

Appendix of UniLumos

Anonymous Author(s)

Affiliation

Address

email

In this appendix, we provide additional details to complement the main paper. First, **Sec. A** shows detailed ablation results to evaluate the contribution of each component in our framework. Then, we explain the motivation behind introducing physics-guided feedback in **Sec. B**. Next, we present the detailed pipeline of our proposed *LumosData* relighting data construction process in **Sec. C**. We describe the training implements of our model, including the loss functions and their interactions, in **Sec. D**. **Sec. E** provides additional qualitative results to further illustrate the effectiveness of UniLumos. Finally, **Sec. F**, **Sec. G**, and **Sec. H** discuss the limitations of our method, its broader societal impact, and the safeguards we implement to ensure responsible use of generative relighting models.

A Additional Results and Ablative Study

To further validate the effectiveness of UniLumos, we present additional quantitative results and conduct a series of ablative experiments. As summarized in Tab. 1, we divide our analysis into four sections: baseline comparison, ablation of physics-guided components, the effect of using geometry as input, and training data domain analysis.

Table 1: Quantitative comparison. **Bold** number indicate the best performance.

Model	(a) Quality			(b) Temporal Consistency	(c) Lumos Consistency	
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	R-Motion \downarrow	Avg. Score \uparrow	Dense L2 Error \downarrow
1) Comparison of baseline methods						
IC-Light Pre Frame	20.132	0.851	0.133	2.437	0.672	0.432
Light-A-Video (IC-Light + CogVideoX 2B)	19.851	0.859	0.124	1.784	0.641	0.383
Light-A-Video (IC-Light + Wan2.1 T2V 1.3B)	20.784	0.876	0.129	1.582	0.682	0.371
UniLumos (our)	25.031	0.891	0.109	1.436	0.871	0.147
2) Ablative Study						
UniLumos w/o Depth Feedback	23.472	0.883	0.118	1.443	0.870	0.265
UniLumos w/o Normal Feedback	22.115	0.874	0.123	1.446	0.863	0.173
UniLumos w/o All Feedback	21.433	0.862	0.139	1.473	0.859	0.297
UniLumos w/o Path Consistency	25.317	0.902	0.113	1.438	0.875	0.153
3) Feedback as Inputs						
UniLumos w/ Depth*	23.041	0.868	0.114	1.451	0.862	0.179
UniLumos w/ Normal*	25.273	0.876	0.119	1.447	0.869	0.283
UniLumos w/ All*	26.012	0.882	0.104	1.440	0.867	0.153
4) Effect of Training Domain						
Only Video	22.487	0.863	0.119	1.487	0.857	0.173
Only Image	24.471	0.872	0.123	2.429	0.841	0.182

1) Ablative Study. We first investigate the impact of the physics-guided feedback mechanism. Removing both depth and normal supervision (row: w/o All Feedback) results in a substantial drop in both relighting quality and geometric consistency, confirming that our physics-inspired loss plays a crucial role in producing physically plausible results. Interestingly, removing only normal feedback degrades the relighting quality more than removing depth alone, suggesting that surface orientation has a stronger influence on realistic light-shadow behavior. We also evaluate the impact of path consistency learning on accelerating inference. When this component is removed (w/o Path Consistency), the model still achieves comparable SSIM and LPIPS scores, with only a marginal

23 drop in physical consistency. This suggests that path consistency introduces minimal overhead in
24 terms of output quality, yet enables substantial acceleration via few-step inference. Therefore, we
25 adopt this module to enhance efficiency while preserving the fidelity of relighting.

26 **2) Feedback as Model Inputs.** We further explore an alternative design where dense depth and
27 normal maps are provided as additional model inputs (rows marked with *). In this setting, the
28 model receives geometric priors directly rather than through supervision. This approach leads to
29 modest gains in generation quality, achieving the highest PSNR and lowest LPIPS when both depth
30 and normal maps are injected as input. However, we observe two important drawbacks: (1) the
31 generated illumination often lacks realistic shadowing or highlight transitions, resulting in physically
32 implausible outcomes despite improved metrics; (2) the overall pipeline becomes heavier, requiring
33 depth and normal estimation during inference, which is computationally inefficient and unfriendly
34 to practical deployment. Moreover, for end users, supplying such dense maps is impractical in
35 real-world applications. This supports our design choice to use depth and normal as training-time
36 feedback, not as runtime input.

