c_{text} to form $\hat{c}_{\text{text}}$, which enters DiT blocks through cross-attention to coherently guide scene brightness across frames.

This decoupling frees the Spatial Stream from memorizing textures with baked-in illumination, specializing it in geometry and lighting, while the Global Stream anchors structure and semantics to faithfully preserve identity and motion.

3.5 TRAINING STRATEGY

We freeze the pretrained backbone and adapt it using Low-Rank Adaptation (LoRA), while jointly training the 3D-Aware Light Adapter and the Ambient Injector. This enables the model to implicitly approximate complex radiance transport while preserving its generative priors.

Objective Function. Following the Rectified Flow framework (Lipman et al., 2023), we train the model to predict the velocity field $\boldsymbol{v}$ transporting the initial noise distribution to the data distribution:

$$\mathcal{L} = \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{z}}\left[\left\|\boldsymbol{v}_\theta(\boldsymbol{x}_t, \underbrace{D^r, M^r, M^r_{\text{tgt}}, \mathbf{C}, \gamma}_{\text{Spatial Illumination}}, \underbrace{V_S, \alpha, c_{\text{text}}}_{\text{Global Context}}, t) - (\boldsymbol{z} - \boldsymbol{x})\right\|_2^2\right], \tag{7}$$

where $\boldsymbol{x}_t = (1-t)\boldsymbol{x} + t\boldsymbol{z}$ is the interpolated flow state between the ground-truth video $\boldsymbol{x}$ and Gaussian noise $\boldsymbol{z}$. Inside $\boldsymbol{v}_\theta$, these inputs form the transformer input $\boldsymbol{x}_{in}$ (Eq. 6), fusing the source reference V_S with the spatially-modulated flow state $\hat{\boldsymbol{x}}_t$. The network $\boldsymbol{v}_\theta$ learns to align the flow trajectory with spatial geometric and illumination constraints and the global spatio-temporal reference, ensuring generated samples satisfy the target lighting conditions.

3.6 OMNI-VIEW ILLUMINATION (OVI) DATASET

Acquiring real-world videos with paired multi-view trajectories, precise ground-truth geometry, and decoupled lighting control is prohibitively expensive. We therefore construct OVI using Unreal Engine 5 (UE5) (Epic Games, 2022) and Monte-Carlo path tracing to generate photo-realistic, physically disentangled video streams (Fig. 21). Inspired by LightLab (Magar et al., 2025), we separately render an ambient stream V_{amb} with the target light off and a target-light stream V_{light} with a canonical high-intensity white light and all other lights off. Offline, we combine these linear RGB streams into ground-truth videos with varied light colors, intensities, and ambient levels:

$$V_{relit}(\hat{\alpha}, \hat{\gamma}, \mathbf{C}) = \hat{\alpha} V_{amb} + \hat{\gamma}(V_{light} \odot \mathbf{C}), \tag{8}$$

where $\mathbf{C}$ specifies the target light color, and $\hat{\alpha}$ and $\hat{\gamma}$ are compositing coefficients scaled relative to the canonical rendered intensity. This parameterization provides training pairs with precise intensity and color differences, supervising large-scale illumination variations from switching lights on/off (Fig. 5) to complex chromatic transitions such as neon color changes (Fig. 6). Activating the local



Figure 5: Light On Results. *LumiLens* synthesizes realistic "lights-on" effects by robustly inferring target illumination's specific spatial form and volumetric influence despite ambiguous ambient-lit inputs.



Figure 6: Diverse Chromatic Variations. Given the unlit input video (top), this figure demonstrates precise control over light spectral properties.

light source from an "off" state, *LumiLens* synthesizes diverse, vivid and highly saturated chromaticities, faithfully rendering the resulting physically plausible specular reflections and complex global illumination (e.g., color bleeding) on the subject and background.

Randomizing $\mathbf{C}, \hat{\gamma}, \hat{\alpha}$ yields eight variants per sequence, expanding 16,800 base sequences to approximately **134,400 training samples** (49 frames at 10 FPS). Input and output videos use display-referred sRGB; linear compositing is confined to offline data preparation, followed by tone mapping and sRGB encoding (Appendix B.2).

4 Experimental Results

4.1 Settings of Experiments

Baselines and Evaluation Metrics. To the best of our knowledge, *LumiLens* is the first method for fine-grained joint control of spatially-varying local illumination and viewpoint in monocular videos. We evaluate its decoupled learning and joint control in three tasks restricted to light color change (including light *on*/*off*) and camera control, adapting state-of-the-art baselines as follows:

Task 1: Joint Control of Spatially-Varying Illumination and Viewpoint. We compare against Light-X (Liu et al., 2025a) and V-RGBX (Fang et al., 2025b), with EX-4D (Hu et al., 2025a) providing camera control for V-RGBX. NanoBanana Pro (Google DeepMind, 2026) generates reference images for the Nano-Ref configurations of these baselines. On in-the-wild videos, VBench (Huang et al., 2024) evaluates overall perceptual quality under joint lighting and camera control.

Task 2: Fine-grained Parametric Light Control. We compare against RelightVid (Fang et al., 2025a) and frame-by-frame IC-Light (Zhang et al., 2025b) to evaluate temporal consistency and local light interactions. On in-the-wild videos, we measure temporal consistency with CLIP-S (Radford et al., 2021), motion preservation with RAFT-based optical-flow differences (Teed & Deng, 2020), geometric preservation with Chamfer Distance (CD), and intrinsic albedo consistency with IAC.

Task 3: Camera Control and Reflection Consistency. We compare against ReCamMaster (Bai et al., 2025) and EX-4D (Hu et al., 2025a) to evaluate camera control and view-dependent appearance. We report CLIP-V for semantic fidelity, RotErr, TransErr, and CamMC for camera accuracy (Zheng et al., 2024), and GIM-based matching pixels (Mat. Pix.) for view consistency (Shen et al., 2024).

Table 1: Quantitative results on Synthetic Validation Dataset. Ours (A)/(B): ablation Models A/B.

	Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓	CD↓
Task1	Light-X	13.892	0.442	0.472	679.59	-
	V-RGBX+EX-4D	13.335	0.385	0.501	701.82	-
	Ours (A)	17.361	0.579	0.342	472.33	-
	Ours (B)	16.126	0.561	0.369	483.90	-
	Ours	**21.761**	**0.877**	**0.216**	**370.47**	-
Task2	RelightVid	15.274	0.471	0.454	331.63	0.583
	IC-Light	12.738	0.394	0.518	380.91	0.659
	Ours	**22.410**	**0.882**	**0.209**	**236.22**	**0.198**
Task3	ReCamMaster	18.454	0.730	0.202	539.62	-
	EX-4D	19.371	0.724	0.191	513.29	-
	Ours	**23.139**	**0.898**	**0.179**	**401.23**	-

Metrics on Synthetic Data. With ground truth available, all three tasks are evaluated using PSNR and SSIM for pixel-level accuracy, LPIPS for perceptual similarity, and FVD (Unterthiner et al., 2019) for video distribution quality. For relighting, CD between point clouds reconstructed with MegaSAM (Li et al., 2025) additionally measures geometric preservation. Detailed baseline configurations and metric definitions are provided in Appendices C.2 and C.3, respectively.

Evaluation dataset. Our **Synthetic Held-out Set**, generated as in Sec. 3.6, contains four unseen scenes with 20 distinct light sources; each light has two random camera trajectories and two light color states for rigorous quantitative benchmarking. Our **In-the-Wild Benchmark** assesses robustness and open-domain generalization using 100 high-quality videos from Pexels (Pexels, 2025), KoalaDataset (Wang et al., 2025), and AI-generated Sora clips (Brooks et al., 2024). Its diverse dynamic scenes involving humans, animals, and complex object interactions provide a challenging qualitative testbed. Training and inference details are provided in Appendix C.1.

4.2 Comparative Analysis

Quantitative Evaluation on Synthetic Data. LumiLens outperforms all baselines across all three tasks (Table 1). Its highest PSNR/SSIM and lowest LPIPS show superior high-frequency reconstruction; Light-X and RelightVid produce blur or inconsistent lighting with higher perceptual errors. Its significantly lower FVD validates the Global Stream's anchoring of scene dynamics against flickering in frame-wise editors such as IC-Light. Its lowest CD reflects 3D structure preservation through the Spatial Stream's geometric constraints, whereas video inpainting distorts structure.

Perceptual and Camera-Control Evaluation. Tables 2 and 5 evaluate in-the-wild videos; Table 6 combines real and synthetic camera-control results.

Task 1: LumiLens surpasses Light-X and V-RGBX on **VBench** (Huang et al., 2024) (Table 2), maintaining high aesthetic quality and temporal coherence that validate joint conditioning. Light-X often produces "sliding" shadows as it struggles to coordinate camera movement with lighting changes.

Task 2: LumiLens achieves the highest CLIP-S and lowest motion-preservation error, IAC, and CD (Table 5), demonstrating temporally consistent relighting that preserves motion, intrinsic albedo, and scene geometry.

Task 3: LumiLens achieves the lowest translation error and best motion consistency, with rotation error competitive with EX-4D (Table 6), supporting precise camera control and view-consistent appearance. LumiLens also achieves lower CamMC at 60°–90° in both domains (Appendix C.4).

Qualitative Evaluation. Figure 17 corroborates our quantitative findings with superior relighting under dynamic cameras. Figures 8 and 9 demonstrate robust generalization and versatility across challenging indoor/outdoor scenes: physically plausible, diverse light–scene interactions and seamless single- and multi-light control under varied trajectories. See the supplementary video for dynamic results.

Table 2: Task 1: Quantitative results for Choreographing Light and Lens on in-the-wild videos. Ours (A)/(B): ablation Models A/B (Sec. 4.3).

Method	Subject Consistency↑	Background Consistency↑	Motion Smoothness↑	Aesthetic Quality↑	Imaging Quality↑
Light-X	94.11	93.21	98.36	55.26	71.62
Light-X (Nano-Ref)	94.97	93.76	98.86	56.89	72.11
V-RGBX + EX-4D	94.01	92.33	98.02	54.88	71.54
V-RGBX (Nano-Ref) + EX-4D	94.21	93.01	98.43	55.02	71.49
Ours (A)	94.26	92.90	98.66	53.81	71.80
Ours (B)	94.33	92.81	98.12	54.25	71.23
Ours	**95.18**	**94.58**	**99.37**	**58.17**	**72.81**

4.3 Ablation Studies and Analysis

Tables 1 and 2 evaluate architectural choices on synthetic and in-the-wild videos.

