
IllumiCraft: Unified Geometry and Illumination Diffusion for Controllable Video Generation (Supplementary Material)

Anonymous Author(s)

Affiliation

Address

email

1 A Overview

2 In this supplementary material, we provide the following content:

- 3 (a) Details of HDR environment map transformation in Section B.
- 4 (b) Details of the presented lighting encoder in Section C.
- 5 (c) Additional experimental results in Section D.
- 6 (d) Additional visualization results in Section E.

7 B HDR Map Transformation

8 Let \mathcal{I}_c denote the first (base) video frame, from which we extract a reference chrome ball image using
9 DiffusionLight [1]. Then we relabel \mathcal{I}_c as \mathcal{I}_0 and denote each subsequent frame by \mathcal{I}_t for $t \in [1, T]$,
10 where T is the total number of frames (49 in our experiments). We use Video-Depth-Anything [2] to
11 provide per-frame depth maps $D_t(u, v)$. Our goal is to synthesize the appearance of the chrome ball
12 in each \mathcal{I}_t by estimating the camera’s 3D motion and warping \mathcal{I}_0 accordingly.

13 B.1 Sparse Feature Tracking

14 From the base frame \mathcal{I}_0 , we detect up to $N = 200$ reliable 2D corners $\mathbf{p}_i^0 = (u_i^0, v_i^0)$, $t \in [1, N]$
15 using the OpenCV Shi-Tomasi detector and track them into the current frame \mathcal{I}_t using the OpenCV
16 Lucas-Kanade optical flow, which produces $\mathbf{p}_i^t = (u_i^t, v_i^t)$; points with failed tracks or missing depth
17 are discarded, leaving a robust set of correspondences for motion estimation.

18 B.2 3D Motion Estimation via Constrained Affine Fit

19 For each correspondence $(\mathbf{p}_i^0, \mathbf{p}_i^t)$, we first read the depths $z_i^0 = D_0(u_i^0, v_i^0)$ and $z_i^t = D_t(u_i^t, v_i^t)$
20 from the base and current depth maps, then lift them to the 3D points of normalized camera by $\mathbf{X}_i^0 =$
21 $z_i^0 K^{-1}[u_i^0, v_i^0, 1]^\top$ and $\mathbf{X}_i^t = z_i^t K^{-1}[u_i^t, v_i^t, 1]^\top$, where K is the intrinsic matrix. We estimate
22 the rigid affine transform (R, t) by minimizing $\sum_i \|R \mathbf{X}_i^0 + t - \mathbf{X}_i^t\|^2$ subject to $R^\top R = I$ and
23 $\det R = 1$ using the OpenCV `estimateAffine3D`. To enforce temporal smoothness, we dampen the
24 raw 3x4 transform $M_{3d} = [R \mid t]$ toward the 3x4 identity affine matrix \hat{I} via $M_{3d} \leftarrow \hat{I} + \alpha (M_{3d} - \hat{I})$
25 with $\alpha = 0.05$, and finally re-orthogonalize R via SVD while clamping both rotation angle and
26 translation magnitude.

“Turquoise waves crash basalt rocks, *dull stormy light*” “A woman gazes at a lit candle, *low-key candlelight*”

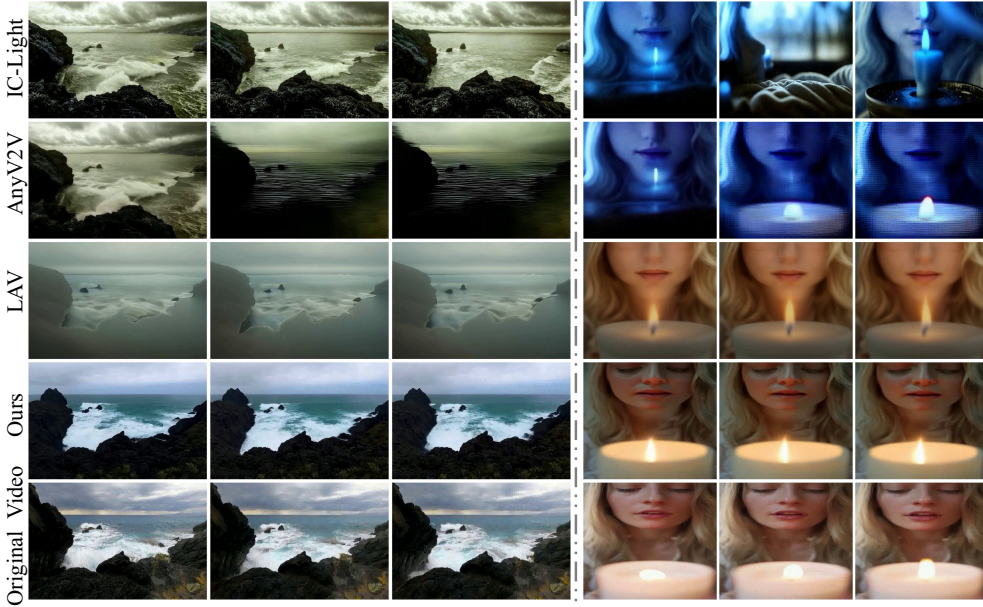


Figure 1: **Visual results under the text-conditioned setting.** We compare IC-Light [3], AnyV2V [4], Light-A-Video [5] (abbreviated LAV in the figure), and our proposed method, IllumiCraft.