37 **3) Effect of Training Domain.** To assess the benefits of our unified framework across modalities,
38 we perform domain-specific training: one version trained only on video pairs, and another only on
39 image pairs. Training exclusively on videos leads to poor generalization in relighting quality, while
40 training only on images degrades temporal consistency. UniLumos, when trained jointly on both
41 domains, achieves the best balance—delivering high-quality and temporally coherent results across
42 inputs. This highlights the strength of our unified formulation and the importance of cross-domain
43 training in learning robust, generalizable relighting behaviors.

44 **B Physics-Plausible Feedback**

45 To further clarify the motivation and design behind our physics-plausible feedback mechanism, we
46 present a breakdown of key questions and considerations addressed during its development.

47 **Q1: What is the motivation for introducing physical constraints in relighting?**

48 The primary goal of relighting is to generate visually plausible illumination under new lighting condi-
49 tions. However, many diffusion-based methods lack explicit physical modeling, leading to artifacts
50 such as overexposed highlights, misaligned shadows, or inconsistent light directions. Introducing
51 physical constraints serves as a refinement mechanism that aligns generated light with the scene’s
52 underlying geometry. This helps enforce realism and spatial consistency in illumination, which is
53 especially critical under complex lighting or HDR scenarios.

54 **Q2: Why are depth and normal maps chosen as the targets for physical supervision?**

55 Depth and surface normals are among the most accessible and general-purpose dense scene attributes.
56 By design, these estimations intentionally suppress fine-scale lighting effects to focus on intrinsic
57 geometry. This makes them ideal for supervising relighting, where the goal is to decouple geometry
58 from illumination and enforce spatial structure in lighting behavior. In the proposed UniLumos,
59 we align the predicted lighting with reference geometry by minimizing the L2 norm error between
60 generated and reference-aligned depth/normal maps via a pre-trained dense estimation model (e.g.,
61 Lotus [2]) with frozen parameters. This provides a simple yet effective metric to quantify physical
62 plausibility.

63 **Q3: Why are alternative physical signals—such as albedo, shadow, or material—not used instead?**

64 While albedo, shadow masks, and material properties can provide rich supervision, they come with
65 significant drawbacks. Albedo and shadow estimation often rely on inverse rendering and suffer
66 from domain sensitivity or ambiguity. Material annotations are expensive and dataset-dependent.
67 Moreover, many of these properties are entangled with illumination, making them less reliable as
68 supervisory signals. In contrast, depth and normals can be predicted from monocular images with
69 high availability and generalize well across scenes, offering a favorable balance between supervision
70 quality and computational cost.

71 **Q4: Why are depth and normal maps used as training-time constraints rather than as model inputs?**

While it is possible to condition the model directly on estimated depth and normal maps, doing so increases the input dimensionality and model complexity. It would also introduce a dependency on external estimators during inference, complicating the pipeline and potentially propagating errors. Instead, we use them as supervision signals during training. This design keeps the inference pipeline simple—relying only on image and lighting condition inputs—while still allowing the model to learn geometry-aware behaviors. The supervision acts as a form of inductive bias, guiding the model toward physically plausible outputs without requiring additional input channels at the test phase.

C Details of Datasets

Step 1: Subject Mask. Given an input video $\mathbf{V}_{\text{real}} \in \mathcal{R}^{[T+1, H, W, 3]}$, we first extract per-frame subject masks $\mathbf{M} \in \mathcal{R}^{[T+1, H, W]}$ using BiRefNet [5]. These subject masks allow us to isolate the target subject foreground and the target background.

