UniLumos: Fast and Unified Image and Video Relighting with Physics-Plausible Feedback

Pengwei Liu[1,2,*], **Hangjie Yuan**[†2,3,1,*], **Bo Dong**[2,3], **Jiazheng Xing**[1,2,4], **Jinwang Wang**[2,3,1], **Rui Zhao**[4], **Weihua Chen**[†2,3], **Fan Wang**[2]
[1]Zhejiang University, [2]DAMO Academy, Alibaba Group, [3]Hupan Lab,
[4]National University of Singapore
$*$ Equal contributions. [†] Corresponding author.
{yuanhangjie.yhj, kugang.cwh}@alibaba-inc.com

Abstract

Relighting is a crucial task with both practical demand and artistic value, and recent diffusion models have shown strong potential by enabling rich and controllable lighting effects. However, as they are typically optimized in semantic latent space, where proximity does not guarantee physical correctness in visual space, they often produce unrealistic results—such as overexposed highlights, misaligned shadows, and incorrect occlusions. We address this with **UniLumos**, a unified relighting framework for both images and videos that brings RGB-space geometry feedback into a flow-matching backbone. By supervising the model with depth and normal maps extracted from its outputs, we explicitly align lighting effects with the scene structure, enhancing physical plausibility. Nevertheless, this feedback requires high-quality outputs for supervision in visual space, making standard multi-step denoising computationally expensive. To mitigate this, we employ path consistency learning, allowing supervision to remain effective even under few-step training regimes. To enable fine-grained relighting control and supervision, we design a structured six-dimensional annotation protocol capturing core illumination attributes. Building upon this, we propose **LumosBench**, a disentangled attribute-level benchmark that evaluates lighting controllability via large vision-language models, enabling automatic and interpretable assessment of relighting precision across individual dimensions. Extensive experiments demonstrate that UniLumos achieves state-of-the-art relighting quality with significantly improved physical consistency, while delivering a 20x speedup for both image and video relighting. Code is available at https://github.com/alibaba-damo-academy/Lumos-Custom.

1 Introudction

Relighting, altering illumination in images or videos while preserving intrinsic scene attributes such as geometry, reflectance, and content, is a longstanding problem in computer vision and graphics [33, 55]. It underpins a wide range of applications in film production, gaming, and augmented reality, where seamless lighting integration is critical to visual fidelity. Beyond realism, lighting conveys rich aesthetic and semantic cues—it defines atmosphere, evokes emotion, and reinforces narrative structure. Relighting is thus not only a technical challenge but also a creative tool that shapes how characters, objects, and environments are perceived. However, achieving physically consistent lighting remains a core challenge—requiring the alignment of illumination effects with different scene attributes across both space and time. To address this, traditional approaches often rely on inverse rendering pipelines [46, 50, 2] that estimate intrinsic scene properties, such as geometry, reflectance, and environmental lighting, from input images. While these methods provide physically grounded

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

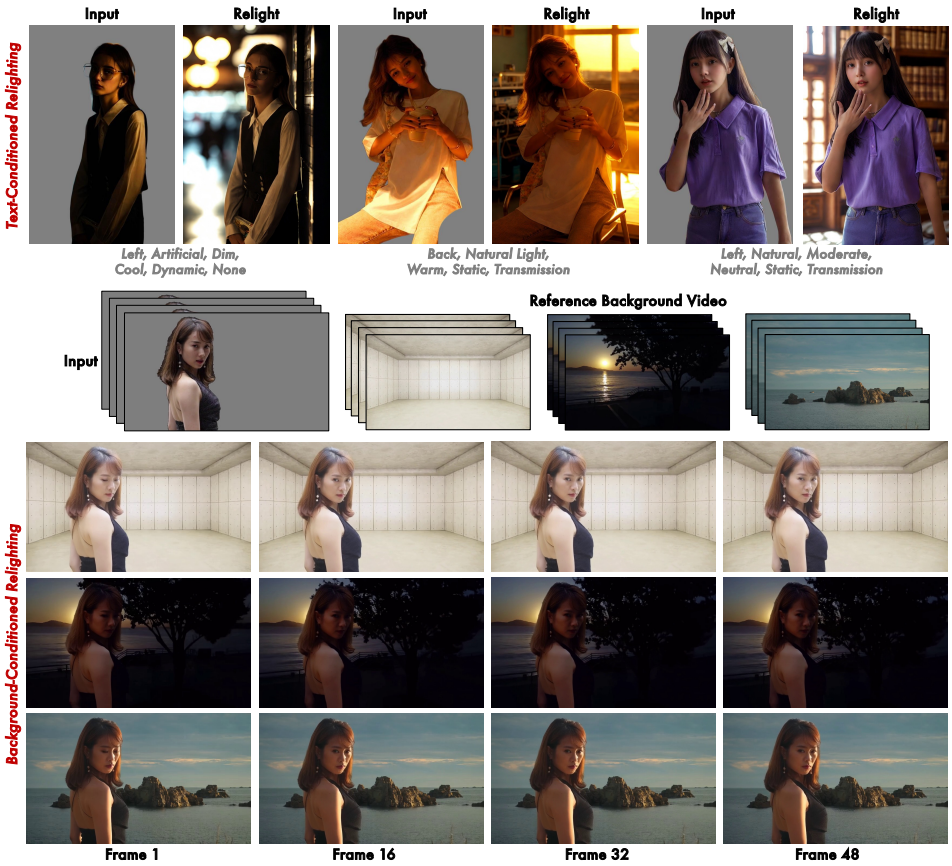

Figure 1: UniLumos performs physically plausible image and video relighting, conditioned on textual prompts and reference videos.

results, they typically require complex inputs—such as high dynamic range images or spherical harmonics coefficients—and are limited to constrained domains. This makes them impractical for real-world applications, where users often provide only a single image, a short video, or a high-level lighting prompt as input. These limitations underscore the need for a new paradigm—one that can deliver high-quality, physically plausible relighting while operating under minimal and naturalistic input conditions.

Recent diffusion-based relighting methods [49, 6, 12, 57] have shown promise by leveraging large-scale image and video datasets to produce diverse and controllable lighting effects under various user-defined conditions, such as reference images or text prompts. However, this strength reveals a fundamental weakness: diffusion models typically operate in semantic latent space, where similarity does not guarantee physical correctness in the visual domain. As a result, they often fail to respect scene geometry, especially in complex scenes with dynamic lighting or temporal constraints. For example, IC-Light [49] and SynthLight [6], which are primarily designed for image relighting, lack both temporal modeling and explicit physical supervision in the visual domain. Instead, they rely on latent-space representations, such as MLP-based embeddings (IC-Light) or multi-stage training (SynthLight), which often lead to artifacts like misaligned shadows, overexposed highlights, or incorrect lighting directions—particularly under complex geometry or extreme illumination. Light-A-Video [12] and RelightVid [12] extend these methods to the video domain, aiming to improve temporal coherence while retaining the visual quality of diffusion-based relighting. Light-A-Video is a training-free framework that combines IC-Light with a pre-trained video diffusion model via iterative alignment. While this improves frame-wise consistency, it incurs high inference costs due to multiple model passes. RelightVid adopts a joint training strategy with a video diffusion backbone, which enhances temporal stability compared to training-free approaches. However, it still operates without explicit physical supervision, resulting in inaccurate light-scene interactions and limited generalization to complex or dynamic environments. In short, existing diffusion-based methods

excel in synthesizing plausible appearances but fall short in enforcing the physical plausibility that is essential for high-quality relighting.