27 B.3 Warping the Reference Chrome Ball

28 For each pixel (x_0, y_0) in the reference chrome ball image \mathcal{I}_0 of size (w, h) , we compute the
 29 mean depth $d_{\text{avg}} = \frac{1}{WH} \sum_{u=0}^{W-1} \sum_{v=0}^{H-1} D^0(u, v)$, map (x_0, y_0) into video-frame coordinates by
 30 $x_v = \frac{W}{w}x_0$ and $y_v = \frac{H}{h}y_0$, lift it to 3D via $\mathbf{X}_0 = d_{\text{avg}}K^{-1}[x_v, y_v, 1]^\top$, apply the affine transform
 31 $\mathbf{X}'_0 = R\mathbf{X}_0 + t$, project back via $[u', v', 1]^\top \propto K\mathbf{X}'_0$, recover warped coordinates $x'_0 = \frac{u'}{W/w}$ and
 32 $y'_0 = \frac{v'}{H/h}$, and finally use the OpenCV `remap` to sample \mathcal{I}_0 at (x'_0, y'_0) , producing the warped chrome
 33 ball in the current frame.

34 B.4 Discussion

35 Our method tracks a few key points on the chrome ball in 3D using depth maps to directly recover
 36 camera motion, avoiding the pitfalls of 2D model fitting. We reduce jitter and ensure smooth results
 37 by gently smoothing each new motion estimate and capping its maximum change. Since the chrome
 38 ball is nearly spherical, using its average depth introduces only negligible error. This fully automatic
 39 pipeline produces accurate, stable warps with minimal manual intervention.

40 **Collected Lighting Prompts.** To generate truly diverse relit videos, we curated 100 unique lighting
 41 prompts (see Table 5 and Table 6). These prompts span both everyday and fantastical illumination
 42 scenarios, ranging from simple indoor scenes to dramatic, otherworldly effects. This variety ensures
 43 that our model delivers outputs that are both quantitatively robust and qualitatively rich.

44 C Lighting Encoder

45 The lighting encoder first reshapes the input HDR video tensor $\mathcal{V}_{\text{hdr}} \in \mathbb{R}^{T \times 32 \times 32 \times 3}$ (with $T = 49$)
 46 into $X_1 \in \mathbb{R}^{49 \times 3072}$, then applies a four-layer MLP, each Linear layer followed by LeakyReLU,
 47 with dimensions $3072 \rightarrow 4096 \rightarrow 4096 \rightarrow 4096 \rightarrow 4608$ (where $4608 = 3 \times 1536$, corresponding
 48 to 3 illumination tokens) to produce $Y_1 \in \mathbb{R}^{49 \times 4608}$. This is reshaped and permuted into $Y_2 \in$
 49 $\mathbb{R}^{3 \times 49 \times 1536}$, processed by a single-layer TransformerEncoder ($d_{\text{model}} = 1536, n_{\text{head}} = 8, \text{dim}_{\text{ff}} =$
 50 2048) yielding $Y_3 \in \mathbb{R}^{3 \times 49 \times 1536}$, and then passed through a depth-wise Conv1d followed by
 51 LeakyReLU and squeeze, to collapse the temporal axis into the final output $Z \in \mathbb{R}^{3 \times 1536}$.