Step 2: Lumos Augmentation. To simulate diverse lighting degradations for training, we relight each subject sequence under multiple lighting conditions using a pre-trained 2D relighting model, such as IC-Light [4]. This operation is applied independently to each frame of the subject region, resulting in a degenerated video $\mathbf{V}_{\text{deg}} \in \mathcal{R}^{[T+1, H, W, 3]}$. To generate rich illumination variations, we refer to the description of light and shadow given by IC-Light [4], as listed in Tab. 2, which serve as the semantic guidance for image-level relighting and the light source directions. For each input video, we randomly sample 5 prompts and 3 directions, forming $5 \times 3 = 15$ unique prompt-direction pairs. The relighting is applied only to the subject region, extracted using the subject masks $\mathbf{M} \in \mathcal{R}^{[T+1, H, W]}$ from **Step 1**. Notably, we randomly sample one degradation condition from the 15 prompt-direction combinations for each subject in each iteration. This strategy reduces training cost while exposing the model to diverse illumination patterns, thereby improving generalization.

Table 2: Lighting-related textual prompts used in Lumos Augmentation from IC-Light [4]. Each prompt can be combined with different canonical light directions during training.

ID	Lighting Prompt	Example Light Direction
1	sunshine from window	None
2	neon light, city	Left Light
3	sunset over sea	Right Light
4	golden time	Top Light
5	sci-fi RGB glowing, cyberpunk	Bottom Light
6	natural lighting	
7	warm atmosphere, at home, bedroom	
8	magic lit	
9	evil, gothic, Yarnam	
10	light and shadow	
11	shadow from window	
12	soft studio lighting	
13	home atmosphere, cozy bedroom illumination	
14	neon, Wong Kar-wai, warm	

Step 3: Gaussian Background. To provide external lighting context during training, we generate a background video $\mathbf{V}_{\text{bg}} \in \mathbb{R}^{[T+1, H, W, 3]}$ to accompany the relit subject. Instead of relying on complex inpainting-based synthesis (e.g., ProPainter [6], DiffuEraser [3]), we adopt a simple yet effective strategy by filling the background with either pure color or Gaussian noise. This design avoids injecting semantic or structural priors, allowing the model to focus solely on illumination learning.

Specifically, for each frame $t \in [1, T + 1]$ and channel $c \in \{R, G, B\}$, we first define the background region using the subject mask $\mathbf{M}_t \in \mathbb{R}^{H \times W}$ obtained in **Step 1**. Let $\Omega_{bg}^t = \{(i, j) \mid \mathbf{M}_t(i, j) = 0\}$ denote the set of background pixels. We compute the mean and standard deviation of background pixel intensities as:

$$\mu_c^t = \frac{1}{|\Omega_{bg}^t|} \sum_{(i, j) \in \Omega_{bg}^t} \mathbf{V}_t(i, j, c), \quad \sigma_c^t = \sqrt{\frac{1}{|\Omega_{bg}^t|} \sum_{(i, j) \in \Omega_{bg}^t} (\mathbf{V}_t(i, j, c) - \mu_c^t)^2}. \quad (1)$$