Color and Intensity Global Injection. Model A tests explicit spatial light encoding by Fourier-encoding target color and intensity as global parameters and injecting them through DiT cross-attention, as in ambient control. Its Spatial Stream receives only a binary light-source mask, losing local intensity/color gradients. Tables 1, 2, and Fig. 13 show uniformly tinted, "flat" shading without high-frequency 3D definition, realistic lighting fall-off, or volumetric effects. This confirms that fine-grained local relighting requires spatially-explicit parameter maps.

Depth Conditioning vs. Warped RGB Conditioning. Model B tests our division of geometry and semantics by replacing warped depth with RGB in the Spatial Stream and removing source-video concatenation in the Global Stream. Tables 1, 2, and Fig. 14 reveal two failures: (i) **"Baked-In" Artifacts:** warped RGB retains source-view shadows, producing ghosting alongside new shadows; (ii) **Texture Loss:** removing the source video loses fine details (e.g., fabric patterns), significantly degrading SSIM and LPIPS. These results validate depth for geometry guidance and the source video for semantic anchoring. Adding source concatenation to warped RGB (Model C) only marginally improves PSNR (Appendix A.2.6).

See Appendices A.2.1–A.2.3 for robustness analyses and C.5 for the user study.

Limitations and Future Work. Despite OVI's logical consistency, implicit radiance transport yields perceptual plausibility without guaranteed photometric accuracy (e.g., Lux or Nits). Monocular occlusions and backbone sequence limits hinder extreme viewpoints (e.g., 360° orbits) and long-term coherence. In occluded close-ups, inaccurate depth can cause minor floaters; stronger depth estimators could improve stability. High-frequency detail remains backbone-limited; future work will explore larger datasets and stronger backbones such as MiniMax H3 (MiniMax, 2026). Fully unobserved off-screen emitters cannot be controlled independently, although α modulates environmental illumination; with more training data, our design can support off-screen directional key lighting. We also aim to move beyond frame-wise generation toward continuous 4D representations with updated light primitives, geometry, and dynamics, supporting loop closure and more rigorous illumination modeling.

5 Conclusion

LumiLens is the first framework for joint camera-trajectory and parametric local-illumination control in monocular video. Its scalable OVI dataset and novel dual-stream conditioning effectively overcome the "Baked-In Reflection Trap" of prior warping-based approaches. Experiments demonstrate high-fidelity, view-dependent radiance capturing complex physical light transport beyond simple 2D style transfer. This establishes a new standard for controllable video generation, enabling authentic, physics-aware virtual cinematography from flat footage.

AI USE STATEMENT

We used generative AI tools to assist with drafting and polishing the manuscript, adjusting its structure, and suggesting titles. The authors reviewed and revised the AI-assisted text and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We prioritize the transparency and accessibility of our research to foster future innovation. All methodological details, ranging from the Omni-View Illumination (OVI) data generation pipeline to the specific training hyperparameters, as well as detailed descriptions of our proposed *LumiLens*, have been rigorously documented in the main paper and the appendix: OVI dataset construction in Appendix B.2, implementation and training details in Appendix C.1, baseline configurations in Appendix C.2, evaluation metrics in Appendix C.3, and per-case control settings and camera trajectories in Appendix D. To facilitate independent verification and further development, we will open-source our full implementation, including the complete codebase, pre-trained model checkpoints, and the curated dataset. This commitment promotes transparency, enables replication, and encourages the research community to extend and improve upon our work in controllable video generation.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3D: taming large video diffusion transformers for 3d camera control. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=0n4bS0R5MM.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from A single video. *CoRR*, abs/2503.11647, 2025. doi: 10.48550/ARXIV.2503.11647. URL https://doi.org/10.48550/arXiv.2503.11647.

Quentin Bammey. Experiments on deep single-image portrait relighting. *Image Process. Line*, 12: 420–434, 2022. doi: 10.5201/IPOL.2022.429. URL https://doi.org/10.5201/ipol.2022.429.

Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1670–1687, 2015. doi: 10.1109/TPAMI.2014.2377712. URL https://doi.org/10.1109/TPAMI.2014.2377712.

Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *CoRR*, abs/2501.02690, 2025a. doi: 10.48550/ARXIV.2501.02690. URL https://doi.org/10.48550/arXiv.2501.02690.

Weikang Bian, Xiaoyu Shi, Zhaoyang Huang, Jianhong Bai, Qinghe Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Hongsheng Li. Relightmaster: Precise video relighting with multi-plane light images. *CoRR*, abs/2511.06271, 2025b. doi: 10.48550/ARXIV.2511.06271. URL https://doi.org/10.48550/arXiv.2511.06271.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. doi: 10.48550/ARXIV.2311.15127. URL https://doi.org/10.48550/arXiv.2311.15127.

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.

Chris Careaga and Yağız Aksoy. Physically controllable relighting of photographs. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715402. doi: 10.1145/3721238.3730666. URL https://doi.org/10.1145/3721238.3730666.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Junuk Cha, Mengwei Ren, Krishna Kumar Singh, He Zhang, Yannick Hold-Geoffroy, Seunghyun Yoon, Hyunjoon Jung, Jae Shin Yoon, and Seungryul Baek. Text2relight: Creative portrait relighting with text guidance. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 1980–1988. AAAI Press, 2025. doi: 10.1609/AAAI.V39I2.32194. URL https://doi.org/10.1609/aaai.v39i2.32194.

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. LBM: latent bridge matching for fast image-to-image translation. *CoRR*, abs/2503.07535, 2025. doi: 10.48550/ARXIV.2503.07535. URL https://doi.org/10.48550/arXiv.2503.07535.

Dong-Yu Chen, Yixin Guo, Shuojin Yang, Tai-Jiang Mu, and Shi-Min Hu. Beyond inpainting: Unleash 3d understanding for precise camera-controlled video generation. *CoRR*, abs/2601.10214, 2026. doi: 10.48550/ARXIV.2601.10214. URL https://doi.org/10.48550/arXiv.2601.10214.

Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. Scribblelight: Single image indoor relighting with scribbles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5720–5731. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00537. URL https://openaccess.thecvf.com/content/CVPR2025/html/Choi_ScribbleLight_Single_Image_Indoor_Relighting_with_Scribbles_CVPR_2025_paper.html.

Epic Games. *Unreal Engine 5 - a real-time 3D creation tool*. Epic Games, Inc., Cary, NC, 2022. URL https://www.unrealengine.com/en-US/unreal-engine-5. Available at: https://www.unrealengine.com/en-US/unreal-engine-5.

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. *CoRR*, abs/2501.16330, 2025a. doi: 10.48550/ARXIV.2501.16330. URL https://doi.org/10.48550/arXiv.2501.16330.

Ye Fang, Tong Wu, Valentin Deschaintre, Duygu Ceylan, Iliyan Georgiev, Chun-Hao Paul Huang, Yiwei Hu, Xuelin Chen, and Tuanfeng Yang Wang. V-rgbx: Video editing with accurate controls over intrinsic properties, 2025b. URL https://arxiv.org/abs/2512.11799.

Google DeepMind. Gemini Image Pro (nano-banana pro). https://deepmind.google/models/gemini-image/pro/, 2026. Accessed: 2026-01-21.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Fx2SbBgcte.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *CoRR*, abs/2404.02101, 2024. doi: 10.48550/ARXIV.2404.02101. URL https://doi.org/10.48550/arXiv.2404.02101.

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. *CoRR*, abs/2506.15673, 2025. doi: 10.48550/ARXIV.2506.15673. URL https://doi.org/10.48550/arXiv.2506.15673.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. EX-4D: extreme viewpoint 4d video synthesis via depth watertight mesh. *CoRR*, abs/2506.05554, 2025a. doi: 10.48550/ARXIV.2506.05554. URL https://doi.org/10.48550/arXiv.2506.05554.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 2005–2015. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.00193. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hu_DepthCrafter_Generating_Consistent_Long_Depth_Sequences_for_Open-world_Videos_CVPR_2025_paper.html.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 21807–21818. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 25096–25106. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02371. URL https://doi.org/10.1109/CVPR52733.2024.02371.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 10486–10496. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00981. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_MegaSaM_Accurate_Fast_and_Robust_Structure_and_Motion_from_Casual_CVPR_2025_paper.html.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. Illumicraft: Unified geometry and illumination diffusion for controllable video generation. *CoRR*, abs/2506.03150, 2025b. doi: 10.48550/ARXIV.2506.03150. URL https://doi.org/10.48550/arXiv.2506.03150.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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*, 2025a.

Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. *CoRR*, abs/2503.20785, 2025b. doi: 10.48550/ARXIV.2503.20785. URL https://doi.org/10.48550/arXiv.2503.20785.

Yang Liu, Chuanchen Luo, Zimo Tang, Yingyan Li, Yuran Yang, Yuanyong Ning, Lue Fan, Zhaoxiang Zhang, and Junran Peng. Tc-light: Temporally coherent generative rendering for realistic world transfer. *CoRR*, abs/2506.18904, 2025c. doi: 10.48550/ARXIV.2506.18904. URL https://doi.org/10.48550/arXiv.2506.18904.

Yue Ma, Jiangming Wang, Yucheng Wang, Xilai Wang, Zhiyuan Li, Xinyu Wang, Hongyu Liu, Ruofan Liang, Songchun Zhang, Yuxuan Xue, and Qifeng Chen. LiveLight: Real-time streaming video relighting with interactive control. *CoRR*, abs/2608.01771, 2026. doi: 10.48550/ARXIV.2608.01771. URL https://doi.org/10.48550/arXiv.2608.01771.

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. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715402. doi: 10.1145/3721238.3730696. URL https://doi.org/10.1145/3721238.3730696.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities, July 2026. URL https://www.minimax.io/blog/minimax-h3.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Pexels. Pexels. https://www.pexels.com, 2025. Accessed: 2026-01-14.

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 *International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3d-informed world-consistent video generation with precise camera control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 6121–6132. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00574. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_GEN3C_3D-Informed_World-Consistent_Video_Generation_with_Precise_Camera_Control_CVPR_2025_paper.html.

Soumyadip Sengupta, Angjoo Kanazawa, Carlos Domingo Castillo, and David W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces 'in the wild'. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6296–6305. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00659. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Sengupta_SfSNet_Learning_Shape_CVPR_2018_paper.html.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3d worlds. *CoRR*, abs/2604.13036, 2026. doi: 10.48550/ARXIV.2604.13036. URL https://doi.org/10.48550/arXiv.2604.13036.

Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. In *The Twelfth International Conference on Learning Representations*, 2024.

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. Style transfer for headshot portraits. *ACM Trans. Graph.*, 33(4), July 2014. ISSN 0730-0301. doi: 10.1145/2601097.2601137. URL https://doi.org/10.1145/2601097.2601137.

Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Trans. Graph.*, 36(4), 2017. doi: 10.1145/3072959.3126847. URL https://doi.org/10.1145/3072959.3126847.

Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. *CoRR*, abs/2509.15130, 2025. doi: 10.48550/ARXIV.2509.15130. URL https://doi.org/10.48550/arXiv.2509.15130.

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E. Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79:1–79:12, 2019. doi: 10.1145/3306346.3323008. URL https://doi.org/10.1145/3306346.3323008.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pp. 402–419. Springer, 2020. doi: 10.1007/978-3-030-58536-5_24. URL https://doi.org/10.1007/978-3-030-58536-5_24.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rylgEULtdN.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan

Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 8428–8437. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00789. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Koala-36M_A_Large-scale_Video_Dataset_Improving_Consistency_between_Fine-grained_Conditions_CVPR_2025_paper.html.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In Andres Burbano, Denis Zorin, and Wojciech Jarosz (eds.), *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, pp. 114. ACM, 2024. doi: 10.1145/3641519.3657518. URL https://doi.org/10.1145/3641519.3657518.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: create anything in 4d with multi-view video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 26057–26068. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02427. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_CAT4D_Create_Anything_in_4D_with_Multi-View_Video_Diffusion_Models_CVPR_2025_paper.html.

Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 442–452. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00050. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xing_LumiNet_Latent_Intrinsics_Meets_Diffusion_Models_for_Indoor_Scene_Relighting_CVPR_2025_paper.html.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M. Patel, and Yiqun Mei. Wonder: Video world model done better. *CoRR*, abs/2607.26037, 2026. doi: 10.48550/ARXIV.2607.26037. URL https://doi.org/10.48550/arXiv.2607.26037.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In Andres Burbano, Denis Zorin, and Wojciech Jarosz (eds.), *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, pp. 113. ACM, 2024. doi: 10.1145/3641519.3657481. URL https://doi.org/10.1145/3641519.3657481.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Ying Shan, and Kai Han. RayPE: Ray-space positional encoding for 3d-aware video generation. *CoRR*, abs/2606.27345, 2026. doi: 10.48550/ARXIV.2606.27345. URL https://doi.org/10.48550/arXiv.2606.27345.

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *CoRR*, abs/2503.05638, 2025. doi: 10.48550/ARXIV.2503.05638. URL https://doi.org/10.48550/arXiv.2503.05638.

Jianshu Zeng, Yuxuan Liu, Yutong Feng, Chenxuan Miao, Zixiang Gao, Jiwang Qu, Jianzhang Zhang, Bin Wang, and Kun Yuan. Lumen: Consistent video relighting and harmonious background replacement with video generative models. *CoRR*, abs/2508.12945, 2025. doi: 10.48550/ARXIV.2508.12945. URL https://doi.org/10.48550/arXiv.2508.12945.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 2050–2062. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.00197. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_ReCapture_Generative_Video_Camera_Controls_for_User-Provided_Videos_using_Masked_CVPR_2025_paper.html.

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, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=u1cQYxRI1H.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *CoRR*, abs/2410.15957, 2024. doi: 10.48550/ARXIV.2410.15957. URL https://doi.org/10.48550/arXiv.2410.15957.

Sixiao Zheng, Zimian Peng, Yanpeng Zhou, Yi Zhu, Hang Xu, Xiangru Huang, and Yanwei Fu. Vidcraft3: Camera, object, and lighting control for image-to-video generation. *CoRR*, abs/2502.07531, 2025. doi: 10.48550/ARXIV.2502.07531. URL https://doi.org/10.48550/arXiv.2502.07531.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-a-video: Training-free video relighting via progressive light fusion. *CoRR*, abs/2502.08590, 2025. doi: 10.48550/ARXIV.2502.08590. URL https://doi.org/10.48550/arXiv.2502.08590.

A VISUAL RESULTS

A.1 PARAMETRIC CONTROL AND RESULT GALLERIES

Fine-grained Parametric Control. Figure 7 illustrates fine-grained parametric control under target trajectories, including simultaneous dual-light, independent single-light, and ambient modulation.

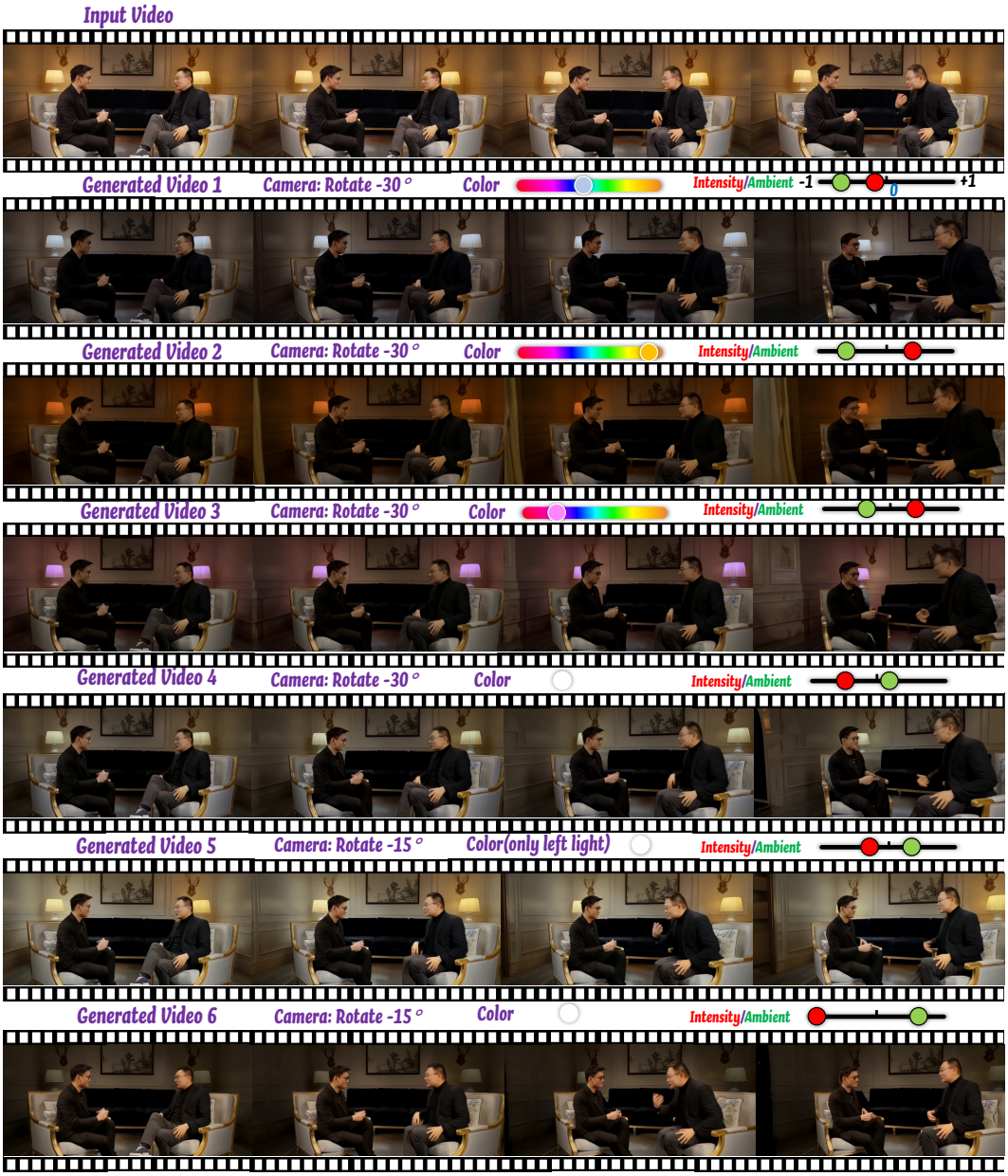


Figure 7: *LumiLens* enables fine-grained parametric control under target trajectories, supporting simultaneous dual-light, independent single-light, and ambient modulation. Notably, the "Intensity/Ambient" bar regulates luminosity: positive values enhance lighting, while negative values attenuate it.

Result Galleries. To showcase the robust generalization and versatility of *LumiLens*, we provide extensive result galleries in Figs. 8 and 9. These galleries demonstrate our framework's capacity to synthesize physically plausible, highly diverse light-scene interactions across a wide range of challenging indoor and outdoor scenarios.

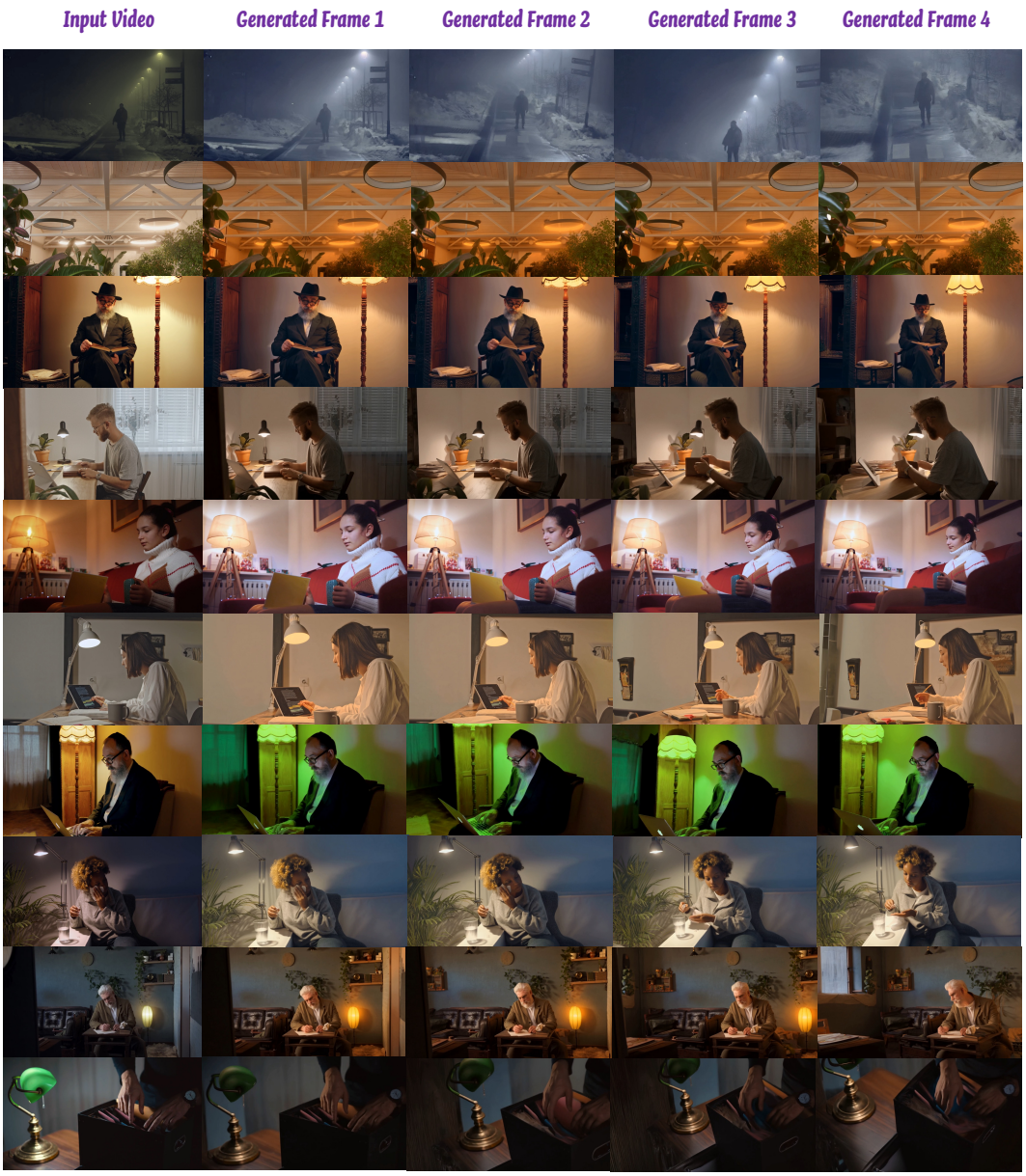


Figure 8: Result galleries. Non-consecutive frames; see the supplementary video for more intuitive results.