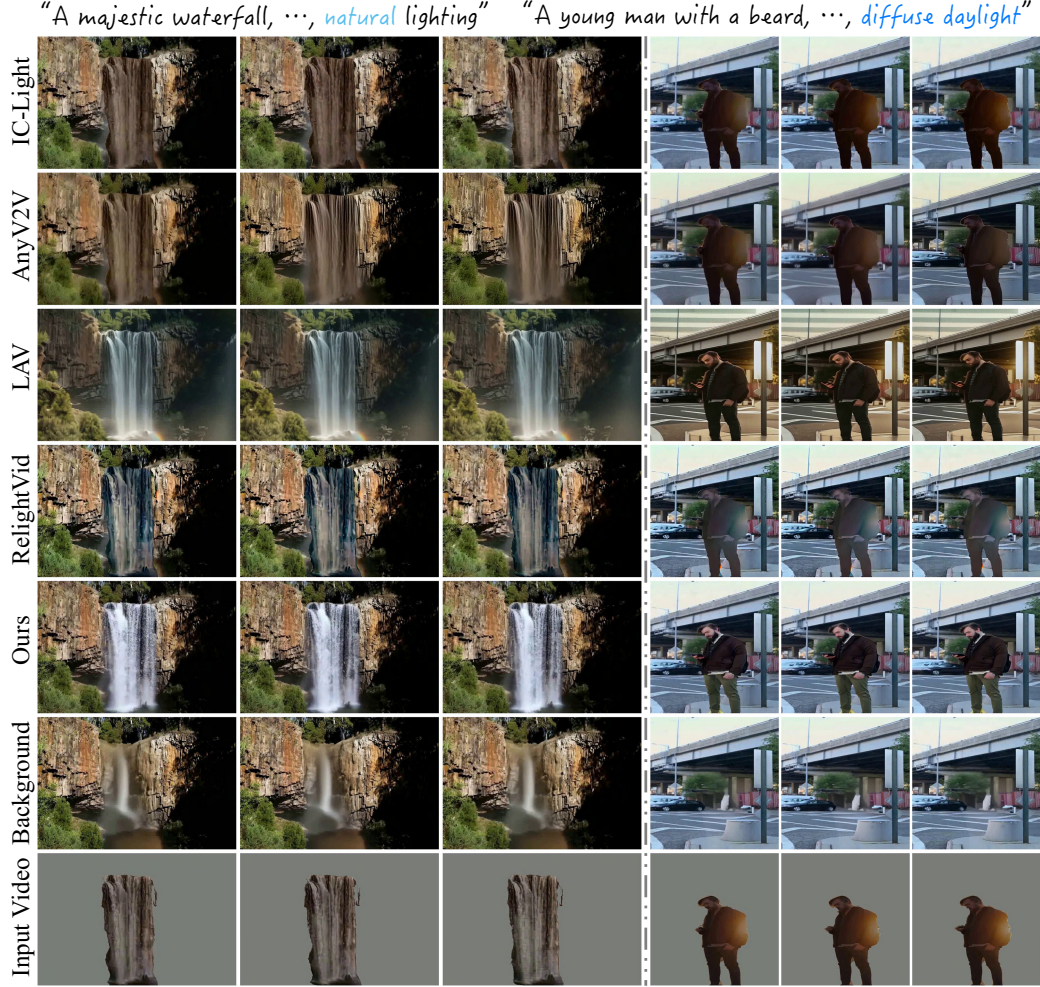


Figure 2: **Visual results under the background-conditioned setting.** We compare IC-Light [3], AnyV2V [4], Light-A-Video [5], RelightVid [5] and our proposed method, **IllumiCraft**.

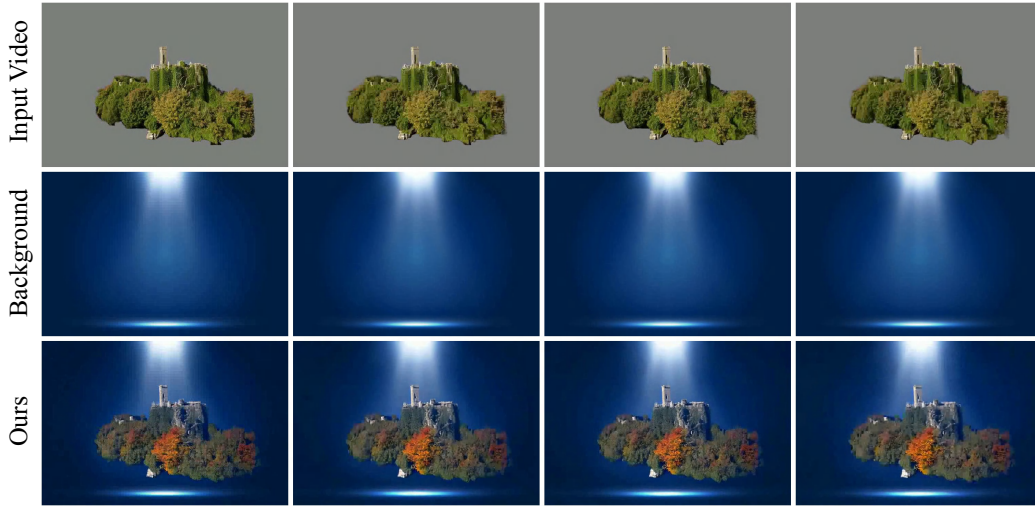
52 D Additional Experimental Results

53 D.1 Comparison with Existing Methods

54 **Text-Conditioned Video Relighting.** Figure 1 qualitatively benchmarks four video relighting
 55 methods on two different illumination conditions, dull stormy light and a low-key candle light,
 56 showing side-by-side comparisons of IC-Light [3], AnyV2V [4], Light-A-Video [6] and **IllumiCraft**.
 57 In the coastal scene, LAV yields overly smooth, desaturated outputs, AnyV2V introduces temporal
 58 jitters and erratic color shifts. IC-Light also causes color shifts and cannot preserve the fine details in
 59 the original video frames. In contrast, **IllumiCraft** preserves the original structure, faithfully renders
 60 prompt-specific cues (e.g., turquoise waves, intimate candlelight glow), and maintains temporal
 61 stability without artifacts. This demonstrates superior fidelity and consistency over all baselines.

62 **Background-Conditioned Video Relighting.** As shown in Figure 2, we compare two relighting
 63 scenarios, a majestic waterfall under natural lighting (left) and a bearded man under diffuse daylight
 64 (right), over four baselines (IC Light [3], AnyV2V [4], Light-A-Video [6] and RelightVid [5]) and
 65 our method. RelightVid introduces banding and creates unnatural illumination on the waterfall.
 66 IC Light and AnyV2V preserve the overall brightness, but blur fine details such as droplets, hair,
 67 and clothing. Light-A-Video desaturates tones, oversmooths the water spray, and alters the portrait
 68 background, causing artifacts. In contrast, our method follows each prompt precisely, achieving a

"An ancient stone castle on an island, ..., *bright chalky spotlight in misty blue haze*"



"branch of an apple tree in full bloom, ..., *towering LED floodlight, stark snowy illumination*"