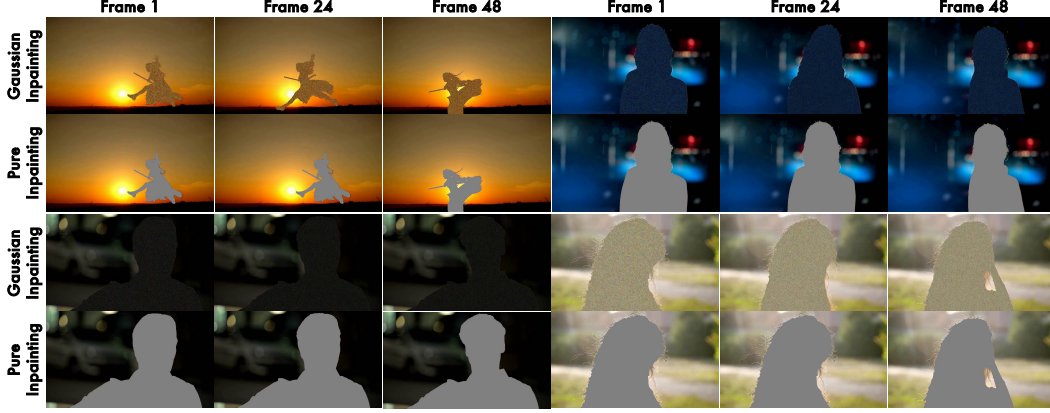


Figure 1: Comparison of background inpainting strategies of four representative cases. Here, *Gaussian Inpainting* fills the background using random noise sampled with the same mean and variance as the subject region, ensuring statistical consistency. *Pure Inpainting* directly fills the background with a uniform color (i.e., gray), without modeling spatial or color variation. The Gaussian strategy provides more realistic signal distribution and accelerates early-stage convergence in training.

105 We then fill the background with pixel-wise samples from a Gaussian distribution:

$$\mathbf{V}_{bg}^t(i, j, c) \sim \mathcal{N}(\mu_c^t, (\sigma_c^t)^2), \quad \forall (i, j) \in \Omega_{bg}^t. \quad (2)$$

106 This procedure ensures that the background region maintains a similar color distribution to the
 107 original video while avoiding structural detail that may bias learning. For comparison, we also test a
 108 variant that uses pure-color background, where each background pixel is set to μ_c^t (i.e., $\sigma_c^t = 0$).

109 *In practice, we observe that such statistically consistent placeholders—particularly Gaussian-filled*
 110 *ones—accelerate early-stage convergence during training.* We attribute this to the reduced visual
 111 complexity and improved normalization behavior, which make the model less sensitive to background
 112 variation. The resulting \mathbf{V}_{bg} serves as a clean, distribution-aligned conditioning signal for the
 113 relighting network.

114 **Step 4: Caption Augmentation.** In addition to relighting augmentation, we generate lighting-aware
 115 captions to provide rich semantic supervision aligned with physical lighting behavior. Specifically, we
 116 leverage Qwen2.5-VL [1], a vision-language model with fine-grained visual reasoning capabilities, to
 117 analyze each input video and generate structured captions describing its lighting attributes. The input
 118 to Qwen2.5-VL consists of the original video and its corresponding scene-level caption. We then
 119 apply a custom-designed prompt (see Listing 1) to steer the model toward predicting six categories of
 120 lighting-related labels as shown in Tab. 3, including all subcategories and their physical interpretations.
 121 The output of this process is a structured caption \mathbf{C} for each video, formatted as a dictionary mapping
 122 the six categories to their predicted labels (see example in Listing 1). These structured captions serve
 123 as auxiliary supervision and evaluation labels in later stages, helping the model better align with
 124 interpretable physical lighting semantics. They also enhance downstream controllability and facilitate
 125 attribute-based retrieval or evaluation.

```

126 SYSTEM_PROMPT = """
127 You are a helpful, respectful and honest assistant. Always answer as helpfully as
128 possible, while being safe. Your answers should not include any harmful,
129 unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure
130 that your responses are socially unbiased and positive in nature. \n\nIf a
131 question does not make any sense, or is not factually coherent, explain why
132 instead of answering something not correct. If you don't know the answer to a
133 question, please don't share false information.
134 """
135
136 PROMPT = """
137 Role: You are an expert in image/video light and shadow analysis, good at analyzing
138 light and shadow from multiple angles.
139 
```

Table 3: Classification criteria and definitions for light-related scene attributes.