The result gallery continues in Fig. 9. Non-consecutive frames are shown for illustration; please refer to the supplementary video for more intuitive dynamic results.



Figure 9: Result galleries. Non-consecutive frames; see the supplementary video for more intuitive results.

A.2 Ablation, Robustness, and Generalization

A.2.1 Robustness to Noisy Depth

Since our method relies on depth estimation, its resilience to input degradation is also evaluated. Varying levels of Gaussian noise are introduced to the depth maps:

$$\tilde{D}_S = D_S + \epsilon \cdot D_S, \qquad \epsilon \sim \mathcal{N}(0, \text{rate}), \tag{9}$$

where *rate* controls the perturbation strength. To rigorously stress-test the model, challenging scenes featuring fine geometric structures (e.g., the slender supporting stem of a floor lamp), which are notoriously prone to depth artifacts, are specifically selected.

Table 3: Ablation on noisy depth. Quantitative results on Synthetic Validation Dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓
Ours(rate 0%)	**21.761**	**0.877**	**0.216**	**370.47**
Ours(rate 1%)	21.398	0.852	0.221	376.11
Ours(rate 2%)	21.273	0.844	0.229	378.82
Ours(rate 3%)	20.162	0.801	0.237	382.20
Ours(rate 4%)	19.681	0.785	0.249	390.44
Ours(rate 5%)	18.039	0.729	0.266	401.27

Observation: As visualized in Fig. 10, even under moderate noise levels that typically corrupt thin structures, *LumiLens* demonstrates graceful degradation: despite geometric jitter, the global semantic anchor effectively stabilizes the generation, ensuring coherent illumination and motion consistency. Performance under different noise levels is summarized in Table 3.

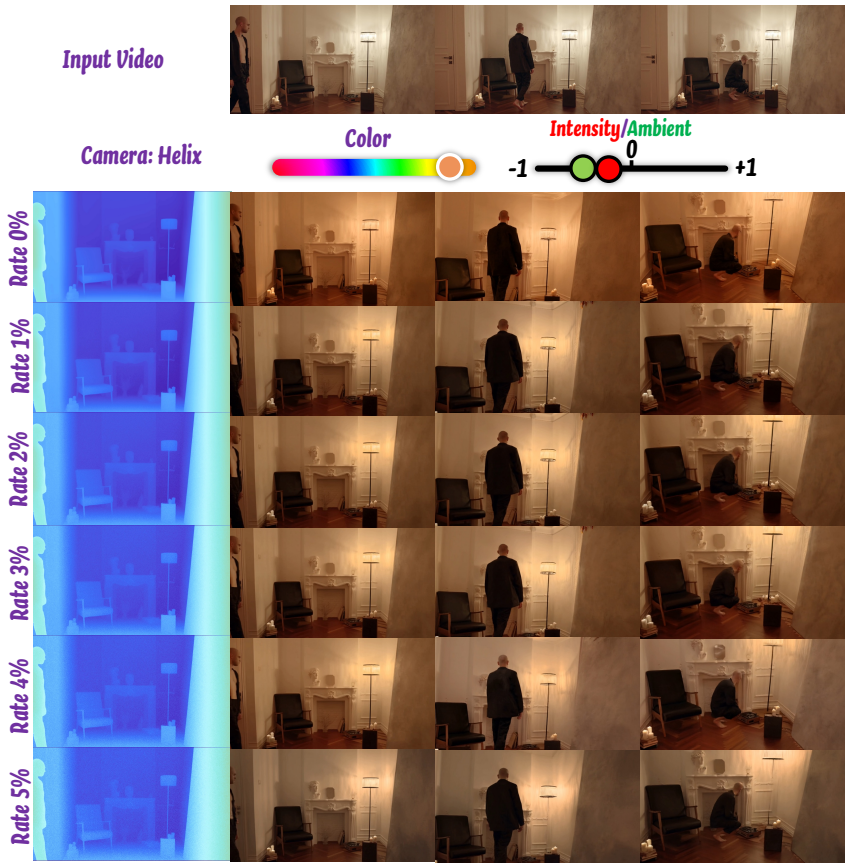


Figure 10: Analysis on Noisy Depth Robustness.

A.2.2 Robustness to Light Source Occlusion

Scenarios with light source occlusion, where the emitter is physically blocked by scene objects, are further evaluated. As demonstrated in Fig. 11, *LumiLens* resolves ray visibility, rendering plausible shadow casting and natural ambient darkening consistent with the occlusion geometry.



Figure 11: Analysis on Light Occlusion Robustness.

A.2.3 Robustness on Non-Lambertian Surfaces

Standard relighting struggles with non-Lambertian surfaces. Scenes featuring mirrors, ceramic tiles, and wet flooring are evaluated to verify our model's capability in handling view-dependent effects. **Observation:** As shown in Fig. 12, our model renders specular highlights that correctly shift with the camera view. Crucially, reflections exhibit chromatic consistency, updating synchronously with light source illumination changes. This confirms that *LumiLens* implicitly models complex and view-dependent radiance transport, rather than relying on trivial texture memorization.

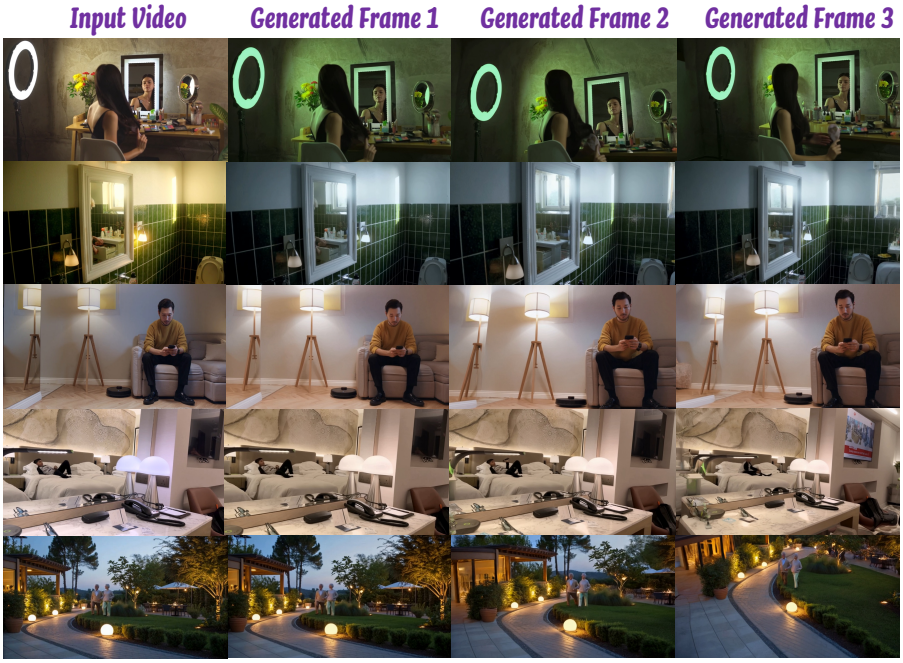


Figure 12: Video generation results on non-Lambertian scenes.

A.2.4 Color and Intensity Global Injection (Model A)

Model A treats target light color and intensity as global parameters, encoded via Fourier features and injected into the DiT via cross-attention; the Spatial Stream receives only a binary light-source mask. As shown in Fig. 13, this variant fails to generate realistic lighting fall-off and volumetric effects. The global injection forces a uniform tint across the scene, resulting in "flat" shading that lacks high-frequency 3D definition.

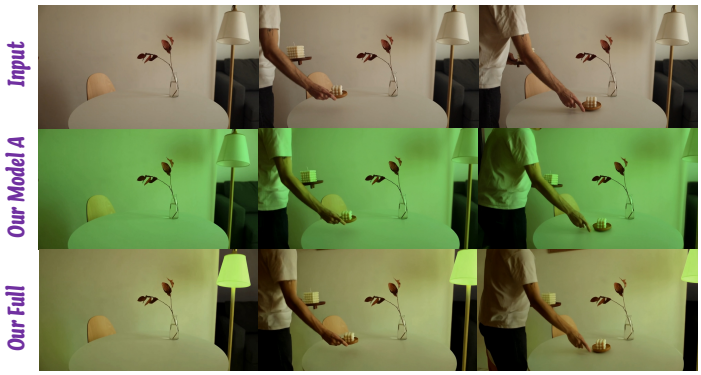


Figure 13: Ablation on Color and Intensity global injection (Model A).

A.2.5 Warped RGB Conditioning (Model B)

Model B replaces warped depth in the Spatial Stream with warped RGB frames and removes concatenation of the source video in the Global Stream. Figure 14 shows the resulting "Baked-In" artifacts and texture loss: warped RGB carries source-view shadows, causing ghosting where old and new shadows coexist, while omitting the source video causes fine texture loss.



Figure 14: Ablation on Warped RGB Condition (Model B).

A.2.6 Interaction with Source Concatenation

Model C retains the warped-RGB condition of Model B and restores source-video concatenation in the Global Stream. Table 4 shows only a marginal PSNR increase from 16.126 to 16.292, since pose-aligned RGB already carries appearance. With warped depth, the source video instead provides the appearance reference. This comparison examines the interaction between warp modality and source concatenation.