To bridge the gap between generative flexibility and physical correctness, we propose **UniLumos**, a unified relighting framework for both images and videos that brings RGB-space geometry feedback into a flow-matching backbone. Unlike existing diffusion-based approaches that operate purely in latent space, UniLumos introduces a physics-plausible feedback mechanism that supervises generation with dense geometric signals—specifically, depth and surface normals—estimated from its outputs. These lighting-invariant cues serve as ideal supervision signals, enabling the model to explicitly align illumination with the scene structure, significantly improving shadow alignment, shading consistency, and spatial coherence. Nevertheless, this feedback requires high-quality outputs for supervision in visual space, making standard multi-step denoising computationally expensive. To mitigate this, we employ path consistency learning [13], allowing supervision to remain effective even under few-step training regimes.

Beyond model-level improvements, existing relighting methods lack structured illumination descriptions and dedicated evaluation metrics. Generic generation scores (e.g., FID, LPIPS) fail to capture lighting-specific errors such as shadow misalignment, intensity mismatch, or incorrect light direction. To address this, we introduce LumosData, a scalable data pipeline that extracts diverse relighting pairs from real-world videos. At its core is a structured six-dimensional annotation protocol covering direction, light source type, intensity, color temperature, temporal dynamics, and optical phenomena—enabling both fine-grained conditioning during training and physically grounded evaluation at test time. Building upon this, we propose **LumosBench**, a disentangled attribute-level benchmark that evaluates lighting controllability via large vision-language models, enabling automatic and interpretable assessment of relighting precision across individual dimensions.

Our contributions are summarized as follows:

- **Unified Relighting with Physics-Plausible Feedback:** We propose UniLumos, a unified relighting framework for both images and videos that incorporates RGB-space geometry feedback into a flow-matching backbone, explicitly aligning lighting effects with the scene structure to enhance the physical plausibility of relighting.
- **Structured Illumination Annotation and Evaluation Benchmark:** We design a structured six-dimensional annotation protocol that captures core illumination attributes, enabling fine-grained control and supervision. Building upon this, we introduce LumosBench, a disentangled attribute-level benchmark that leverages large vision-language models to automatically and interpretably evaluate relighting controllability across individual lighting dimensions.
- **Extensive Validation:** Extensive experiments demonstrate that UniLumos achieves state-of-the-art relighting quality with significantly improved physical consistency, while delivering a 20x speedup for both image and video relighting.

2 Related Work

Video Diffusion Models. Recent advances in video diffusion models [3, 4, 7, 38] have enabled the generation of temporally coherent video sequences conditioned on various inputs such as text or images. In the field of text-to-video (T2V) generation [45], most methods extend existing text-to-image diffusion backbones with additional modules that capture temporal dynamics across frames. In contrast, a few approaches train video diffusion models from scratch to directly learn spatiotemporal priors [38]. For image-to-video (I2V) tasks, where static images are animated with plausible motion, several methods propose specialized architectures tailored for image animation [35, 40]. Other strategies offer lightweight, plug-and-play adapters that can be integrated into pre-trained models. Stable Video Diffusion [2], for example, fine-tunes T2V models for I2V tasks, achieving state-of-the-art performance. Beyond synthesis quality, a growing body of work emphasizes controllability, allowing users to guide generation with fine-grained constraints [52, 48].

Relighting Methods. Recent advances in deep neural networks have significantly improved lighting control for 2D and 3D visual content, especially in portrait relighting. Methods such as Relightful Harmonization [32], SwitchLight [21], ConceptSliders [14], Intrinsic Image Diffusion [22], Neural Gaffer [20], DI-Light [44], SynthLight [6], and IC-Light [49] demonstrate progress in realism and

controllability. While numerous portrait relighting approaches exist [31, 47, 42, 5], most of them rely heavily on portrait-specific priors. In contrast, UniLumos is designed as a general-purpose relighting framework that is not constrained to any particular object category. With the rise of diffusion-based generative models, approaches like LumiSculpt [51] extend lighting control to text-to-video (T2V) generation. Moreover, RelightVid [12] and Light-A-Video [57] implemented video relighting based on IC-Light. However, achieving both precise lighting control and high visual quality in video remains challenging due to the trade-off between spatial realism and temporal consistency.

Feedback Learning in Generative Models. Feedback learning has become a powerful tool to improve output alignment in generative models, from language systems trained with human preferences [36, 30] to visual diffusion models guided by aesthetic or attribute-based rewards [9, 43, 37, 25], *e.g.*, InstructVideo [43] and DRaFT [10]. In visual domains, feedback can also be physical—for example, using geometric cues to guide generation toward realism. Recent advances in distillation and consistency training [15, 34, 29, 39, 53, 13] have accelerated diffusion inference by reducing the number of denoising steps, enabling models to recover high-quality RGB outputs with just a few iterations. However, most existing techniques focus on appearance synthesis and overlook geometry-aware feedback, which typically requires high-fidelity outputs and is incompatible with few-step inference. UniLumos bridges this gap by combining physically grounded supervision with path-consistency learning, enabling efficient and physically plausible relighting under fast sampling regimes.

3 Preliminaries

Problem Formulation. Given an image or video $\mathbf{S}_1 \in \mathbb{R}^{T\times H\times W\times C}$ with intrinsic scene properties (e.g., geometry, reflectance, content) under initial illumination $\mathbf{L}_1$, the goal of relighting is to modify the illumination within a subject region specified by a binary mask $\mathbf{M} \in \{0,1\}^{T\times H\times W}$ to match a target lighting condition $\mathbf{C}$. The condition $\mathbf{C}$ may take the form of an image, video, or text description and implicitly defines a desired illumination field $\mathbf{L}_2$. The relit output $\mathbf{S}_2 \in \mathbb{R}^{T\times H\times W\times C}$ should exhibit lighting consistent with $\mathbf{L}_2$ in the masked region $\mathbf{M}$ while preserving the intrinsic attributes of $\mathbf{S}_1$. This can be formulated as a conditional generation problem:

$$\mathbf{S}_2 = f_\theta(\mathbf{S}_1, \mathbf{C}, \mathbf{M}), \tag{1}$$

where f_θ is the relighting model parameterized by θ.

Flow Matching. To efficiently model complex illumination transformations in relighting, we build upon *Wan2.1* [38], a foundation video generation model based on flow matching [28, 11]. Flow matching formulates generative modeling as learning velocity fields between noise $x_0 \sim \mathcal{N}(0, I)$ and data x_1, using a linear interpolation:

$$x_t = t \cdot x_1 + (1-t)\cdot x_0, \quad v_t = \frac{\mathrm{d}x_t}{\mathrm{d}t} = x_1 - x_0, \tag{2}$$

where $t \in [0,1]$ is sampled from a logit-normal distribution. The model learns to predict v_t from x_t, conditioned on timestep t and context c (e.g., text embeddings), by minimizing the mean squared error:

$$\mathcal{L}_0 = \mathbb{E}_{x_0,x_1,t,c}\left\|v_\theta(x_t,t,c) - v_t\right\|_2^2. \tag{3}$$

Path Consistency Learning. To further accelerate inference, we adopt path consistency learning [13], which encourages consistent velocity predictions under larger integration steps. Given a velocity field v_θ and step size $d > 0$, we recursively define:

$$x_{t+d} = x_t + d \cdot v_\theta(x_t, t, d), \tag{4}$$

Moreover, enforce two-step consistency using:

$$\mathcal{L}_{\text{fast}} = \mathbb{E}_{x_t,t,d}\left\|v_\theta(x_t,t,2d) - \frac{1}{2}\left[v_\theta(x_t,t,d) + v_\theta(x_{t+d},t+d,d)\right]\right\|_2^2. \tag{5}$$

This objective enables the model to learn shortcut-consistent velocity fields without separate teacher-student stages, allowing for fast and high-quality generation with arbitrary step budgets.