Primary Category	Subcategory	Definition
Direction of Light	Front Light Side Light Back Light Top Light Bottom Light Split Light Ambient Light Without Clear Direction	The light source is positioned directly in front of the subject, illuminating it head-on. The light source is positioned at a 90-degree or 45-degree angle to the subject, coming from the side. The light source is located behind the subject, directed towards the camera. The light source is positioned directly above the subject, casting light downwards. The light source is positioned below the subject, casting light upwards. The light source illuminates one side of the subject while leaving the other side in shadow. Ambient light is non-directional, uniformly illuminating the environment from multiple sources.
Light Source Type	Natural Light Artificial Light Rendering Light	Illumination from nature without human intervention, varying with time of day, weather, and location. Human-made light sources (e.g., bulbs, LEDs) used in indoor/outdoor spaces for functional or artistic effects. Digitally simulated light in CGI, games, or animations using techniques like ray tracing.
Light Intensity	Glare Moderate Dim	Extremely bright light over 1000 lumens that can cause discomfort or obscure detail. Balanced lighting (200–1000 lumens), suitable for most activities and comfortable viewing. Low lighting under 200 lumens, often cozy but may reduce visibility and detail recognition.
Color Temperature	Cool Tone Neutral Warm Tone	5000K–10000K; bluish hues, common in daylight or overcast scenes. 4000K–5000K; balanced light with no strong blue or yellow tint. 2000K–4000K; reddish or yellowish hues, typical in sunrise/sunset or indoor lighting.
Light Changes in Time	Static Light Dynamic Light (Intensity Changing) Dynamic Light (Moving Source)	Illumination remains constant in both intensity and direction over time. Light intensity changes gradually over time (e.g., dawn to daylight). Direction of light changes due to movement of light source (e.g., headlights, stage lights).
Optical Phenomena	Transmission (Glass) Refraction/Reflection (Water Surface, Mirror) Scattering (Fog Effect) None	Light passes through transparent materials like glass, with possible scattering or absorption. Light bends or reflects at water or mirror surfaces, altering its direction. Light diffuses through particles like fog or mist, reducing visibility. No significant optical phenomena are observed in the scene.

...

Tasks: Analyze the input image/video, provide corresponding classification results for the following multiple categories, and return them in the specified output format.

- Direction of Light:*
Task 1: Analyze the image and classify the light source direction as front light, side light, back light, top light, bottom light, or split light. Identify the angle of the light source relative to the subject, and describe its effect on shadow formation.
- Light Source Type:*
Task 2: Analyze the image and classify the light source type as either Natural Light, Artificial Light, or Rendering Light.
- Light Intensity:*
Task 3: Analyze the image/video to assess the light intensity present. Classify the light intensity into three categories: Glare, Moderate, and Dim. Special attention should be given to situations where bright light sources may create a glaring effect even in otherwise dim environments.
- Color Temperature:*
Task 4: Analyze the image/video to assess the color temperature present. Classify the color temperature into three categories: Cool Tone, Neutral, and Warm Tone.
- Light Changes in Time:*
Task 5: Analyze the video to assess light changes over time. Classify the light changes into two main categories: Static Light and Dynamic Light. For Dynamic Light, further categorize it into two subtypes: Intensity Gradient and Moving Light Source.
- Optical Phenomena:*
Task 6: Analyze the image/video with a focus on the specific scene to assess the optical phenomena present. Pay close attention to scenarios involving glass, water surfaces, mirrors, and fog. Classify the phenomena into the following categories: Transmission (Glass), Refraction/Reflection (Water Surface, Mirror), Refraction/Reflection (Mirror), Scattering (Fog Effect), and None.

Guidelines:

- Accuracy: Assign each tag to the most appropriate category and subcategory.*
- Multiple Tags: If an action fits multiple categories, assign all relevant tags.*

```

185 3. Comprehensiveness: Capture all detectable dynamic attributes without omissions.
186 4. JSON Validity: Ensure the output JSON is correctly formatted and adheres to the
187   specified structure.
188
189 Example Output:
190 {
191   "Direction of Light": "Front Light",
192   "Light Source Type": "Artificial Light",
193   "Light Intensity": "Moderate",
194   "Color Temperature": "Cool Tone",
195   "Light Changes in Time": "Dynamic Light (Intensity Changing Light)",
196   "Optical Phenomena": "Transmission (Glass)"
197 }
198 """