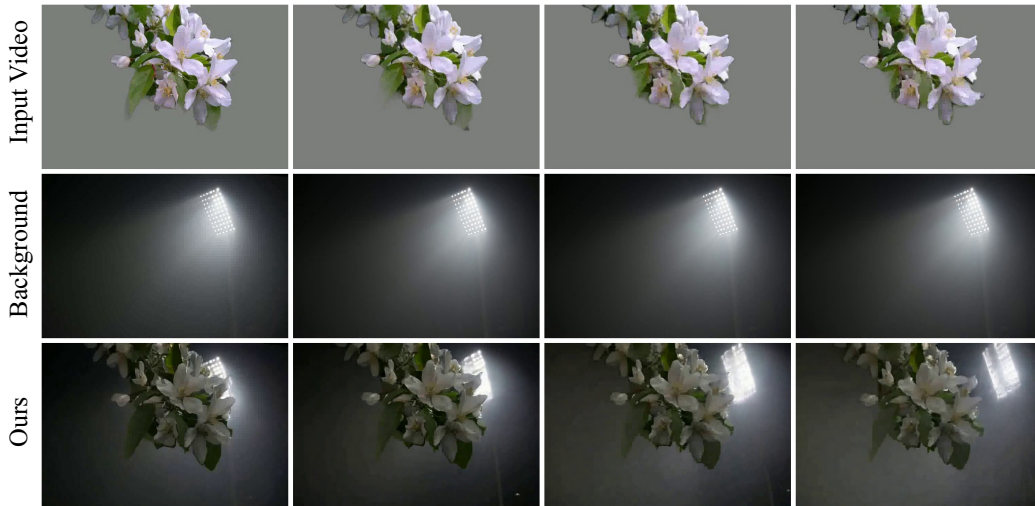


Figure 3: **Failure cases of IllumiCraft.** We show the failure cases generated by our method.

Table 1: Impact of dropping 3D tracking videos (text-only).

Possibility	FVD (\downarrow)	TA (\uparrow)	TC (\uparrow)
10%	2285.32	0.3303	0.9893
20%	2251.21	0.3332	0.9939
30%	2186.40	0.3342	0.9948
40%	2234.35	0.3325	0.9915

Table 2: Effect of dropping 3D tracking videos (background).

Possibility	FVD (\downarrow)	TA (\uparrow)	TC (\uparrow)
10%	1154.32	0.3231	0.9902
20%	1102.21	0.3273	0.9935
30%	1072.38	0.3292	0.9945
40%	1098.35	0.3278	0.9937

69 high-fidelity waterfall and sharp rock edges with rock-solid frame-to-frame consistency, enhancing
70 detail preservation and temporal coherence in both scenarios.

71 D.2 Ablation Study

72 **Impact of Dropping 3D Tracking Videos.** We evaluate the impact of randomly dropping 3D
73 tracking videos during training under both text-conditioned (Table 1) and background-conditioned
74 (Table 2) settings. A 30% drop rate offers the best balance across visual quality, text alignment, and

Table 3: Effect of dropping the reference image (text-only).

Possibility	FVD (\downarrow)	TA (\uparrow)	TC (\uparrow)
5%	2232.46	0.3331	0.9939
10%	2186.40	0.3342	0.9948
20%	2175.23	0.3341	0.9943
30%	2158.32	0.3338	0.9941

Table 4: Effect of dropping the reference image (background).

Possibility	FVD (\downarrow)	TA (\uparrow)	TC (\uparrow)
5%	1065.83	0.3284	0.9941
10%	1072.38	0.3292	0.9945
20%	1105.28	0.3275	0.9928
30%	1127.32	0.3269	0.9925

frame consistency. In the text-conditioned scenario, it reduces FVD, increases the text alignment score to 0.3342, and improves the temporal coherence score to 0.9948. In the background-conditioned setting, the same drop rate also lowers FVD and boosts text alignment to 0.3292. These results suggest that a 30% drop rate consistently yields optimal performance across all key metrics.

Effect of Dropping Reference Image. As shown in Tables 3 and 4, a 10% possibility of dropping reference images during training achieves the best overall trade-off in both text-only and background-only settings. In the text-only condition, while the 10% drop rate does not result in the lowest FVD, it delivers the highest text alignment (0.334) and the highest temporal coherence (0.995), making it the most balanced choice. In the background-only scenario, the same 10% drop rate lowers FVD to 1072.38, raises text alignment to 0.3292, and maintains coherence at 0.9945. Overall, a 10% drop rate maximizes visual realism, alignment, and frame consistency across both settings.

D.3 Failure Cases

As shown in Figure 3, in the top example (an ancient stone castle illuminated by a bright chalky spotlight in misty blue haze), our relighting sometimes shifts and misaligns the lower foliage, resulting in oversaturated greens. In the bottom example (a flowering apple branch moving in front of a towering LED floodlight), when the branch crosses the illuminated region, parts of the floodlight are occluded and mistakenly treated as foreground, causing unwanted changes in the floodlight’s appearance. To address these issues, we plan to expand our curated dataset to include more scenes with dynamic occlusions and strong directional lighting.