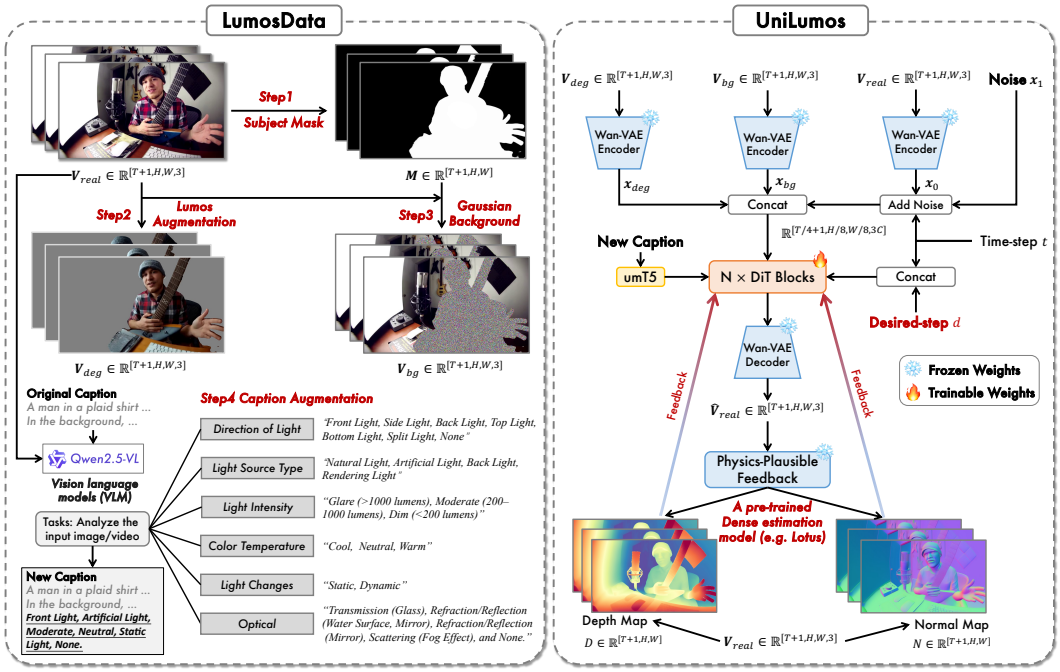

Figure 2: The overall pipeline of **UniLumos**. The left is *LumosData*, our proposed data construction pipeline, which consists of four stages for generating diverse relighting pairs from real-world sources. The right shows the architecture of UniLumos, a unified framework for image and video relighting, designed to achieve physically plausible illumination control.

4 Methodology

We present **UniLumos**, a unified framework for physically plausible image and video relighting, as illustrated in Fig. 2. Built upon Wan 2.1 [38], a flow-matching diffusion model for video generation, UniLumos relights images and videos under user-specified lighting conditions—including reference images, video clips, or text prompts—while preserving scene content and temporal coherence.

To bridge the gap between semantic generation and physical correctness, UniLumos incorporates two key innovations: (1) a physics-plausible feedback that supervises the model with geometry signals from RGB space, and (2) a structured illumination annotation protocol that enables fine-grained control and evaluation. We jointly train the model with geometry-aware supervision and lighting-conditioned objectives, achieving high-quality and efficient few-step inference.

4.1 Physics-Plausible Feedback

While most relighting methods rely on photometric reconstruction or latent-space consistency, such signals offer limited geometric grounding—often resulting in misaligned shadows, implausible shading, and incorrect light directions. To address this, UniLumos enforces consistency between generated illumination and underlying scene geometry, promoting more realistic light–scene interactions.

As illustrated in Fig. 2 (right), we introduce a physics-plausible feedback that guides the generation process using geometry-aware supervision. This component complements the flow-matching architecture with explicit structural priors, enhancing physical plausibility without altering the model's inference inputs. We adopt depth and surface normals as our supervisory targets due to their generality, accessibility, and strong disentanglement from illumination. Unlike shadow masks or material properties, which are often ambiguous, entangled with lighting, or costly to obtain, monocular depth and normal maps capture intrinsic scene structure. They can be reliably estimated by a pre-trained dense estimator (e.g., Lotus [17]).

Specifically, after decoding the predicted latent variable into RGB frames via the *Wan-VAE Decoder* [38], we extract estimated depth $\hat{\mathbf{D}} \in \mathbb{R}^{T\times H\times W}$ and normals $\hat{\mathbf{N}} \in \mathbb{R}^{T\times H\times W}$ using a frozen

dense estimator. These are compared against pseudo-ground-truth maps $(\mathbf{D}, \mathbf{N})$ from the reference input to compute the geometry-aware feedback loss:

$$\mathcal{L}_{\text{phy}} = \mathbb{E}_{x_0, x_1, t, c}\left[\mathbf{M} \odot \left(\frac{\|\hat{\mathbf{D}} - \mathbf{D}\|_2}{\|\mathbf{D}\|_2} + \frac{\|\hat{\mathbf{N}} - \mathbf{N}\|_2}{\|\mathbf{N}\|_2}\right)\right], \tag{6}$$

where $\mathbf{M} \in \mathbb{R}^{T \times H \times W}$ denotes the foreground subject mask. This feedback encourages the model to align its lighting predictions with consistent structural interpretation while keeping inference lightweight and geometry-free.

However, the proposed physics-plausible feedback requires supervision in the RGB domain, which relies on high-quality predictions that are typically only available after full-step denoising has been completed. This poses a major computational bottleneck for standard diffusion models. To mitigate this, we adopt path consistency learning [13], which reformulates denoising as a velocity regression task, thereby supporting practical training under a few-step regimes. Enforcing consistency between intermediate outputs and final predictions enables reliable geometric feedback without sacrificing inference efficiency.

4.2 Structured Illumination Annotation and Evaluation Benchmark

In the problem formulation, the relighting task involves conditioning on a target illumination descriptor $\mathbf{C}$. However, most prior work treats $\mathbf{C}$ as unstructured prompts—such as text, images, or reference frames—offering limited control or interpretability. Moreover, conventional evaluation metrics, such as FID or LPIPS, focus on perceptual similarity but fail to capture lighting-specific discrepancies, such as shadow misalignment or intensity mismatches.

To address this, we construct LumosData, a scalable dataset pipeline that enriches $\mathbf{C}$ with structured lighting semantics. As shown in Fig. 2 (left), we extract relighting pairs from real-world videos. Given an input sequence $\mathbf{V}_{\text{real}} \in \mathbb{R}^{[T+1,H,W,3]}$, we first obtain subject masks $\mathbf{M} \in \{0,1\}^{[T+1,H,W]}$ using BiRefNet [54] to isolate the foreground. We then apply a pre-trained relighting model such as IC-Light [49] to generate synthetic relit versions $\mathbf{V}_{\text{deg}}$ under diverse lighting conditions, guided by a curated prompt set. To avoid entanglement with background semantics, we inpaint the background using Gaussian noise, ensuring clean illumination signals without introducing artifacts.

Beyond this relighting pipeline, LumosData introduces a structured six-dimensional annotation protocol that covers direction, intensity, color temperature, light source type, temporal dynamics, and optical phenomena.