Table 4: Effect of source-video concatenation with warped RGB on the synthetic joint-control task.

Variant	Warped condition	Source video	PSNR↑
Model B	RGB	No	16.126
Model C	RGB	Yes	16.292
LumiLens	Depth	Yes	**21.761**

A.2.7 Geometry-Guided Attention

Figure 15 compares source-video attention between *LumiLens* and the RGB variant: *LumiLens* attends precisely to the corresponding source-video patches and produces consistent outputs, whereas the RGB variant produces distorted structures.

To further validate that *LumiLens* inherently comprehends the underlying 3D scene geometry—a prerequisite for physically plausible relighting—we visualize the internal cross-attention maps. Specifically, we compare our dual-stream architecture against an RGB-only guided baseline, examining the attention weights corresponding to localized geometric structures (e.g., the "wheel" in Fig. 15).

As illustrated, the RGB-only baseline exhibits a diffuse and rough attention distribution. The attention weights on the target wheel are weakly activated and severely dispersed across the surrounding background, making them almost indistinguishable from unrelated regions. This indicates that without explicit geometric constraints, the network's attention is easily distracted by complex 2D textures and entangled source illumination, failing to establish accurate spatial correspondences.

In stark contrast, *LumiLens* produces a remarkably sharp and highly concentrated attention map tightly bounded to the object's physical contours. The attention weights on the wheel are significantly higher than the rest of the scene, with distinct boundaries that cleanly separate the foreground structure from the background. This sharp activation provides compelling evidence that our explicit geometric guidance (e.g., warped depth) effectively anchors the network's attention to the correct 3D geometry. By eliminating attention dispersion and suppressing spurious 2D correlations, *LumiLens* ensures that newly synthesized light transport effects are precisely grounded in the underlying physical structure, ultimately leading to highly realistic and view-consistent re-rendering.

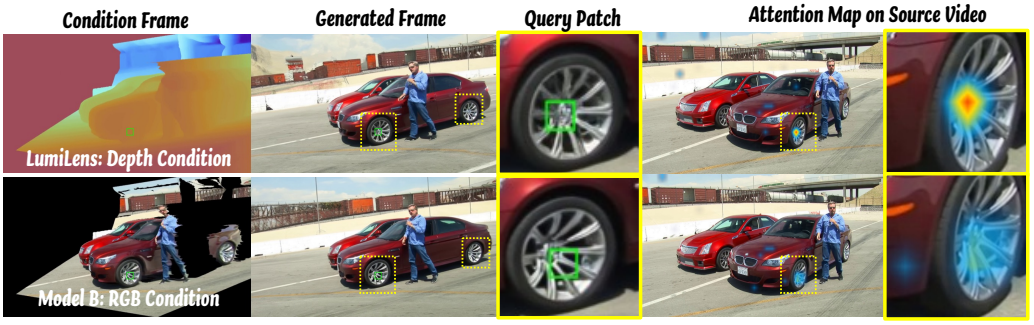


Figure 15: Geometry-guided attention. The green box marks the query patch; source-video attention is more concentrated around the wheel with depth conditioning than with the RGB-guided baseline.

A.2.8 OUT-OF-DOMAIN EMITTERS

Although trained exclusively on conventional lighting fixtures, *LumiLens* seamlessly generalizes to modulate the illumination of out-of-domain objects, such as computer monitors or windows, as shown in Fig. 16. This emergent zero-shot capability reveals a fundamental strength of our design: it effectively decouples the spatial modeling of radiance from intrinsic object semantics. Because our Spatial Stream provides purely geometric conditioning (e.g., masks and depth) devoid of object identity, the network avoids overfitting to the texture correlations associated with "lamps." Instead, it internalizes the universal, physics-based mechanisms of light transport. By comprehending this underlying geometric cause-and-effect, *LumiLens* learns to treat any user-defined masked region as a valid physical emitter, transforming arbitrary semantic entities into physically plausible light sources.

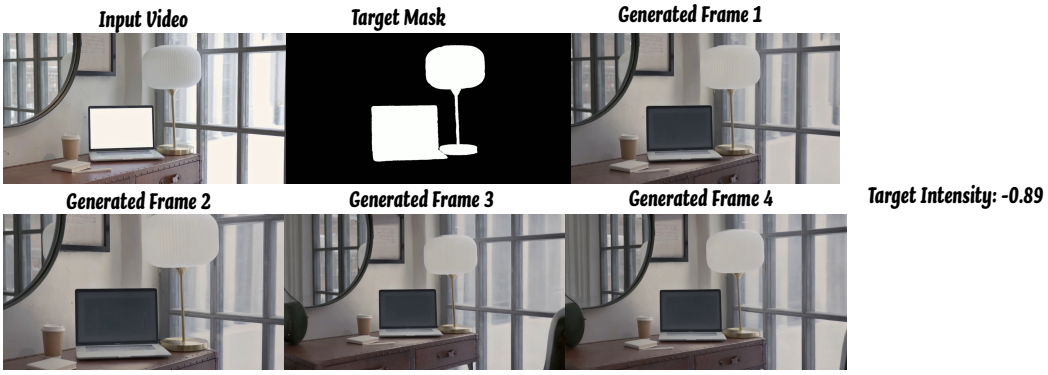


Figure 16: Out-of-domain illumination control for emitters such as monitors and windows.

A.3 Comparisons with Baselines

Figure 17 presents visual comparisons for Tasks 1 and 2. **Disentanglement & Spatially-varying Control.** Comparisons in Tasks 1 and 2 reveal fundamental limitations in current baselines. RelightVid and Light-X (without NanoBanana reference) act merely as global 2D filters, uniformly tinting the scene without spatial logic. Even when relying on high-quality references from closed-source models, Light-X and V-RGBX fail to disentangle lighting from geometry, resulting in severe artifacts and identity degradation rather than physical re-rendering. In contrast, *LumiLens* demonstrates robust decoupled controllability: it synthesizes true spatially-varying effects—such as dynamic cast shadows that elongate with geometry and specular highlights that glide across surfaces—while preserving fine structural details and subject identity.



Figure 17: Result Comparisons. RelightVid performs relighting only, and V-RGBX uses EX-4D for camera control. Non-consecutive frames; see the supplementary video for more intuitive results.

Faithful Viewpoint & Structural Integrity. In Task 3 (Fig. 18), distinct failure modes are observed in baselines. ReCamMaster often struggles to faithfully adhere to the user-specified camera trajectories, leading to significant pose drift. EX-4D, heavily reliant on warping, exhibits characteristic symptoms of the "Baked-In Reflection Trap": severe structural stretching artifacts and static highlights that remain rigidly attached to textures. These artifacts mirror the warping failures analyzed in our Ablation Study (Fig. 14). In contrast, *LumiLens* maintains faithful trajectory adherence, enabling it to synthesize clean, structurally intact geometry. With precise viewpoint control, textures naturally align with the underlying 3D shape rather than being blindly pasted onto distorted meshes.

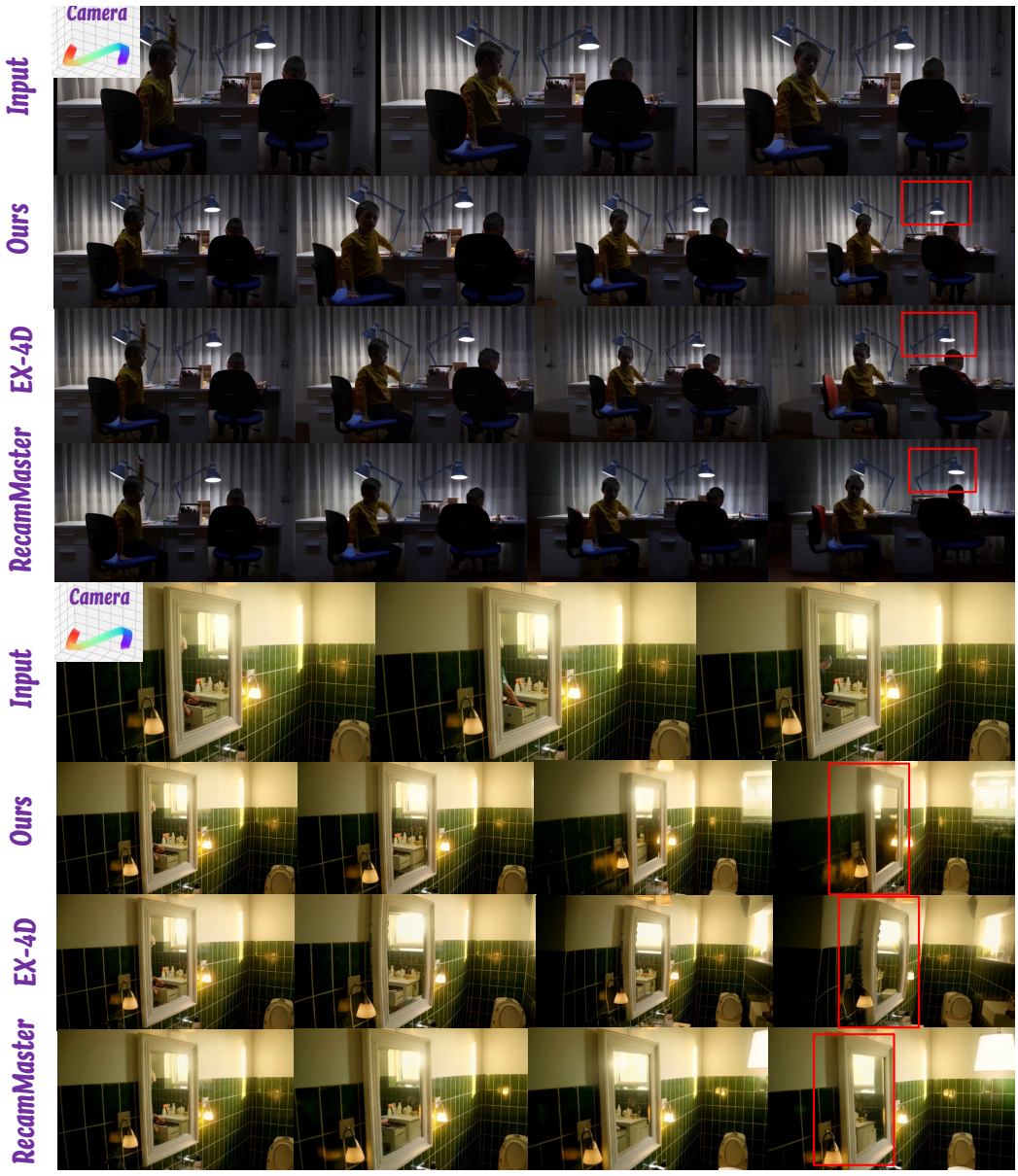


Figure 18: Comparisons on Recam-only Task.