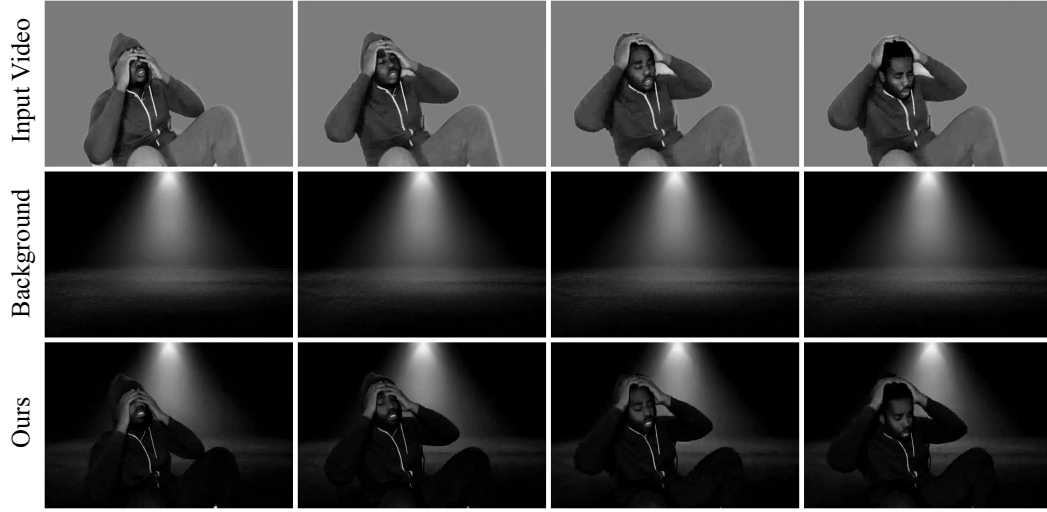
E Additional Visualization Results

Figure 4, 5, 6, 7, 8 and 9 show additional relit videos generated by IllumiCraft. Across a wide range of custom backgrounds, including varied spotlight arrangements, colored backdrops, and complex scene contents, IllumiCraft adapts seamlessly to diverse illumination scenarios, producing smooth shading transitions, consistent specular highlights, and accurate shadows. These examples demonstrate its versatility in handling challenging lighting conditions while faithfully preserving original background details, resulting in high-quality relit videos without visible artifacts.

References

- [1] Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. DiffusionLight: Light probes for free by painting a chrome ball. In *CVPR*, 2024. 1
- [2] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025. 1
- [3] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. IC-Light GitHub Page, 2024. 2, 3
- [4] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhui Chen. AnyV2V: A tuning-free framework for any video-to-video editing tasks. *arXiv:2403.14468*, 2024. 2, 3
- [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. *arXiv:2501.16330*, 2025. 2, 3

"A hooded man in *hard-edged luminous spotlight* over a dark floor, *suspenseful stage mood*"



"An elephant with textured skin, ..., *razor-sharp blue shafts*, *futuristic*, ..."

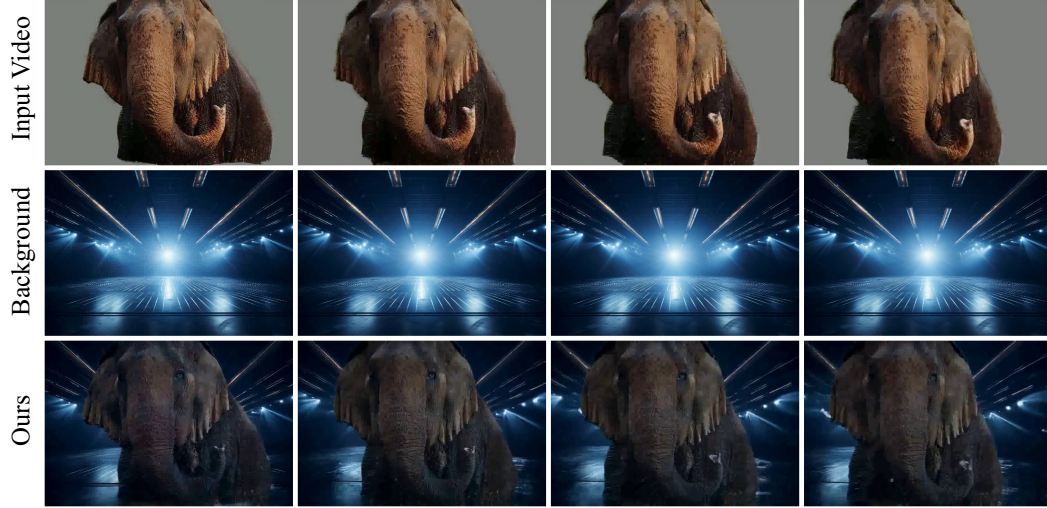


Figure 4: **Visual results of IllumiCraft.** Our method produces high-fidelity, prompt-aligned videos that adapt to diverse lighting conditions, including dramatic spotlight effects.

- 114 [6] Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi
 115 Dong, Yuhang Zang, Yuhang Cao, et al. Light-a-video: Training-free video relighting via
 116 progressive light fusion. *arXiv:2502.08590*, 2025. 3