These attributes are automatically generated using vision-language models (e.g., Qwen2.5-VL [1]) with carefully designed prompts, and are integrated into $\mathbf{C}$ to provide an enriched semantic label. This protocol serves dual purposes: (1) Fine-grained conditioning: During training, the model is guided by explicit lighting attributes embedded in $\mathbf{C}$, promoting more interpretable and controllable generation across scenarios. (2) Attribute-aligned Benchmark: Building upon the same attribute protocol, we construct LumosBench. This automatic benchmark uses vision-language models to assess whether generated outputs accurately reflect intended lighting conditions, enabling interpretable, attribute-level controllability evaluation beyond pixel-based metrics.

Each training tuple is structured as $(\mathbf{V}_{\text{deg}}, \mathbf{V}_{\text{bg}}, \mathbf{M}, \mathbf{C}) \rightarrow \mathbf{V}_{\text{real}}$, where the model learns to restore realistic lighting consistent with the semantic cue $\mathbf{C}$ and structural context. LumosData introduces rich diversity in content and illumination by leveraging a variety of real-world sources. Built on Panda70M [8], we curate $\sim$110K high-quality video pairs and augment training with 1.2M additional relit images using IC-Light. This combination supports robust learning of physically plausible relighting without relying on expensive hardware or manual annotations. See more details in Appendix B.

4.3 Joint Learning Objective and Training Strategy

Model Implementation. The inputs of aligned videos $(\mathbf{V}_{\text{deg}}, \mathbf{V}_{\text{bg}}, \mathbf{V}_{\text{real}})$ are passed through a *Wan-VAE Encoder* [38] to obtain semantic latent representations $(\mathbf{x}_{\text{deg}}, \mathbf{x}_{\text{bg}}, \mathbf{x}_0)$. During training, we generate the noisy latent input $\mathbf{x}_t$ via Eq. 2, and concatenate it with the strong conditional signals $\mathbf{x}_{\text{deg}}$ and $\mathbf{x}_{\text{bg}}$ along the channel dimension. This combined tensor is injected into the DiT blocks of the Wan backbone. Additionally, to support path-consistency learning, the diffusion step t and

the expected denoising step d are appended as temporal condition vectors. All new projection and fusion layers are initialized with zero weights to preserve compatibility with the pre-trained Wan initialization and ensure stable optimization from the outset.

Joint Objective. Our training objective integrates three complementary losses to balance appearance fidelity, geometric consistency, and fast inference. The full loss is defined as:

$$\mathcal{L} = \lambda_0 \mathcal{L}_0 + \lambda_1 \mathcal{L}_{\text{fast}} + \lambda_2 \mathcal{L}_{\text{phy}}, \tag{7}$$

where $\mathcal{L}_0$ is the standard flow-matching loss that aligns the predicted velocity field with the ground-truth field, $\mathcal{L}_{\text{fast}}$ is the path consistency loss that improves model performance under few-step denoising regimes, and $\mathcal{L}_{\text{phy}}$ is a physics-guided loss that supervises the RGB outputs using estimated depth and normal maps. We adopt fixed weights of $\lambda_0 = 1.0$ and $\lambda_1 = \lambda_2 = 0.1$ for all experiments. This unified objective encourages the model to generate relit results that are photorealistic, temporally smooth, and physically grounded while supporting efficient inference without sacrificing output quality.

Training Strategy. To balance physical supervision and training efficiency, we adopt a selective optimization strategy inspired by path consistency scheduling [13]. In each training iteration, we divide the batch based on supervision type, following an 80/20 split to avoid prohibitive costs from full supervision while still maintaining effective learning signals. As shown in Alg. 1, 20% of each batch is allocated to compute the path consistency loss $\mathcal{L}_{\text{fast}}$, which involves three forward passes and one backward pass to enforce consistency across timesteps. The remaining 80% is used for the standard flow-matching loss $\mathcal{L}_0$, with 50% of these samples further supervised using RGB-space geometry feedback via $\mathcal{L}_{\text{phy}}$ (i.e., depth and normal alignment). This probabilistic scheduling ensures high training throughput while allowing the model to benefit from multi-level supervision. To further enhance illumination diversity during training, we apply randomized lighting augmentations on the degraded subject $\mathbf{V}_{\text{deg}}$, which introduces realistic lighting variability without the need for explicitly paired captures.

Algorithm 1 Loss Sampling Strategy per Iteration

Require: Batch size B, total training samples
1: **for** each training iteration **do**
2: Randomly sample 20% of batch $\rightarrow \mathcal{L}_{\text{fast}}$
3: Compute path consistency loss:
4: 3× forward, 1× backward
5: Remaining 80% $\rightarrow \mathcal{L}_0$
6: Among those, 50% $\rightarrow$ RGB reconstruction
7: Compute physics-guided loss $\mathcal{L}_{\text{phy}}$
8: **end for**

5 Experiments

5.1 Experimental Details

Training Details. We adopt the Wan2.1-T2V-1.3B-480P [38] as the base model, and initialize all new trainable parameters with zeros to minimize its influence at the beginning of training. We use the AdamW optimizer with the learning rate of 1e-5 for training the entire framework. All the models are trained with a batch size of 8 for 5,000 iterations on 8 NVIDIA H20 GPUs (with 96GB RAM).

Baselines. We compare UniLumos against a range of representative image and video relighting methods. For image-based relighting, we include SwitchLight [21], DiLightNet [44], IC-Light [49], and SynthLight [6], which leverage various forms of latent modeling or light disentanglement to relight single images. For video relighting, we apply IC-Light via frame-by-frame and include Light-A-VideoCogVideoX-2B[41] and another using Wan 2.1 T2V-1.3B [38]. These baselines together represent state-of-the-art performance across both image and video relighting settings.

Dataset. For testing, we selected samples from the internal dataset, processed using the method described in Sec. B. These samples were evenly split: half for image generation at 768x512 resolution, and half for video generation at 480p resolution (832x480), with each video sample containing 49 frames. To further demonstrate the model's generalization to non-human scenes, we conducted additional evaluations on two public object-centric relighting benchmarks: StanfordOrb [24] and Navi [19], which include objects and sculptures under a variety of lighting environments and are completely disjoint from our training data, and see more results in Appendix C.1.

Evaluation metrics. We evaluate relighting performance across three key dimensions: (1) visual fidelity: We assess image quality using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity

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

Model	(a) Quality			(b) Temporal Consistency	(c) Lumos Consistency	
	PSNR ↑	SSIM ↑	LPIPS ↓	R-Motion↓	Avg. Score ↑	Dense L2 Error ↓
Image Relighting						
SwitchLight [21]	20.483	0.901	0.094	-	0.717	0.388
DiLightNet [44]	21.894	0.860	0.131	-	0.682	0.401
IC-Light [49]	24.316	0.884	0.108	-	0.703	0.447
SynthLight [6]	25.572	0.905	0.102	-	0.791	0.214
UniLumos	**26.719**	**0.913**	**0.089**	-	**0.912**	**0.103**
Video Relighting						
IC-Light Per Frame [49]	20.132	0.851	0.133	2.437	0.672	0.432
Light-A-Video [57] + CogVideoX[41]	19.851	0.859	0.124	1.784	0.641	0.383
Light-A-Video [57] + Wan2.1[38]	20.784	0.876	0.129	1.582	0.682	0.371
UniLumos	**25.031**	**0.891**	**0.109**	**1.436**	**0.871**	**0.147**