Comparison with ScribbleLight. While *ScribbleLight* (Choi et al., 2025) is a notable open-source framework for interactive image local relighting, we excluded it from the main paper's baseline comparisons in favor of *IC-Light* (Zhang et al., 2025b). This decision was driven by the significant gap in generation fidelity and physical plausibility. As shown in Fig. 19, *ScribbleLight* relies on sparse user scribbles to guide colorization. However, we observed that this paradigm often fails to model complex light transport, resulting in outputs that appear flat, cartoon-like, or overly stylized,

rather than photorealistic. It tends to "paint over" the original textures with the target color (acting more like a segmentation-based color transfer) rather than simulating how light interacts with the underlying 3D geometry. Given our objective of high-fidelity video re-rendering, *ScribbleLight* does not meet the quality threshold required for a meaningful comparison. This observation aligns with the comparative analysis in *LightLab* (Magar et al., 2025), where our reproduction confirms that the method struggles to generate photorealistic lighting effects compatible with real-world footage. Consequently, we selected *IC-Light*, which represents the current state-of-the-art in diffusion-based relighting, to serve as a more rigorous and competitive baseline.



Figure 19: Additional visual comparison with ScribbleLight. IC-Light remains the frame-wise relighting baseline in the main quantitative evaluation.

A.4 OVI Dataset Visualization

Data Preparation Example. Figure 20 illustrates the placement of characters and light fixtures in an OVI scene. Detailed generation settings are provided in Appendix B.2.



Figure 20: Example scene used for OVI data preparation.

Dataset Examples. Figure 21 illustrates the OVI dataset, which combines fine-grained parametric illumination control and synchronized cameras with dynamic objects. Construction details are in Appendix B.2.

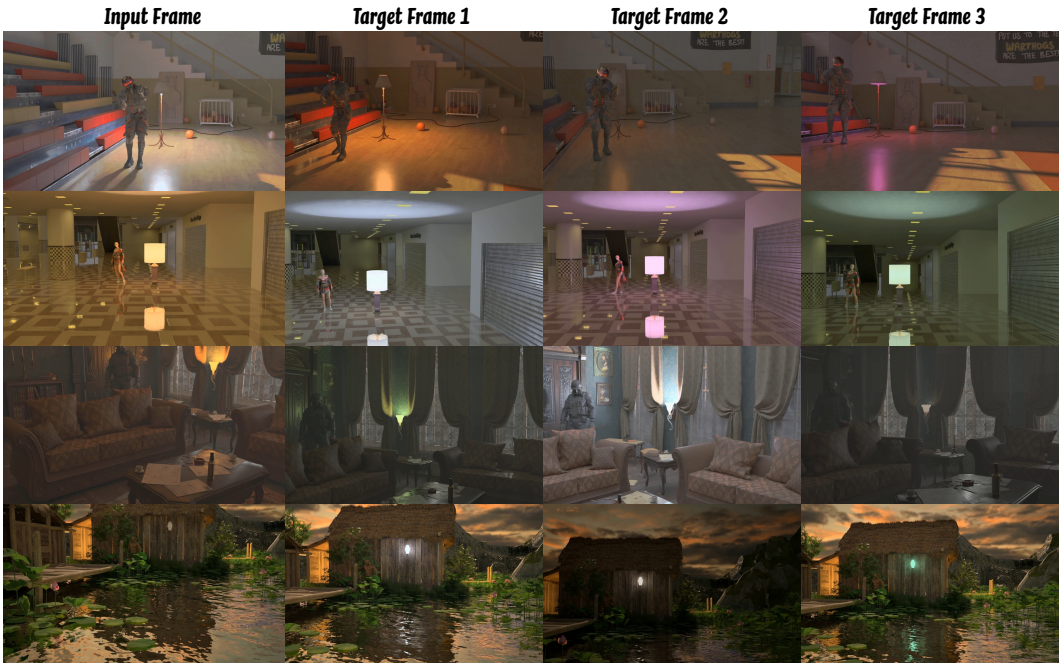


Figure 21: Illustration of OVI Dataset. We propose OVI data curation pipeline to generate scalable datasets with fine-grained parametric illumination control and synchronized cameras that include dynamic objects.

B PRELIMINARIES AND METHOD DETAILS

B.1 TEXT-TO-VIDEO PRELIMINARIES

Text-to-Video Base Models. This study is conducted using the Wan 2.2 pre-trained text-to-video foundation model (Wan et al., 2025). The architecture comprises a three-dimensional Variational Auto-Encoder (VAE) $\mathcal{E}$ (Kingma & Welling, 2014) for latent space mapping and a series of Diffusion Transformer (DiT) blocks for sequence modeling. Each basic DiT block consists of 2D spatial self-attention, 3D spatial-temporal attention, cross-attention, and a feed-forward network (FFN). The text prompt embedding c_{text} is obtained by a T5 encoder ε_{T5} (Raffel et al., 2020) and injected into the model through cross-attention. The diffusion transformer is trained under the Rectified Flow framework (Lipman et al., 2023), enabling the generation of data sample $\boldsymbol{x}$ from an initial Gaussian sample $\boldsymbol{z} \in \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. For a data point $\boldsymbol{x}$, its noised version at timestep t is

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x} + t\boldsymbol{z}. \tag{10}$$

The training objective is defined by a simple mean squared error (MSE) loss:

$$\mathcal{L}_{RF}(\theta) = \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{z}} \left\| \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_{\text{text}}) - (\boldsymbol{z} - \boldsymbol{x}) \right\|_2^2, \tag{11}$$

where the velocity $\boldsymbol{v}_\theta$ is parameterized by the network θ.

B.2 OVI DATASET CONSTRUCTION

Asset Collection and Sources. A diverse asset library is curated, comprising 120 high-quality scenes (80 indoor, 40 outdoor) and 80 distinct light fixture models (e.g., street lamps, wall sconces for outdoors; floor, table, and pendant lamps for indoors). Each fixture is compatible with Point, Spot, and Rect light sources to ensure generalization across light types. Assets include 500 animated characters and 1,000 motion clips. The environments and fixture assets are collected from `https://www.fab.com` and publicly available online sources.

Hierarchical Procedural Layout Generation. A four-stage hierarchical layout algorithm is designed to automate the generation of aesthetically pleasing and physically valid dynamic scenes:

(1) **Context-Aware Light Placement:** The pipeline first identifies a "Target Light", either selecting an existing light source within the scene or automatically placing a new fixture based on semantic context (e.g., placing a table lamp on a desk). A headless UE5 scripting pipeline traverses scenes, identifies lamps, lanterns, and candles through semantic keyword matching and socket analysis, and attaches Point, Rect, or Spot lights to predefined sockets. Intensity is sampled from $U[I_{\min}, I_{\max}]$, color temperature from $U[2800\text{K}, 4500\text{K}]$, and barn-door angles are randomized. Emissive and non-emissive material instances are swapped between ambient-only and target-light-only rendering modes.

(2) **Camera Space Analysis:** A valid initial camera position relative to the target light is identified via rigorous collision and occlusion detection, ensuring a clear line of sight and optimal composition. A spherical camera "safe zone" around this position prevents clipping into scene geometry. Candidate poses constrain camera–target distance to $[d_{\min}, d_{\max}]$ and clamp pitch. A sphere trace of radius R_{safe} rejects camera collisions. A 4×4 raycasting grid rejects views with clutter ratio above 0.3. Initial-view visibility is checked from the camera center and four corner rays.

(3) **Character Instantiation:** Based on the camera's view frustum and the target light's position, a cylindrical "activity zone" is calculated. A random character with a random motion clip is instantiated within this zone, ensuring visibility without intersecting scene objects. Avatar placement follows the rule of thirds, balancing the target light across the frame. Vertical ray tracing locates the ground, and capsule overlap tests reject intersections with the environment or target object.

(4) **Cinematographic Camera Trajectories:** Twelve predefined camera movement patterns (e.g., pan, dolly, truck) are implemented within the calculated safe sphere. A "Look-At" constraint ensures the target light remains within the camera frustum. Crucially, strict visibility is not enforced, allowing the light fixture to be naturally occluded by characters or scene geometry during movement, thereby simulating realistic inference scenarios. Directed cinematic primitives include Helix-In, S-Curve Dolly, Parabola Dive, and Orbit-Ascend. Camera rotation uses a linearly interpolated virtual look-at target for temporal stability.

It is crucial to note that while these presets are utilized to systematically cover the motion space during data generation, the trained *LumiLens* model effectively generalizes to support arbitrary, user-defined, frame-by-frame continuous camera trajectories during inference.

Decoupled Rendering and Configuration. Inspired by LightLab (Magar et al., 2025), for each frame of a generated sequence, two physically separated linear RGB streams are rendered: (1) An **Ambient Stream (V_{amb})**: Rendered with the target light turned *off*, illuminated solely by the environment map and other background lights. (2) A **Target Light Stream (V_{light})**: Rendered with the target light set to a canonical high intensity and pure white color, while all other light sources are turned *off*. This decomposition underlies subsequent linear-space lighting compositing.

The Movie Render Queue produces synchronized ambient, target-light, and geometric-data passes at 10 FPS. The first two use the UE5 Path Tracer with 512 spatial samples per pixel, eight light bounces, and denoising enabled via `r.PathTracing.Denoiser=1`. A fixed exposure offset of -5.0 EV prevents HDR clipping. Depth and segmentation use rasterization: a world-depth post-process material exports absolute depth, and stencil values identify actors. Lighting, shadows, fog, and atmospheric effects are disabled for geometric-data rendering. The radiance passes disable the filmic tone curve, motion blur, bloom, lens flares, vignette, and chromatic aberration, and are exported as 16-bit floating-point multilayer OpenEXR files with lossless PIZ compression. JSON manifests drive Python/MRQ jobs, with fixed engine settings such as maximum mesh detail and disabled texture streaming.

Compositing and Sequence-Level Tone Mapping. Training videos are composed according to Eq. 8. Here, $\hat{\alpha}$ and $\hat{\gamma}$ serve as compositing coefficients scaled relative to the canonical rendered intensity (e.g., 0.5 denotes 50% of the maximum brightness). The offline coefficients are sampled as $\hat{\alpha} \sim U(0.1, 1.0)$ and $\hat{\gamma} \sim U(0.0, 1.1)$. Colors are sampled from temperatures in $[2500\text{K}, 9000\text{K}]$ or random RGB values. Simple linear modulation of a canonical white basis yields scalable, diversely lit videos from one render pass, avoiding repeated rendering and drastically reducing computational costs.