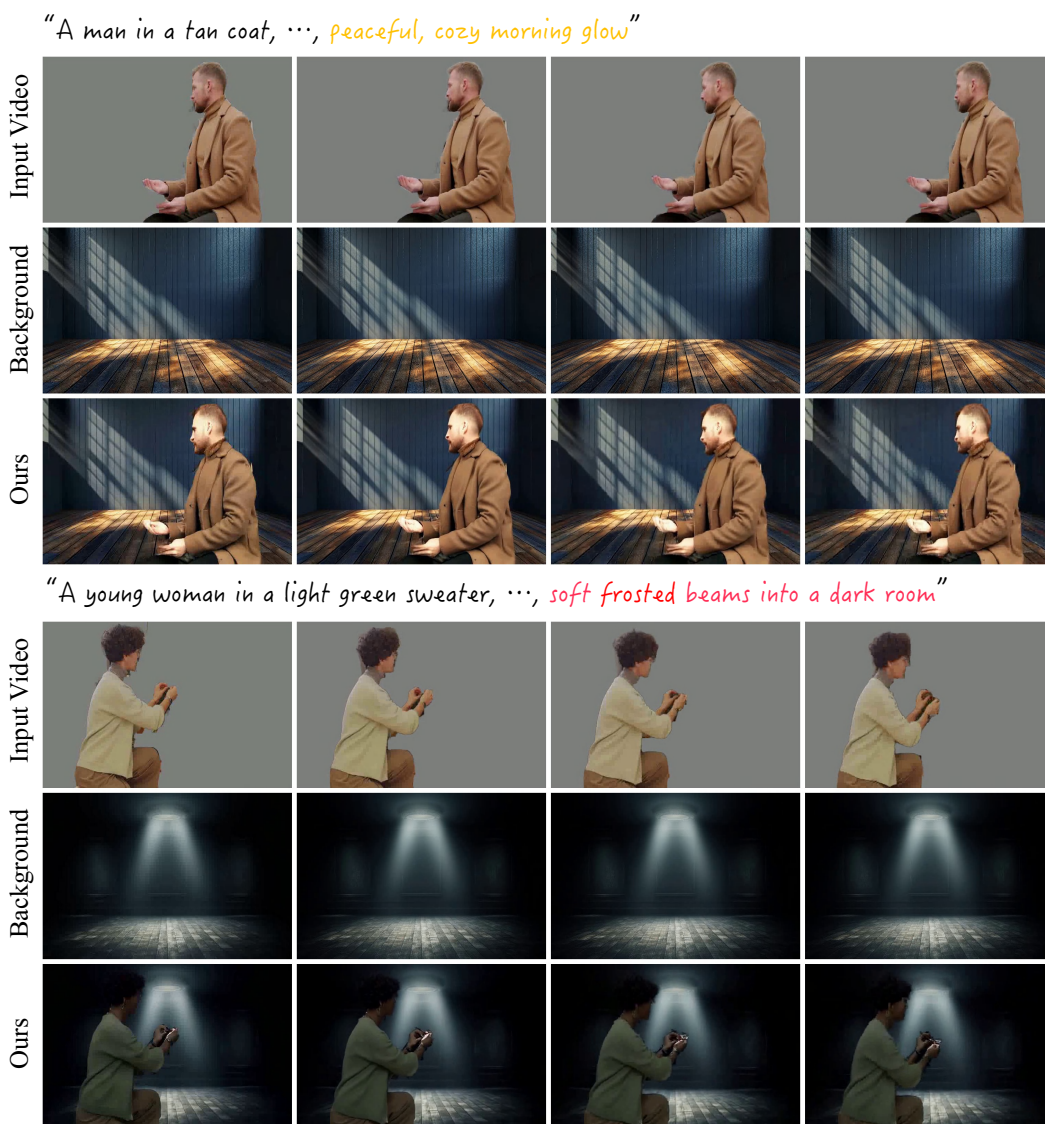


Figure 5: **Visual results of IllumiCraft.** Our method produces high-fidelity, prompt-aligned videos that adapt to diverse lighting conditions, including dramatic spotlight effects.

"A glowing, translucent sphere, ..., crisp radiant overhead, a solemn, cinematic stage vibe"



"A traditional Russian church, ..., warm-white wall sconces, a cozy, minimalist gallery feel"

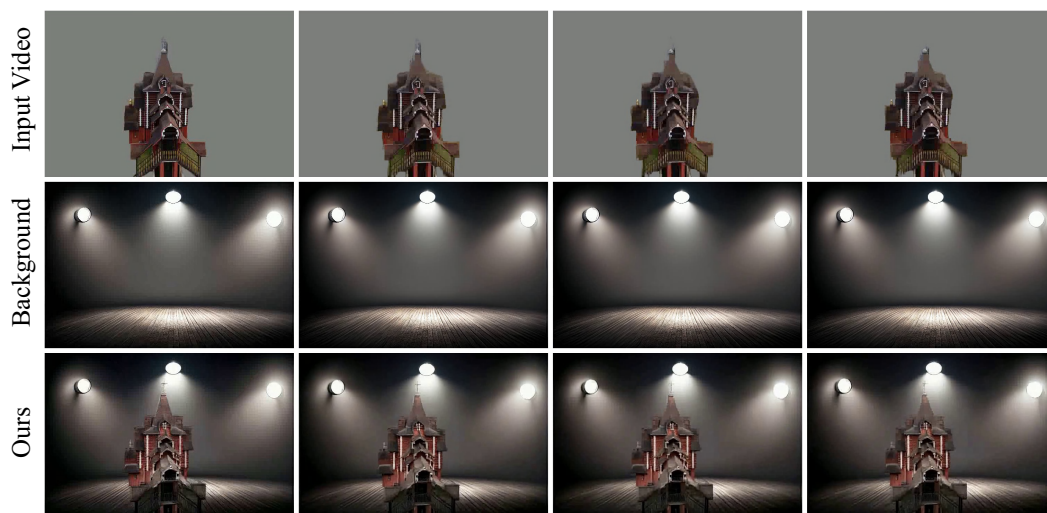
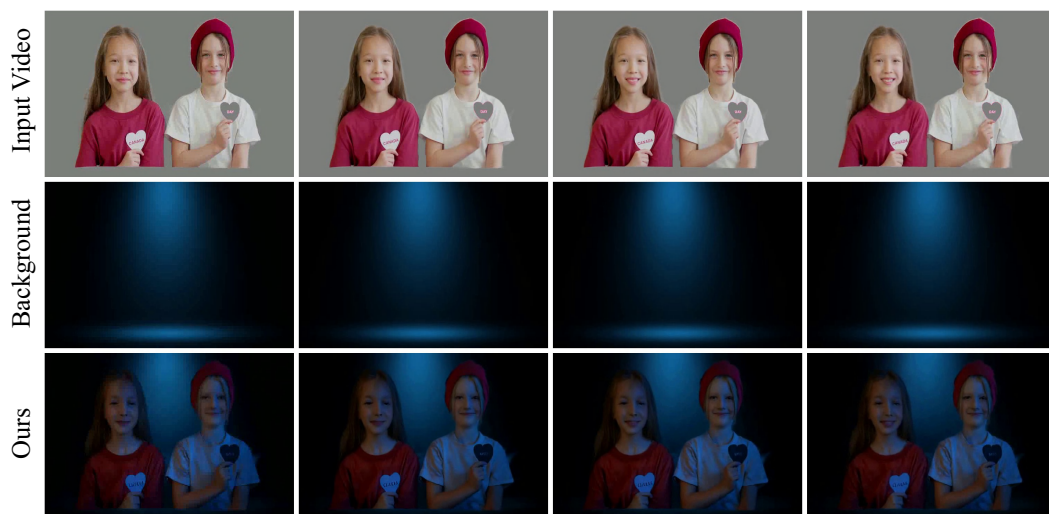