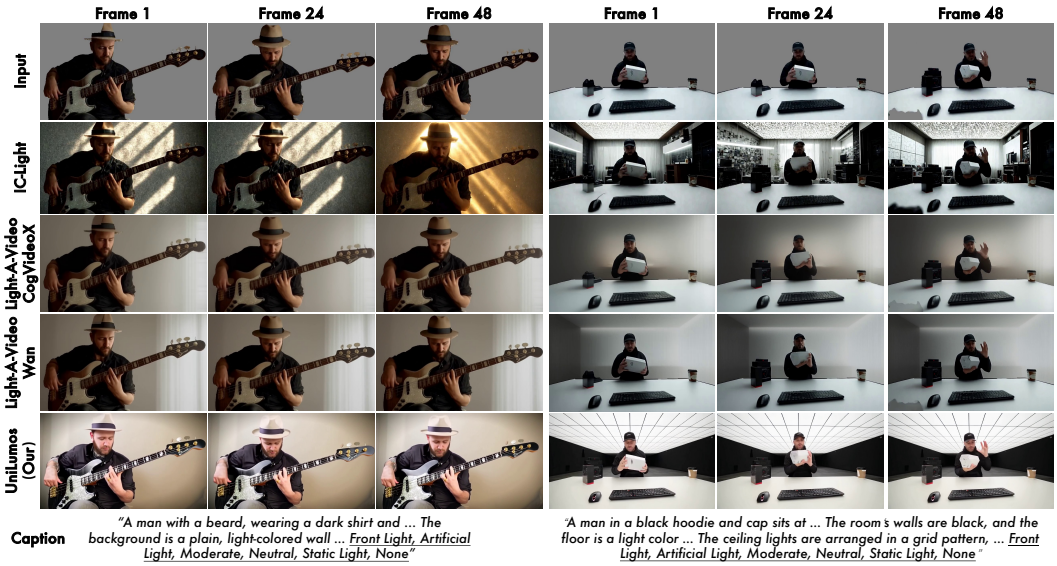

Figure 3: Qualitative comparison of baseline methods. Each method takes a subject video and a textual illumination description as input, generating the related subject with the corresponding background under the specified lighting condition.

Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). For video relighting, we report the average metric across all frames. (2) temporal consistency: Following VBench [18], we adopt the R-Motion metric, which measures temporal smoothness using motion priors from a pre-trained video frame interpolation model [27]. This captures the coherence of lighting transitions across frames. (3) lumos consistency: (i) Lumos Score, computed by applying the same caption-based lighting annotation protocol as in our LumosData construction. Six lighting attributes are predicted and compared with targets, and each is weighted equally to yield an average consistency score. (ii) Dense L2 Error, which quantifies the relative L2 error between predicted and reference depth/normal maps, estimated via a pre-trained geometry model (e.g., Lotus [17]). This provides a physically grounded measure of illumination-geometry alignment.

5.2 Main Results

Quantitative Evaluation. As shown in Tab. 1, UniLumos delivers consistent improvements across three key dimensions: visual fidelity, temporal consistency, and physically grounded lighting alignment. (1) Visual Fidelity. UniLumos produces higher-quality relighting results across both images and videos. Benefiting from structured lighting supervision and geometry-guided feedback, our model generates outputs with clearer shading, sharper details, and more coherent illumination compared to prior works. (2) Temporal Consistency. For video relighting, UniLumos ensures smoother transitions and reduced flickering artifacts. Our use of flow-matching architecture and path consistency learning helps maintain stable lighting across frames, addressing a key limitation in frame-wise

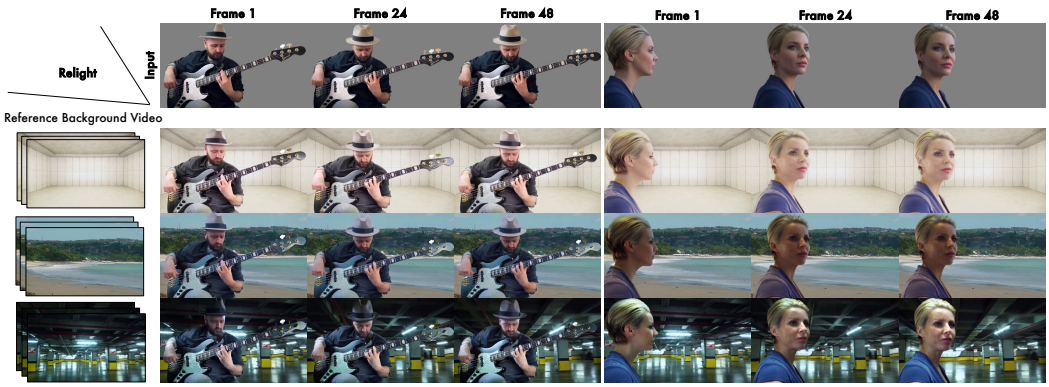

Figure 4: UniLumos performs physically plausible video relighting conditioned on different reference videos.

or training-free methods. (3) Lumos Consistency. Going beyond appearance-based evaluation, UniLumos aligns well with intended lighting semantics. Through structured caption conditioning and physics-guided training, the model better preserves lighting direction, tone, and geometry—validated by both vision-language alignment and dense geometric error metrics.

Efficiency. To assess inference efficiency, we evaluate the video relighting task under a standardized setting, generating 49-frame videos at a 480p resolution. As shown in Fig. 5, UniLumos achieves a significant speedup compared to prior methods, benefiting from its geometry-free inference and few-step generation. While existing models, such as Light-A-Video or IC-Light, require either iterative frame-by-frame processing or complex sampling schedules, UniLumos completes generation over 20 times faster without sacrificing visual fidelity or physical plausibility. This efficiency advantage makes it well-suited for real-time relighting applications and scalable deployment scenarios.

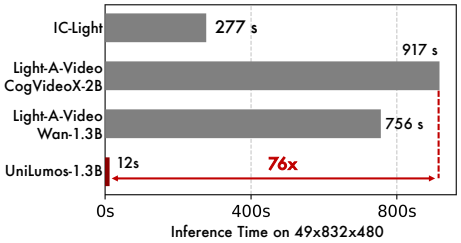

Figure 5: Comparison of inference time costs of different methods under the same settings.

Qualitative Results We present qualitative comparisons in Fig. 3 and Fig. 4, highlighting the advantages of UniLumos in terms of lighting realism, temporal coherence, and controllability. (1) Lighting Quality and Controllability: In Fig. 3, UniLumos produces lighting effects that better match the target description, capturing nuanced directional shadows, color tone, and intensity. Competing methods either fail to reflect the intended lighting change or produce overly uniform results that lack realism. (2) Temporal Consistency: Compared to baseline methods such as frame-wise IC-Light and Light-A-Video, UniLumos achieves smoother frame transitions without flickering or structural distortion. This benefit arises from the joint modeling of space and time, which is further reinforced by physics-aware supervision and path consistency training. (3) Foreground Detail Preservation: UniLumos preserves fine subject details—such as facial structure and clothing texture—better than baselines. For instance, Light-A-Video occasionally introduces deformation or identity drift, while our model maintains high fidelity over long sequences. (4) Relighting with Reference Videos: Fig. 4 showcases UniLumos conditioned on different reference videos. The model successfully adapts both global lighting direction and subtle spatial variations across scenes, demonstrating strong generalization under diverse illumination cues.

LumosBench To evaluate the fine-grained controllability of lighting generation, we introduce LumosBench, a structured benchmark that targets six core illumination attributes defined in our annotation protocol. Unlike prior works that treat lighting holistically or implicitly, LumosBench provides a disentangled, attribute-level evaluation, enabling precise diagnosis of model behavior under controllable lighting conditions. See more results in Appendix C.2.