However, since rendering is performed in linear color space, standard frame-by-frame tone mapping can cause temporal flickering and inconsistent brightness perception (e.g., auto-exposure darkening the scene as light intensity increases). To address this, a sequence-level tone mapping strategy is implemented: global exposure statistics are computed across the entire clip, and a fixed exposure is applied to all frames. Exposure statistics are shared across frames and trajectories within a rendered batch to avoid flicker. This ensures dynamic range changes introduced by $\hat{\gamma}$ align with physical reality. The composed linear radiance is exposure-normalized, Reinhard tone-mapped, and gamma-encoded for display-referred sRGB video. Inference takes and returns sRGB videos; no linear-space compositing is performed at inference. Consequently, a linear change in a relative control need not yield a linear output-pixel response. Controls express relative edits rather than absolute photometric targets. By randomizing $\mathbf{C}, \hat{\gamma}, \hat{\alpha}$ (8 variations per sequence), the base dataset of 16,800 sequences is inflated to approximately **134,400 training samples** (49 frames at 10 FPS), providing robust supervision for modeling realistic and complex light-scene interactions under diverse camera trajectories.

Training-View Warping and Depth Normalization. For each dynamic scene, eight randomized camera trajectories supply RGB and raw depth. Nvdiffrast rasterizes target-view depth and visibility masks using source-view ground-truth depth. This camera sampling is distinct from the eight illumination variants per base sequence described above. Depth is clipped to $[0.5, 100]$ and normalized logarithmically:

$$\bar{D}_i^r = \frac{\log D_i^r - \log D_{\min}^r}{\log D_{\max}^r - \log D_{\min}^r}, \tag{12}$$

where $D_{\min}^r$ and $D_{\max}^r$ are the minimum and maximum over the depth values being normalized. Data visualizations in this appendix use the `spectral_r` colormap. Colormapping applies to depth; the target-view visibility/disocclusion mask is binary and replicated over three channels for encoding.

C Additional Experimental Details and Results

C.1 Implementation details.

LumiLens is built upon the frozen Wan2.2-TI2V-5B backbone (Wan et al., 2025). The method exclusively optimizes three components (totaling 200M parameters): the 3D-aware Light Adapter, the Ambient Injector, and the LoRA layers of rank 32. Training is conducted on video sequences with a resolution of 704×1280 and a temporal length of 49 frames, using the AdamW optimizer. The learning rate follows a step-decay schedule: it is initialized as 1×10^{-4} for the first 3,000 steps and reduced to 5×10^{-5} thereafter. A two-stage training curriculum is employed: a 1,500-step warm-up stage where only camera trajectories are varied under fixed lighting, followed by 5 epochs of full joint training. The training process converges in 6 days on 32 NVIDIA H20 GPUs. Inference requires 50 seconds per clip using 50 denoising steps on an H800 GPU, positioning it within the top-tier of inference speeds among all evaluated baselines. Appendix D provides the per-case control settings and camera-trajectory parameterization.

C.2 Baseline Configurations

Task 1: Joint Control of Spatially-Varying Illumination and Viewpoint. Since no existing method directly supports this specific joint task, two recent works are adapted as baselines: 1) **Light-X** (Liu et al., 2025a), a framework for global illumination and camera control, which serves to benchmark the difference between global and local light guidance under camera movement. 2) **V-RGBX** (Fang et al., 2025b), a video editing method based on intrinsic decomposition, adapted to evaluate the quality of local lighting changes derived from intrinsic attribute manipulation. As IC-Light (Zhang et al., 2025b) offers limited local relighting capability, the stronger **NanoBanana Pro** (Google DeepMind, 2026) also generates reference images for these two baselines (Nano-Ref).

Task 2: Fine-grained Parametric Light Control. To assess the precision of parametric local lighting control, comparisons are made against: 1) **RelightVid** (Fang et al., 2025a), a leading video relighting method. Since it focuses on global lighting, this comparison highlights the advantages of our spatially-varying local control. 2) **IC-Light**, a state-of-the-art image-based relighting method, applied frame-by-frame to evaluate temporal consistency and the quality of local light interaction.

Task 3: Camera Control and Reflection Consistency. To validate our claim regarding the "Baked-in Reflection Trap", *LumiLens* is compared against representative camera control methods: 1) **ReCamMaster** (Bai et al., 2025), which utilizes implicit camera pose injection. 2) **EX-4D** (Hu et al., 2025a), which relies on explicit camera guidance. These comparisons demonstrate the superiority of the proposed re-rendering paradigm over warping-based approaches in handling view-dependent specularities and warping distortion.

C.3 Evaluation metrics.

To comprehensively assess performance across different tasks and domains, specific metric protocols are employed for synthetic and real-world datasets.

Metrics on Synthetic Data. For quantitative comparisons on the Synthetic Held-out Set (where ground truth is available), all three tasks are evaluated using a comprehensive set of metrics. To assess visual quality and perceptual fidelity, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are reported for pixel-level accuracy, alongside Learned Perceptual Image Patch Similarity (LPIPS) for deep feature similarity. For video-level consistency, Fréchet Video Distance (FVD) (Unterthiner et al., 2019) is employed to measure the distributional alignment between generated and real sequences. Furthermore, to rigorously evaluate geometric structural preservation on relighting task, a point-cloud-based assessment is adopted. Leveraging MegaSAM (Li et al., 2025), dynamic 3D geometry is reconstructed from both generated and reference videos to extract per-frame point clouds. Structural discrepancies are then quantified by computing the Chamfer Distance (CD), with the average value reported.

Chamfer Distance. For reconstructed point sets X and Y, we use

$$\mathrm{CD}(X, Y) = \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} \|x - y\|_2 + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} \|y - x\|_2. \tag{13}$$

Table 5: Task 2: Quantitative results for relighting on in-the-wild videos.

Method	CLIP-S↑	Mo. Pre.↓	CD↓	IAC↓
RelightVid	0.983	0.281	0.4169	2.811
Light-X(Nano-Ref)	0.986	0.257	0.3920	3.408
IC-Light	0.979	0.492	0.5122	3.674
Ours	**0.995**	**0.191**	**0.217**	**1.371**

Metrics on Real-world Data. Since ground truth is unavailable for in-the-wild footages, task-specific metrics are adopted to evaluate perceptual quality, physical consistency, and controllability:

Task 1: VBench (Huang et al., 2024), a comprehensive video generation benchmark, holistically assesses perceptual quality under joint lighting and camera control.

Task 2: To verify that local relighting does not disrupt temporal stability or content identity, the following are measured: 1) **Temporal Consistency**, calculated as the average cosine similarity of CLIP (Radford et al., 2021) embeddings between consecutive frames, denoted as CLIP-S; 2) **Motion Preservation**, computed as the difference between the optical flow of the generated video and that of the source video, estimated via RAFT (Teed & Deng, 2020); 3) **Chamfer Distance (CD)**; and 4) **Intrinsic Albedo Consistency (IAC)**, defined as the MSE between albedo maps extracted from source and relighted videos using V-RGBX (Fang et al., 2025b).

Task 3: The evaluation focuses on trajectory precision and view consistency: 1) **CLIP-V**: CLIP similarity between corresponding frames of the generated video and the source video is computed to measure semantic fidelity. 2) **Camera Accuracy**: The SOTA camera parameter estimation model MegaSaM is employed to extract camera rotation R and translation T for each frame. Following CamI2V (Zheng et al., 2024), Rotation Distance (RotErr), Translation Error (TransErr), and Camera Motion Consistency (CamMC) are computed. 3) **View Synchronization**: To evaluate view consistency (Bai et al., 2025), the image matching method GIM (Shen et al., 2024) counts matching pixels whose confidence exceeds a predefined threshold (Mat. Pix.).

Camera Metric Definitions. Let (R_i, T_i) denote the commanded camera-to-world pose and $(\tilde{R}_i, \tilde{T}_i)$ the pose estimated from generated frame i. Rotation and translation errors are accumulated over n frames:

$$\text{RotErr} = \sum_{i=1}^{n} \cos^{-1}\left(\frac{\text{tr}(\tilde{R}_i R_i^{\mathsf{T}}) - 1}{2}\right), \tag{14}$$

$$\text{TransErr} = \sum_{i=1}^{n} \|\tilde{T}_i - T_i\|_2, \tag{15}$$

$$\text{CamMC} = \sum_{i=1}^{n} \|[\tilde{R}_i \mid \tilde{T}_i] - [R_i \mid T_i]\|_2. \tag{16}$$

Rotation error is in radians. For matching pixels, videos are normalized to 1024×576 and GIM matches with confidence above 0.95 are counted.

C.4 Task 2 and Task 3 Results

Task 2: Fine-grained Parametric Light Control. Results for task 2 are shown in Table 5. This task focuses on modulating illumination without altering scene content or motion. 1) **Temporal and Motion Stability:** *LumiLens* achieves the highest **CLIP-S** and the lowest **Motion Preservation** error (RAFT). This confirms that our local relighting is temporally consistent and strictly adheres to the original motion fields, unlike RelightVid, which introduces unwanted warping artifacts. 2) **Material and Structural Preservation:** Most notably, our method achieves the lowest **IAC** score and **CD**. High IAC scores in baselines indicate that they "paint" lighting directly onto the texture, corrupting the underlying albedo. Our low IAC score demonstrates that *LumiLens* successfully preserves intrinsic material properties during relighting. Concurrently, the superior CD metric confirms that our method maintains the underlying 3D geometric structure, ensuring physically plausible edits without introducing warping artifacts or baked-in texture corruption.

Table 6: Task 3: Quantitative results for camera control, averaged over synthetic and in-the-wild videos.

Method	CLIP-V↑	RotErr↓	TransErr↓	CamMC↓	Mat.Pix.↑
ReCamMaster	0.920	5.211	1.236	8.891	531.33
EX-4D	0.915	**1.971**	0.295	2.592	852.09
Ours	**0.925**	2.005	**0.281**	**2.410**	**966.28**

Task 3: Camera Control and Reflection Consistency. Table 6 reports average camera control metrics across real and synthetic datasets. *LumiLens* achieves the lowest translation error and best motion consistency, while maintaining a rotation error highly competitive with EX-4D. Table 7 breaks down CamMC by domain and camera-motion range. LumiLens achieves the lowest CamMC in the $60°$–$90°$ range on both synthetic and real videos, supporting its robustness under large viewpoint changes. For the full joint task, prior methods lack our conditioning design and cannot perform it, so the gains reflect architecture and task-oriented training. For camera control alone, Model B, architecturally and methodologically equivalent to EX-4D, still underperforms *LumiLens* when trained on the same data within the same framework (Tables 1 and 2). The gain thus comes from geometry–appearance decoupling rather than training data. Consequently, *LumiLens* secures the highest visual quality and structural fidelity, demonstrating its capability to couple precise camera trajectories with physically plausible, view-consistent appearances. By synthesizing correct view-dependent effects, *LumiLens* ensures consistent feature matching across views through its re-rendering framework.