Figure 6: **Visual results of IllumiCraft.** Our method produces high-fidelity, prompt-aligned videos that adapt to diverse lighting conditions, including dramatic spotlight effects.

"Two cheerful young girls, ..., broad cool-teal wash, a tranquil, moody spotlight effect"



"A two-tiered round wooden table, ..., slender white spotlights, an eerie, suspenseful scene"

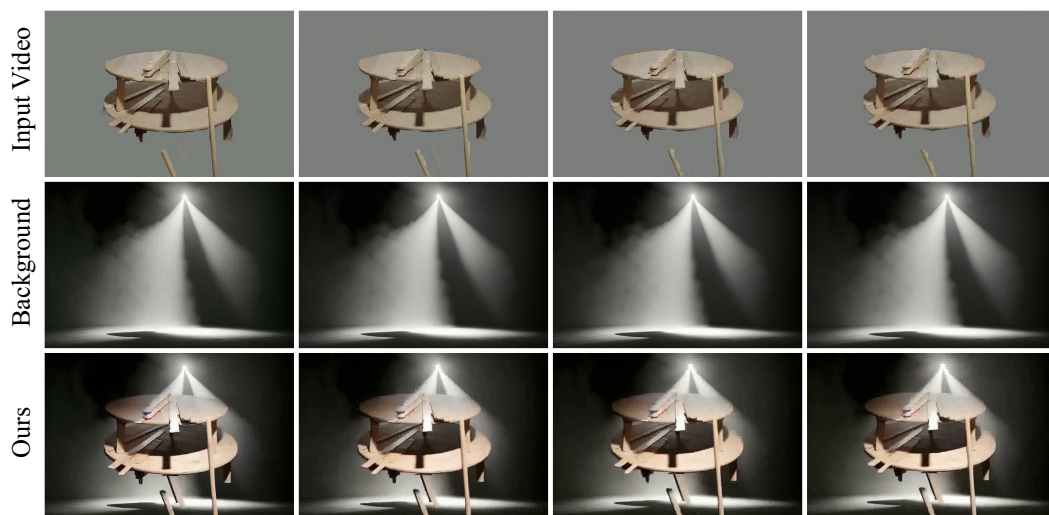
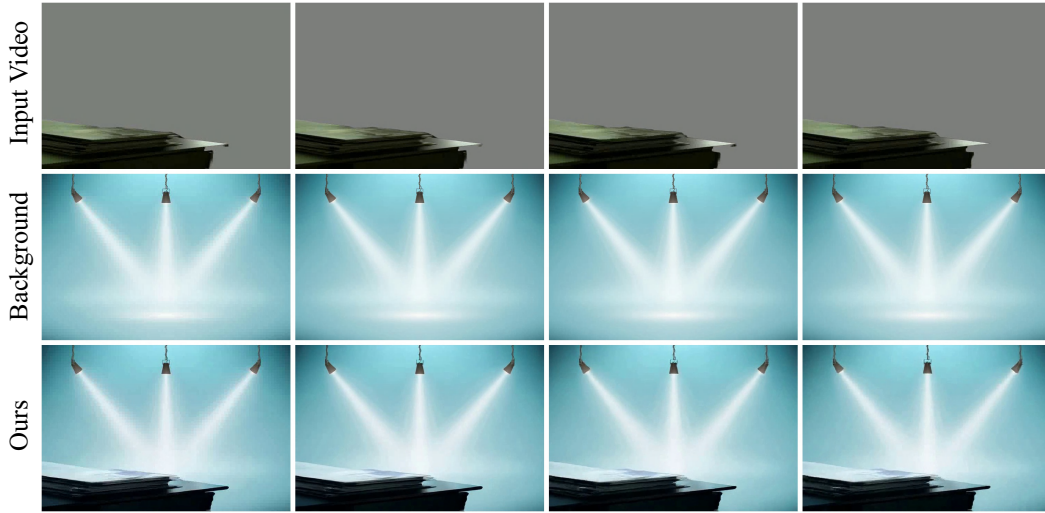


Figure 7: **Visual results of IllumiCraft.** Our method produces high-fidelity, prompt-aligned videos that adapt to diverse lighting conditions, including dramatic spotlight effects.