5.3 Ablation Study

As shown in Tab. 2 and Fig. 6, we conduct ablation studies to analyze the effectiveness of different components. **Physics-Guided Feedback.** Removing both depth and normal feedback (w/o All

Table 2: Quantitative comparison. **Bold** number indicate the best performance.

Model	(a) Quality			(b) Temporal Consistency	(c) Lumos Consistency	
	PSNR ↑	SSIM ↑	LPIPS ↓	R-Motion↓	Avg. Score ↑	Dense L2 Error ↓
Ablative Study						
w/o Depth Feedback	23.472	0.883	0.118	1.443	0.870	0.265
w/o Normal Feedback	22.115	0.874	0.123	1.446	0.863	0.173
w/o All Feedback	21.433	0.862	0.139	1.473	0.859	0.297
w/o Path Consistency	**25.317**	**0.902**	0.113	1.438	**0.875**	0.153
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
UniLumos	25.031	0.891	**0.109**	**1.436**	0.871	**0.147**

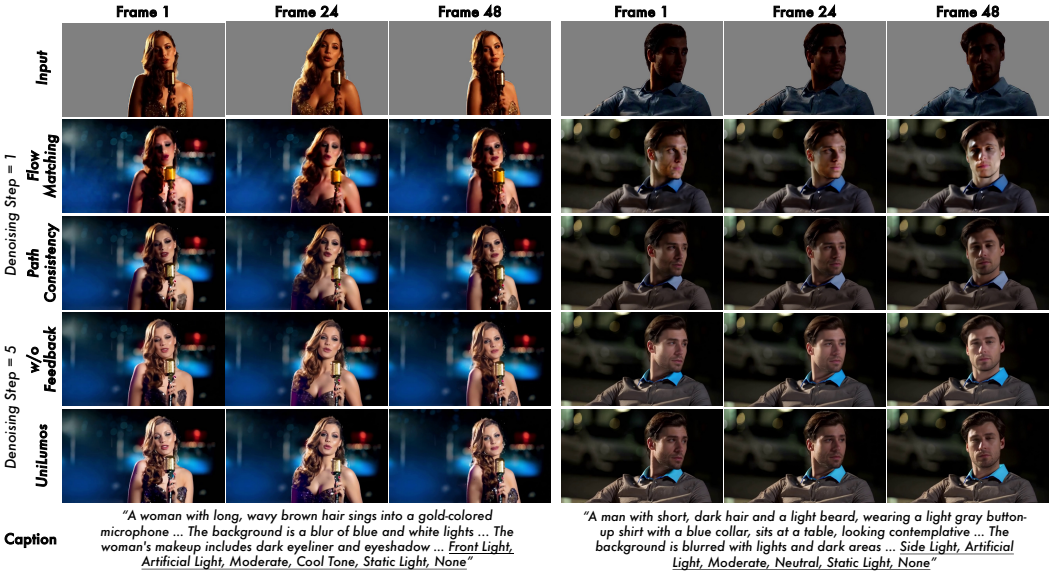

Figure 6: Ablation study. We compare the effects of different components under few-step denoising. For 1-step, we show the impact of flow-matching with and without path consistency. For 5-step, we visualize results before and after introducing physics-plausible feedback.

Feedback) leads to significant degradation in both image quality and physical consistency, confirming the necessity of our physics-guided loss. Notably, omitting only normal supervision causes a larger drop than removing depth, suggesting that surface orientation plays a more critical role than distance in shaping light–shadow interactions. **Path Consistency Learning.** Excluding this component (w/o Path Consistency) yields only minor drops in physical metrics while maintaining competitive SSIM and LPIPS scores. This shows that path consistency incurs little performance cost but offers substantial efficiency benefits in few-step regimes, justifying its inclusion. **Training Modality.** To evaluate the effectiveness of our unified training paradigm, we compare domain-specific variants: training solely on videos leads to poor visual quality, while image-only training sacrifices temporal smoothness. In contrast, our unified approach strikes the best balance—achieving high-quality and temporally coherent relighting across both input types.

6 Conclusion

We introduce UniLumos, a unified framework for physically plausible image and video relighting. It aligns illumination with scene geometry via RGB-space depth and normal supervision, improving shadow accuracy and spatial consistency. To enhance controllability and evaluation, we propose a structured six-dimensional lighting annotation protocol, enabling fine-grained conditioning and physically grounded assessment through VLMs. Experiments show that UniLumos achieves superior relighting quality, physical consistency, and inference efficiency.

Acknowledgment

This work was supported by Damo Academy through Damo Academy Research Intern Program.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 22563–22575, 2023.

[5] Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6221–6231, 2024.

[6] Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. Synthlight: Portrait relighting with diffusion model by learning to re-render synthetic faces. arXiv preprint arXiv:2501.09756, 2025.

[7] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512, 2023.

[8] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13320–13331, 2024.

[9] Weifeng Chen, Jiacheng Zhang, Jie Wu, Hefeng Wu, Xuefeng Xiao, and Liang Lin. Id-aligner: Enhancing identity-preserving text-to-image generation with reward feedback learning. arXiv preprint arXiv:2404.15449, 2024.

[10] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. arXiv preprint arXiv:2309.17400, 2023.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first international conference on machine learning, 2024.

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

[13] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. arXiv preprint arXiv:2410.12557, 2024.

[14] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In European Conference on Computer Vision, pages 172–188. Springer, 2024.

[15] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. International Journal of Computer Vision, 129(6):1789–1819, 2021.

[16] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103, 2024.

[17] 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.

[18] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21807–21818, 2024.

[19] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. Advances in Neural Information Processing Systems, 36:76061–76084, 2023.

[20] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. Advances in Neural Information Processing Systems, 37:141129–141152, 2024.

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

[22] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5198–5208, 2024.

[23] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

[24] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: a real-world 3d object inverse rendering benchmark. Advances in Neural Information Processing Systems, 36:46938–46957, 2023.

[25] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192, 2023.

[26] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. arXiv preprint arXiv:2501.10018, 2025.

[27] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9801–9810, 2023.

[28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

[29] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[31] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. ACM Trans. Graph., 40(4):43–1, 2021.

[32] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6452–6462, 2024.

[33] Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. Image based relighting using neural networks. ACM Transactions on Graphics (ToG), 34(4):1–12, 2015.

[34] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512, 2022.

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. Advances in neural information processing systems, 32, 2019.

[36] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. Advances in neural information processing systems, 33:3008–3021, 2020.

[37] Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal metric learning from human feedback for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13559–13568, 2024.

[38] 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.

[39] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. arXiv preprint arXiv:2312.09109, 2023.

[40] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1481–1490, 2024.

[41] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024.

[42] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. ACM Transactions on Graphics (TOG), 41(6):1–21, 2022.

[43] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6463–6474, 2024.

[44] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In ACM SIGGRAPH 2024 Conference Papers, pages 1–12, 2024.

[45] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision, pages 1–15, 2024.

[46] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5453–5462, 2021.

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

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3836–3847, 2023.

[49] 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.

[50] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18643–18652, 2022.

[51] Yuxin Zhang, Dandan Zheng, Biao Gong, Jingdong Chen, Ming Yang, Weiming Dong, and Changsheng Xu. Lumisculpt: A consistency lighting control network for video generation. arXiv preprint arXiv:2410.22979, 2024.

[52] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems, 36:11127–11150, 2023.

[53] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation. arXiv e-prints, pages arXiv–2402, 2024.

[54] 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.

[55] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In Proceedings of the IEEE/CVF international conference on computer vision, pages 7194–7202, 2019.