```

Listing 1: Prompt Definition

200 D Details for Experimental Setup

201 To achieve a balance between training efficiency and physical supervision, we adopt
202 a selective training strategy that dynamically
203 adjusts the loss composition and augmentation
204 configuration during optimization. **(1) Relighting Augmentation Sampling.** For each training sample, we randomly
205 apply one of 15 predefined relighting
206 augmentations (see Tab. 2) to construct
207 the degraded subject input \mathbf{V}_{deg} . This
208 stochastic augmentation introduces sufficient
209 lighting variability during training
210 while reducing the computational burden compared to exhaustively applying all augmentation combinations.
211 **(2) Loss Scheduling.** Our training objective combines three types of supervision signals: the standard
212 flow-matching loss \mathcal{L}_0 , the path consistency loss \mathcal{L}_{fast} , and the physics-guided loss \mathcal{L}_{phy}
213 based on geometric alignment. While jointly optimizing all losses in every iteration is possible, it
214 would significantly increase the computational cost and hinder convergence speed. Instead, we adopt
215 a probabilistic scheduling scheme to balance performance and efficiency. As shown in Alg. 1, in each
216 training iteration, we randomly sample 20% of the minibatch to compute the path consistency loss
217 \mathcal{L}_{fast} , which requires three forward passes and one backward pass due to its recursive structure. The
218 remaining 80% of the data are used for the standard flow-matching loss \mathcal{L}_0 . Among those 80%, half
219 of the samples (i.e., 40% of the full batch) are additionally decoded into RGB space and supervised
220 using the physics-guided loss \mathcal{L}_{phy} , with ground-truth depth and normal maps. This training policy
221 allows us to effectively supervise the model across multiple dimensions—appearance, geometry, and
222 temporal consistency—without incurring the full cost of joint supervision on every sample.

226 E Additional Result Visualization

227 We present additional image relighting results in Fig. 2.

228 We present additional background-conditioned video relighting results in Fig. 3 and Fig. 4.



Figure 2: UniLumos performs physically plausible image relighting, conditioned on textual prompts.