"A dark wooden desk, ..., *soft-edged white beams overlapping on a cyan backdrop*"



"A stationary airplane wing, ..., *crisp bright spotlights, minimal, neutral-gallery atmosphere*"

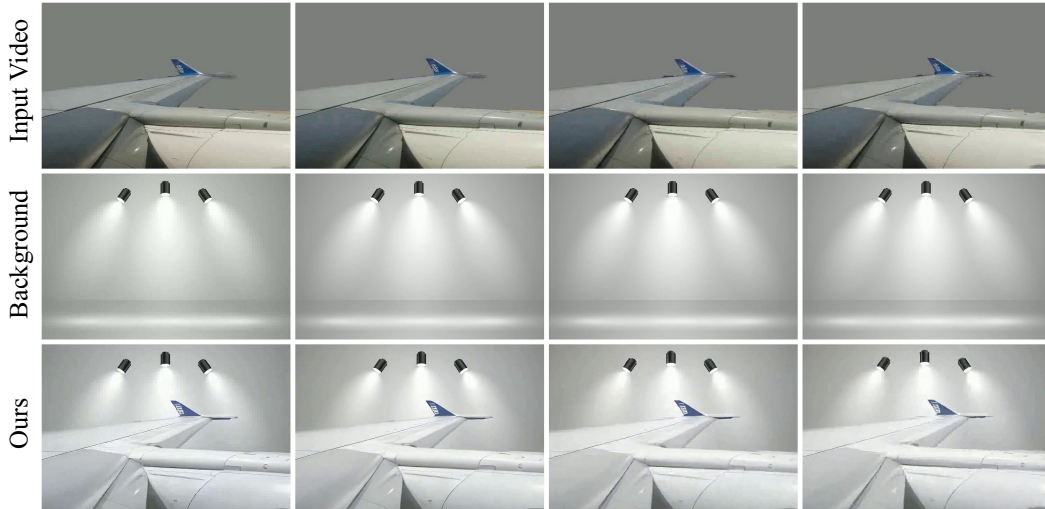


Figure 8: **Visual results of IllumiCraft.** Our method produces high-fidelity, prompt-aligned videos that adapt to diverse lighting conditions, including dramatic spotlight effects.

"A man in a red jacket and blue beanie, ..., *clean radiant pendants, minimalist gallery*"



"Snow-covered pine trees, ..., *towering LED floodlight, stark snowy illumination*"



Figure 9: **Visual results of IllumiCraft.** Our method produces high-fidelity, prompt-aligned videos that adapt to diverse lighting conditions, including dramatic spotlight effects.

Table 5: Our collected lighting prompts (1-50) for relit videos.

#	Lighting Prompt
1	red and blue neon light
2	sunset over sea
3	sunlight through the blinds
4	in the forest, magic golden lit
5	backlighting
6	sunset
7	sunshine, hard light
8	dappled light
9	magic lit, sci-fi RGB glowing, key lighting
10	neon light
11	magic golden lit
12	shadow from window, sunshine
13	sunlight through the blinds
14	neon light
15	cozy bedroom illumination
16	natural lighting
17	soft lighting
18	candle light
19	pink neon light
20	sunlit
21	warm sunshine
22	warm yellow and purple neon lights
23	neon, Wong Kar-wai, warm
24	cyberpunk style and light
25	yellow and purple neon lights
26	sunshine from window
27	moon light
28	soft sunshine
29	dark shadowy light
30	neo punk, city night
31	cyberpunk
32	golden hour light
33	blue hour lighting
34	tungsten light
35	fluorescent office lighting
36	street light at night
37	studio spotlight
38	rim light on subject
39	bokeh city light at night
40	TV screen glow in dark room
41	modern minimalistic LED glow
42	ambient underlit glow
43	soft dusk lighting
44	harsh industrial light
45	warm ambient room light
46	icy blue fluorescent glow
47	mystical twilight shimmer
48	low-key candlelight
49	rainy city neon
50	glowing backlight

Table 6: Our collected lighting prompts (51-100) for relit videos.

#	Lighting Prompt
51	strong lighting
52	rustic lantern light
53	overcast day glow
54	golden twilight shimmer
55	dull stormy light
56	subtle overhead illumination
57	steely warehouse lighting
58	vintage street lamp
59	cool-blue spotlights
60	glowing river reflections
61	sunburst window light
62	morning haze glow
63	afterglow silhouette
64	dawn light shadows
65	broken neon flicker
66	sleek futuristic luminescence
67	soft pastel glow
68	urban dusk illumination
69	underwater blue light
70	interior design spotlight
71	backlit street sign
72	glowing neon arches
73	morning sunbeam
74	rustling leaves with sun rays
75	subdued candlelit ambiance
76	serene twilight light
77	prismatic light effects
78	diffuse daylight
79	geometric LED array
80	rain-soaked neon reflections
81	subterranean light glow
82	bright white spotlights
83	dazzling sun flare
84	vibrant festival lights
85	enchanted aurora borealis
86	subtle office glow
87	twinkling fairy lights
88	chrome and neon reflections
89	fiery red spotlight
90	icy neon glow
91	sun-dappled forest light
92	electric dreamscape glow
93	irradiated room ambiance
94	scattered light beams
95	colored spectrum radiance
96	mystic foggy illumination
97	urban jungle glow
98	warm campfire light
99	bioluminescent glow
100	caustic rippling light