[56] 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.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have made clear claims about contributions and scopes in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have the seperated limitation section in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: The answer NA means that the paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the details of our model and training strategy in the paper for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We offer the implementation code to reproduce our work.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have introduced experiment settings and details, and even add more details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We do not report specific error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have discussed the information about computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss this in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Users should follow usage policies and use models with safety filters and access controls.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original paper that produced the code package and dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide documentation with our code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only use LLM for proof-reading.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix of UniLumos

In this appendix, we provide additional details to complement the main paper. First, we explain the motivation behind introducing physics-plausible feedback in **Sec. A**. Next, we present the detailed pipeline of our proposed LumosData relighting data construction process in **Sec. B**. Then, **Sec. C** contains three additional experimental results on the public dataset (Sec. C.1) and the LumosBench (Sec. C.2). **Sec. D** provides additional qualitative results to further illustrate the effectiveness of UniLumos. Finally, **Sec. E**, **Sec. F**, and **Sec. G** discuss the limitations, broader impacts, and safeguards of our method.

A Physics-Plausible Feedback

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

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

A1: The primary goal of relighting is to generate visually plausible illumination under new lighting conditions. However, many diffusion-based methods lack explicit physical modeling, leading to artifacts such as overexposed highlights, misaligned shadows, or inconsistent light directions. Introducing physical constraints serves as a refinement mechanism that aligns generated light with the scene's underlying geometry. This helps enforce realism and spatial consistency in illumination, which is especially critical under complex lighting or HDR scenarios.

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

A2: Depth and surface normals are among the most accessible and general-purpose dense scene attributes. By design, these estimations intentionally suppress fine-scale lighting effects to focus on intrinsic geometry. This makes them ideal for supervising relighting, where the goal is to decouple geometry from illumination and enforce spatial structure in lighting behavior. In the proposed UniLumos, we align the predicted lighting with reference geometry by minimizing the L2 norm error between generated and reference-aligned depth/normal maps via a pre-trained dense estimation model (e.g., Lotus [17]) with frozen parameters. This provides a simple yet effective metric to quantify physical plausibility.

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

A3: While albedo, shadow masks, and material properties can provide rich supervision, they come with significant drawbacks. Albedo and shadow estimation often rely on inverse rendering and suffer from domain sensitivity or ambiguity. Material annotations are expensive and dataset-dependent. Moreover, many of these properties are entangled with illumination, making them less reliable as supervisory signals. In contrast, depth and normals can be predicted from monocular images with high availability and generalize well across scenes, offering a favorable balance between supervision quality and computational cost.

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

A4: 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.

B 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 [54]. These subject masks allow us to isolate the target subject foreground and the target background.

Table 3: Lighting-related textual prompts used in Lumos Augmentation from IC-Light [49]. 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, Yharnam	
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 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 [49]. 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 [49], as listed in Tab. 3, 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.

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 [56], DiffuEraser [26]), 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 $\Omega^t_{bg} = \{(i,j) \mid \mathbf{M}_t(i,j) = 0\}$ denote the set of background pixels. We compute the mean and standard deviation of background pixel intensities as:

$$\begin{cases} \mu^t_c = \dfrac{1}{|\Omega^t_{bg}|} \displaystyle\sum_{(i,j)\in\Omega^t_{bg}} \mathbf{V}_t(i,j,c), \\ \sigma^t_c = \sqrt{\dfrac{1}{|\Omega^t_{bg}|} \displaystyle\sum_{(i,j)\in\Omega^t_{bg}} \left(\mathbf{V}_t(i,j,c) - \mu^t_c\right)^2} \end{cases} \tag{8}$$

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

$$\mathbf{V}^t_{bg}(i,j,c) \sim \mathcal{N}(\mu^t_c, (\sigma^t_c)^2), \quad \forall (i,j) \in \Omega^t_{bg}. \tag{9}$$

This procedure ensures that the background region maintains a similar color distribution to the original video while avoiding structural detail that may bias learning. For comparison, we also test a variant that uses pure-color background, where each background pixel is set to μ^t_c (i.e., $\sigma^t_c = 0$). In practice, we observe that such statistically consistent placeholders—particularly Gaussian-filled ones—accelerate early-stage convergence during training. We attribute this to the reduced visual complexity and improved normalization behavior, which make the model less sensitive to background variation. The resulting $\mathbf{V}_{bg}$ serves as a clean, distribution-aligned conditioning signal for the relighting network.

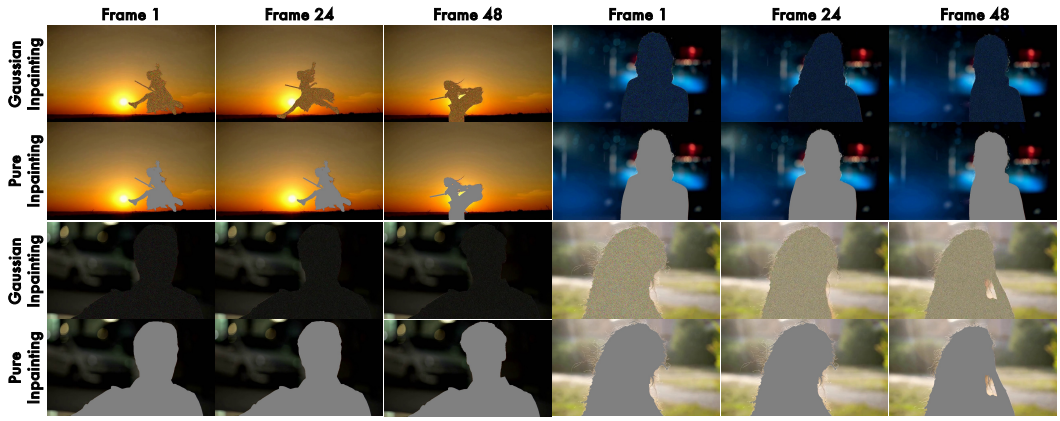

Figure 7: Comparison of background inpainting strategies of four representative cases. Here, *Gaussian Inpainting* fills the background using random noise sampled with the same mean and variance as the subject region, ensuring statistical consistency. *Pure Inpainting* directly fills the background with 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.

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

Table 4: Classification criteria and definitions for light-related scene attributes.

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

```
SYSTEM_PROMPT = """
You are a helpful, respectful and honest assistant. Always answer as helpfully as
    possible, while being safe. Your answers should not include any harmful,
    unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure
    that your responses are socially unbiased and positive in nature.\n\nIf a
    question does not make any sense, or is not factually coherent, explain why
    instead of answering something not correct. If you don't know the answer to a
    question, please don't share false information.
"""

PROMPT = """
Role: You are an expert in image/video light and shadow analysis, good at analyzing
    light and shadow from multiple angles.
...

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

1. 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.

2. Light Source Type:
   Task 2: Analyze the image and classify the light source type as either Natural
       Light, Artificial Light, or Rendering Light.

3. 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.

4. 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.

5. 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.

6. 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:

1. Accuracy: Assign each tag to the most appropriate category and subcategory.
2. Multiple Tags: If an action fits multiple categories, assign all relevant tags.
3. Comprehensiveness: Capture all detectable dynamic attributes without omissions.
4. JSON Validity: Ensure the output JSON is correctly formatted and adheres to the
    specified structure.

Example Output:
{
```

```
    "Direction of Light": "Front Light",
    "Light Source Type": "Artificial Light",
    "Light Intensity": "Moderate",
    "Color Temperature": "Cool Tone",
    "Light Changes in Time": "Dynamic Light (Intensity Changing Light)",
    "Optical Phenomena": "Transmission (Glass)"
}
"""
```

Listing 1: Prompt Definition

C Additional Experimental Results

C.1 Additional Main Results

To further demonstrate the model's generalization to non-human scenes, we conducted additional evaluations on two public object-centric relighting benchmarks: StanfordOrb [24] and Navi [19]. These datasets include objects and sculptures under a variety of lighting environments and are completely disjoint from our training data. StanfordOrb contains canonical 3D scanned objects such as the Stanford bunny and dragon, while Navi includes a wide range of everyday objects like containers, toys, and mugs. As shown in Tab. 5 (StanfordOrb) and Tab. 6 (Navi), UniLumos achieves state-of-the-art results across perceptual (LPIPS), structural (SSIM), and physical (R-Motion) metrics, despite the significant domain gap and without any test-time fine-tuning, outperforming all baselines.