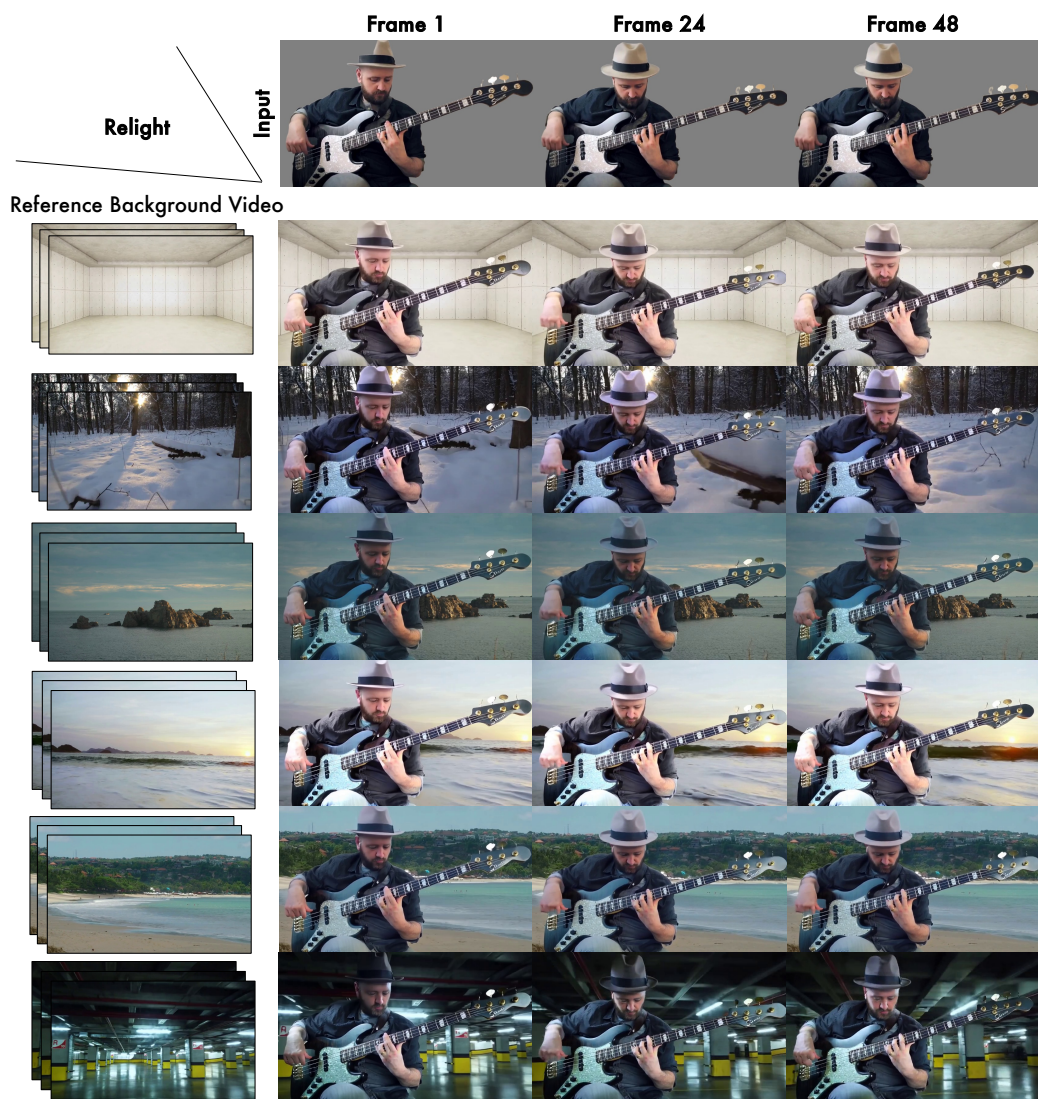


Figure 3: UniLumos performs physically plausible video relighting, conditioned on reference videos.