Table 7: CamMC↓ by domain and camera-motion range. The six domain–motion entries for each method have the mean reported in Table 6.

Domain	Method	0°–30°	30°–60°	60°–90°
Synthetic	ReCamMaster	6.923	8.614	10.272
	EX-4D	**1.882**	2.756	3.342
	Ours	2.149	**2.171**	**2.700**
Real	ReCamMaster	7.486	9.041	11.010
	EX-4D	**2.108**	2.421	3.043
	Ours	2.126	**2.347**	**2.967**

C.5 User Study

To strictly validate perceptual performance, a user study involving 47 participants was conducted, covering four key dimensions (Liu et al., 2025a): Relighting Quality (RQ), which measures fidelity to target illumination and prompt alignment; Video Smoothness (VS); Identity Preservation (IP), ensuring consistency of subject identity and intrinsic albedo; and 4D Consistency (4DC), evaluating spatio-temporal coherence across novel viewpoints. To ensure robust testing, 15 in-the-wild scenes were randomly sampled, and randomized configurations for illumination conditions and camera trajectories were utilized. Results are reported in Table 8.

The interface in Fig. 22 presents the source video, target conditions, and anonymized candidate videos in randomized order. Participants compare relighting quality, video smoothness, identity preservation, and 4D consistency, with a "Hard to judge" option for ambiguous cases.

Table 8: User Study: User preference indicates the percentage of participants who selected our method against baselines. RelightVid performs only relighting.

Method	RQ↑	VS↑	IP↑	4DC↑
Light-X(Nano-Ref)	98.0	97.5	98.1	97.7
V-RGBX(Nano-Ref)+EX-4D	99.7	99.6	100.0	99.1
RelightVid	98.8	98.3	97.9	-

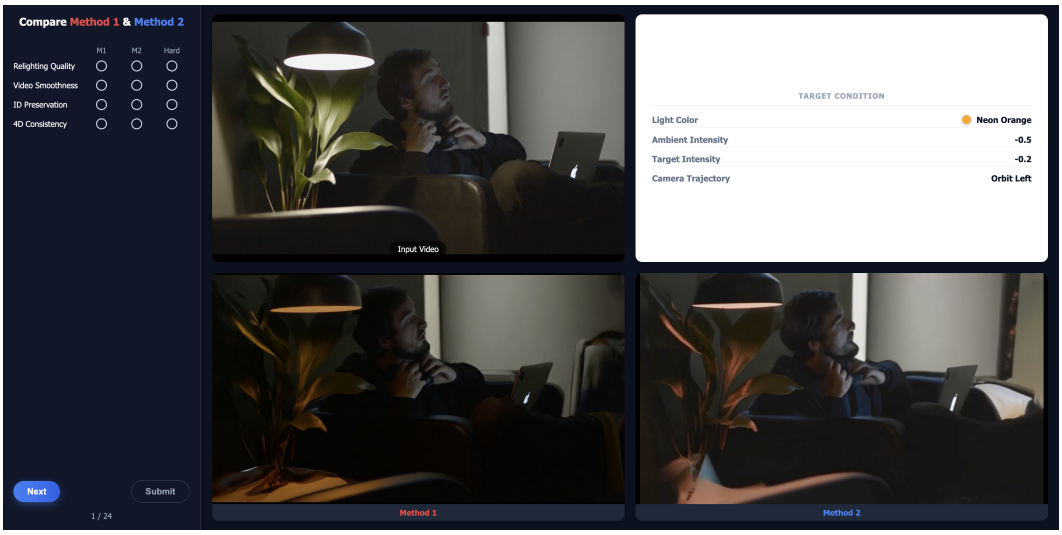


Figure 22: User-study interface with source video, target conditions, and paired candidate results.

D Control Settings and Camera Trajectories

Figures 23–25 list the per-case control settings, including emitter masks, RGB colors, relative local/ambient intensities, and camera trajectories. The case IDs printed in these sheets follow the figure numbering of this paper. An "X" denotes a setting already annotated in the referenced figure, rather than a missing or zero-valued control. Positive intensity values enhance lighting and negative values attenuate it.

D.1 Camera Trajectory Parameterization

Camera controls use a tuple (Mode, Direction, Amplitude), or (Mode, Amplitude) for direction-independent motion. Direction is generally a sign (1 or -1), indicating orientation or initial lateral sway. Amplitude controls the motion's extent, such as orbital radius or dolly distance. The modes used in the visual results are:

Orbit: Circular motion around the scene, with amplitude specifying radius and direction selecting clockwise/counter-clockwise rotation.

Orbit-Ascend: An ascending crane-like shot; direction sets the initial lateral sway and amplitude controls elevation and oscillation.

Zoom-out: Backward translation along the optical axis; amplitude specifies distance, with no direction parameter.

Helix-In: Forward translation combined with a spiral; amplitude controls advancing distance and spiral radius.

S-Curve-Dolly: Forward motion with lateral oscillation; direction sets the initial swing and amplitude its extent, emphasizing parallax.

Parabola-Dive: A sweeping trajectory with amplitude determining depth/span and direction selecting lateral shift. Moderate amplitudes help preserve geometry under large perspective changes.

Rotate: In-place panning; here Direction specifies an angle (e.g., 30°), and Amplitude provides a baseline radius.

These templates are a user interface for generating per-frame extrinsics. LumiLens also accepts arbitrary trajectories supplied directly as JSON camera parameters.

D.2 Per-case Control Settings

Case ID	Target Mask	Target Color	Light Intensity	Ambient Intensity	Camera Trajectory
TeaserFig (A)		255, 255, 255	0.0	X	Orbit, 1, 0.2
TeaserFig (B)		255, 173, 48	-0.46	-0.08	Zoom-out, 0.5
TeaserFig (C)		X	0.51, 0.36, 0.87, 0.66, 0.19, 0.71	-0.25, -0.13, 0.22, 0.35, -0.69, 0.47	Orbit, 1, 0.3
TeaserFig (D)		240, 125, 50 160, 90, 185	0.0, 0.0	-0.17	S_Curve_Dolly, 1, 0.4
TeaserFig (E)		175, 195, 215	-0.49	-0.36	S_Curve_Dolly, -1, 0.3
Fig8 (row 1)		148, 173, 188	0.51	0.19	Helix_In, 1, 0.5
Fig8 (row 2)		237, 148, 62	0.25	-0.27	Zoom-out, 0.3
Fig8 (row 3)		220, 150, 60	-0.28	-0.23	Zoom-out, 0.5
Fig8 (row 4)		210, 170, 130	0.47	-0.79	Rotate, 40, 0.1
Fig8 (row 5)		232, 207, 229	0.45	0.56	Zoom-out, 0.5
Fig8 (row 6)		238, 212, 154	-0.14	-0.21	S_Curve_Dolly, 1, 0.4
Fig8 (row 7)		101, 218, 104	0.11	-0.35	Orbit, 1, 0.2

Figure 23: Control settings for the teaser (Fig. 1) and the first result gallery (Fig. 8).

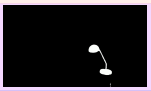
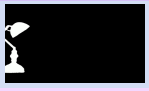
Case ID	Target Mask	Target Color	Light Intensity	Ambient Intensity	Camera Trajectory
Fig 5		X	X	X	Orbit_Ascend, 1, 0.5
Fig 7 (case 5)		X	X	X	Zoom-out, 0.5
Fig 7		X	X	X	Rotate, X, 0.1
Fig9 (row 1)		247, 158, 62	0.49	-0.06	Helix_In, 1, 0.4
Fig9 (row 2)		239, 114, 83	0.24	-0.11	Parabola_Dive, -1, 0.2
Fig9 (row 3)		235, 178, 109	-0.16	-0.29	Helix_In, -1, 0.4
Fig9 (row 4)		244, 128, 206	0.35	0.07	Parabola_Dive, 1, 0.3
Fig9 (row 5)		163, 215, 242	0.46	-0.13	Rotate, -30, 0.1
Fig9 (row 6)		188, 231, 214	0.28	-0.49	Parabola_Dive, 1, 0.2
Fig9 (row 7)		253, 148, 55	0.65	-0.36	S_Curve_Dolly, -1, 0.3
Fig9 (row 8)		248, 211, 89	0.12	0.16	S_Curve_Dolly, 1, 0.5
Fig9 (row 9)		249, 231, 164	0.32	0.85	Orbit, -1, 0.2
Fig9 (row 10)		142, 179, 204	-0.81	0.06	Orbit, 1, 0.2

Figure 24: Control settings for light on/off (Fig. 5), parametric control (Fig. 7), and the second result gallery (Fig. 9).

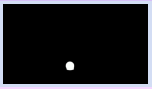
Case ID	Target Mask	Target Color	Light Intensity	Ambient Intensity	Camera Trajectory
Fig14 case2		196, 123, 242	0.21	-0.23	S_Curve_Dolly, 1, 0.3
Fig17 case2		241, 167, 204	-0.46	-0.08	Orbit, 1, 0.2
Fig13		171, 242, 115	0.26	-0.76	S_Curve_Dolly, 1, 0.3
Fig11 (case 1)		218, 145, 237	0.19	-0.26	Rotate, -30, 0.1
Fig11 (case 2)		242, 163, 67	0.42	-0.11	Zoom-out, 0.6
Fig10		247, 228, 201	-0.14	-0.25	Helix_In, 1, 0.4
Fig12 (row 1)		137, 211, 163	-0.15	-0.17	Parabola_Dive, 1, 0.3
Fig12 (row 3)		243, 246, 241	0.39	-0.03	Parabola_Dive, 1, 0.3
Fig12 (row 4)		248, 236, 211	0.27	0.09	Zoom-out, 0.5
Fig12 (row 5)		252, 211, 114	0.05	-0.16	Orbit_Ascend, 1, 0.4
Fig8 (row 8)		211, 197, 183	0.32	0.46	S_Curve_Dolly, 1, 0.3
Fig8 (row 9)		220, 160, 100	0.62	0.12	Rotate, 30, 0.1
Fig8 (row 10)		133, 154, 171	-0.88	-0.26	Orbit_Ascend, 1, 0.5

Figure 25: Control settings for the remaining cases of the first result gallery (Fig. 8), noisy depth (Fig. 10), occlusion (Fig. 11), non-Lambertian surfaces (Fig. 12), Model A (Fig. 13), Model B (Fig. 14), and baseline comparisons (Fig. 17).