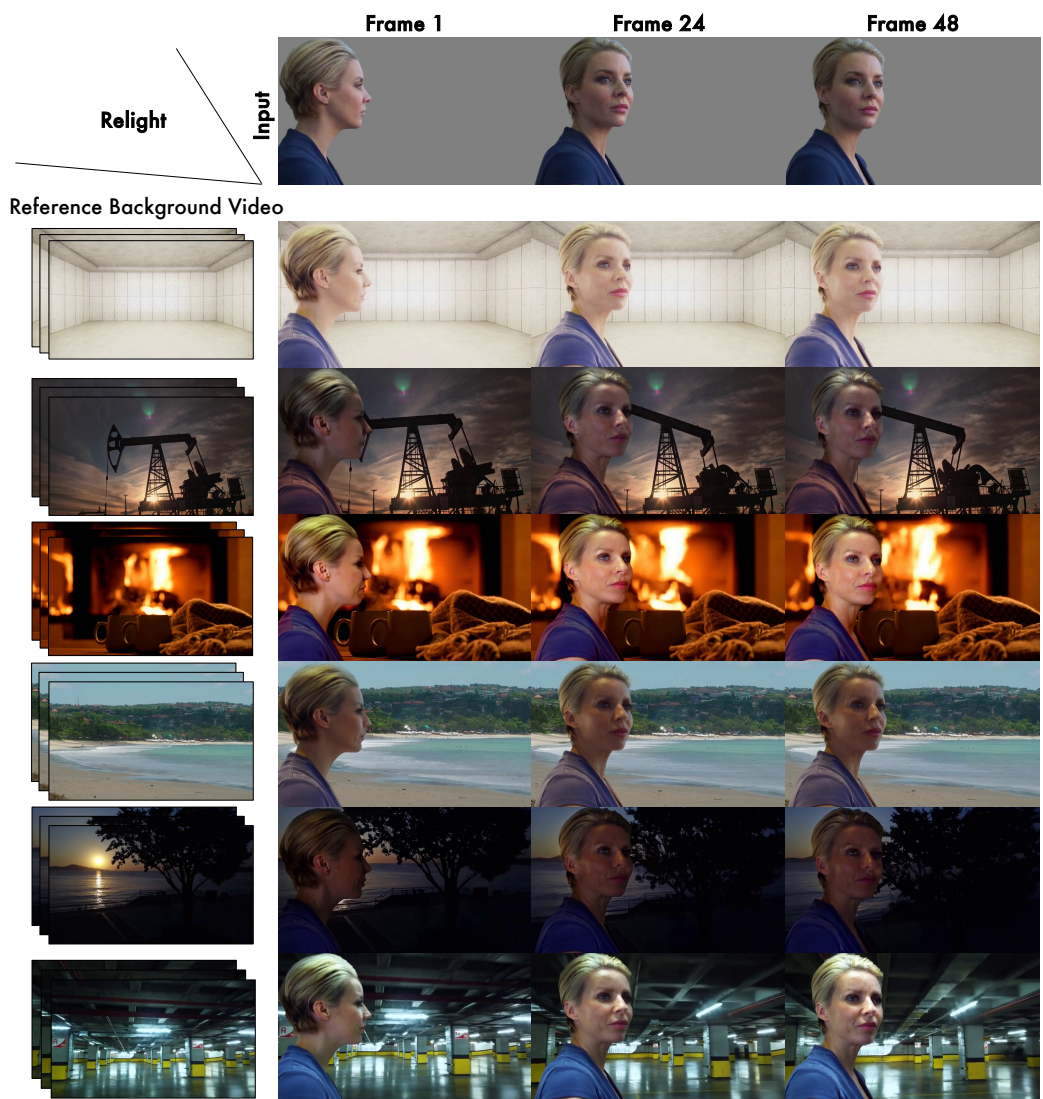


Figure 4: UniLumos performs physically plausible video relighting, conditioned on reference videos.

F Limitation and Future Work

While UniLumos addresses several key limitations of prior relighting methods—specifically, the lack of explicit physical modeling and the inefficiency of multi-step generation—it remains fundamentally constrained by a broader challenge in the field: achieving precise and controllable relighting in a physically measurable manner. In its current form, UniLumos focuses on enforcing geometry-aware consistency (e.g., shadows aligned with depth and normals), but it does not produce fully quantifiable lighting outputs in terms of physical units (e.g., radiance or illuminance levels). Moreover, while our framework supports strong conditioning and generalizes across image and video inputs, future work could explore finer-grained lighting controls (e.g., editable key lights, intensity ramps, or environmental reflections).

G Broader Impact

This work presents a unified relighting framework capable of generating photorealistic images and videos under diverse lighting conditions. As a generative model, UniLumos may support creative and technical applications such as virtual cinematography, augmented reality, and digital design. At the same time, it raises broader questions related to visual authenticity, content manipulation, and the social perception of edited media. While our framework does not perform identity synthesis or hallucinate new visual entities, it enables modification of illumination in ways that may influence narrative tone or perceived realism. This could affect audience interpretation when used in storytelling, advertising, or documentary editing. We thus advocate for transparent usage, including disclosure when relit content is used in downstream applications that impact public perception or interpretation. We believe that responsible innovation requires anticipating downstream risks. In the context of generative relighting, this includes encouraging the development of verification tools, promoting media literacy, and collaborating with industry and policy stakeholders to define norms for fair use and ethical deployment. Our intention is to support creative empowerment without compromising viewer trust or media integrity.

H Safeguards

To support the responsible use of our framework and reduce risks associated with generative technologies, we adopt several safeguards throughout the development and release process. First, UniLumos is explicitly designed to preserve semantic and structural properties of the input, modifying only the illumination. This significantly limits its potential for misuse in identity generation or visual forgery. Second, all training data are sourced from publicly available datasets under appropriate licenses and contain no personally identifiable information. We also recommend that downstream users implement protective measures such as digital watermarking, provenance tracking, and usage disclosures to ensure transparency and traceability. These practices reflect our commitment to safe and accountable AI research. We believe that responsible generative relighting can support creativity while maintaining ethical standards and public trust.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.
- [2] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. arXiv preprint arXiv:2409.18124, 2024.
- [3] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. DiffuEraser: A diffusion model for video inpainting. arXiv preprint arXiv:2501.10018, 2025.
- [4] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In The Thirteenth International Conference on Learning Representations, 2025.
- [5] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. CAAI Artificial Intelligence Research, 3:9150038, 2024.
- [6] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10477–10486, 2023.