Table 5: Quantitative comparison on the StanfordOrb dataset. **Bold** number indicate the best performance.

Model	PSNR ↑	SSIM ↑	LPIPS ↓	R-Motion ↓
IC-Light Per Frame [49]	24.132	0.914	0.126	1.742
Light-A-Video [57] + CogVideoX [41]	25.617	0.923	0.108	1.279
Light-A-Video [57] + Wan2.1 [38]	25.784	0.926	0.104	1.241
UniLumos	**26.512**	**0.934**	**0.097**	**1.103**

Table 6: Quantitative comparison on the Navi dataset. **Bold** number indicate the best performance.

Model	PSNR ↑	SSIM ↑	LPIPS ↓	R-Motion ↓
IC-Light Per Frame [49]	22.021	0.883	0.125	1.974
Light-A-Video [57] + CogVideoX [41]	23.912	0.891	0.121	1.378
Light-A-Video [57] + Wan2.1 [38]	23.474	0.903	**0.116**	1.341
UniLumos	**24.977**	**0.911**	0.120	**1.203**

C.2 LumosBench: An Attribute-level Controllability Benchmark

To evaluate the fine-grained controllability of lighting generation, we introduce LumosBench, a structured benchmark that targets six core illumination attributes defined in our annotation protocol. Unlike prior works that treat lighting holistically or implicitly, LumosBench provides a disentangled, attribute-level evaluation, enabling precise diagnosis of model behavior under controllable lighting conditions.

Specifically, we construct a set of 2k test prompts, each consisting of a video and a structured caption designed to isolate one lighting attribute at a time, while holding other variables constant. These prompts span six categories—direction, light source type, intensity, color temperature, temporal dynamics, and optical phenomena—with multiple subtypes per category (e.g., front/side/back for direction). This design facilitates controlled and interpretable evaluation across lighting axes that are often conflated in prior datasets.

Table 7: Quantitative comparison of the attribute-level controllability. **Bold** number indicate the best performance.

Model	#Params	Direction	Light Source Type	Intensity	Color Temperature	Temporal Dynamics	Optical Phenomena	Avg. Score
General Models								
LTX-Video [16]	1.9B	0.794	0.644	0.487	0.708	0.487	0.403	0.587
CogVideoX [41]	5.6B	0.837	0.692	0.552	0.739	0.532	0.449	0.634
HunyuanVideo [23]	13B	0.863	0.741	0.599	0.802	0.655	0.481	0.690
Wan2.1[38]	1.3B	0.842	0.685	0.436	0.741	0.504	0.433	0.607
Wan2.1[38]	14B	0.871	0.794	0.674	**0.829**	**0.737**	0.505	0.735
Specialized Models								
IC-Light Per-Frame [49]	0.9B	0.793	0.547	0.349	0.493	0.284	0.339	0.468
Light-A-Video [57] + CogVideoX[41]	2.9B	0.787	0.581	0.327	0.536	0.493	0.373	0.516
Light-A-Video [57] + Wan2.1[38]	2.2B	0.801	0.603	0.361	0.582	0.557	0.412	0.553
UniLumos w/o lumos captions	1.3B	0.868	0.774	0.529	0.798	0.543	0.457	0.662
UniLumos	1.3B	**0.893**	**0.847**	**0.832**	0.813	0.662	**0.592**	**0.773**

To assess alignment between intended and generated lighting attributes, we use the vision-language model Qwen2.5-VL [1] to analyze relit outputs and classify whether the target attribute is correctly expressed. Each dimension is scored independently, and the final controllability score is the average across all six dimensions.

General vs. Specialized Models. Tab. 7 presents results for both general-purpose and task-specific relighting models. This benchmark allows us not only to assess overall lighting quality but also to dissect a model's ability to interpret and respond to individual lighting controls. Among general models, Wan 14B shows the highest raw capability, demonstrating the strength of large-scale pretraining for visual generation. Notably, our fine-tuned Wan 1.3B variant achieves substantial gains across all six lighting dimensions, surpassing even much larger models. This highlights the benefit of relighting-specific supervision: fine-tuning on LumosData with structured lighting annotations significantly enhances the model's ability to reason about illumination in a controllable and disentangled manner.

By contrast, specialized relighting models consistently underperform despite being designed for lighting manipulation. This is primarily due to the limitations of their base architectures—typically smaller and trained from scratch or on narrow domains—resulting in weaker generalization to diverse lighting attributes. While they may encode some prior knowledge of lighting physics (e.g., through latent constraints), their restricted modeling capacity hinders semantic alignment with user-intended lighting conditions. These findings underscore the importance of starting from a strong pre-trained backbone and introducing structured, high-level lighting supervision to achieve controllable and physically plausible relighting.

Effectiveness of Structured Captions. Within the specialized group, we conduct an ablation to assess the importance of our proposed lumos captions. **w/o lumos captions** uses only vanilla scene-level captions during training, omitting structured lighting tags. The performance drop—particularly in controllable dimensions like intensity and optical phenomena—confirms that our semantic annotations play a key role in teaching the model fine-grained illumination control. Compared to strong baselines, UniLumos achieves superior scores across nearly all dimensions, demonstrating the impact of LumosBench in pushing model understanding and control of illumination.

D Additional Result Visualization

We present additional image relighting results in Fig. 8.

We present additional background-conditioned video relighting results in Fig. 9 and Fig. 10.

E Limitation and Future Work

UniLumos is still limited by a broader challenge—achieving physically precise and controllable relighting. UniLumos enforces geometry-aware consistency (e.g., shadows aligned with depth and normals) but does not yet produce physically quantifiable lighting outputs such as radiance or illuminance. Future work may explore finer control over lighting, including editable key lights, intensity ramps, and environmental reflections.

Figure 8: UniLumos performs physically plausible image relighting, conditioned on textual prompts.

F Broader Impact

UniLumos can support creative and technical applications such as virtual cinematography, augmented reality, and digital design. However, generative relighting also raises concerns about visual authenticity, content manipulation, and social perception. Our model does not perform identity synthesis or create new visual entities, but it can alter illumination in ways that affect narrative tone or realism. When used in storytelling or media production, such changes may influence audience interpretation. We encourage transparent and responsible use, including disclosure when relit content appears in downstream applications. Responsible innovation in this area requires anticipating potential misuse, supporting verification tools, promoting media literacy, and engaging with stakeholders to define fair and ethical use. Our goal is to enable creative expression while maintaining viewer trust and media integrity.

G Safeguards

To promote responsible use and reduce risks of generative misuse, we include several safeguards in the development and release of UniLumos. All training data come from public datasets with appropriate licenses and contain no personally identifiable information. We also recommend that downstream users apply protective measures such as watermarking, provenance tracking, and clear usage disclosures to support transparency. These practices aim to ensure safe deployment and maintain public trust in generative relighting systems.

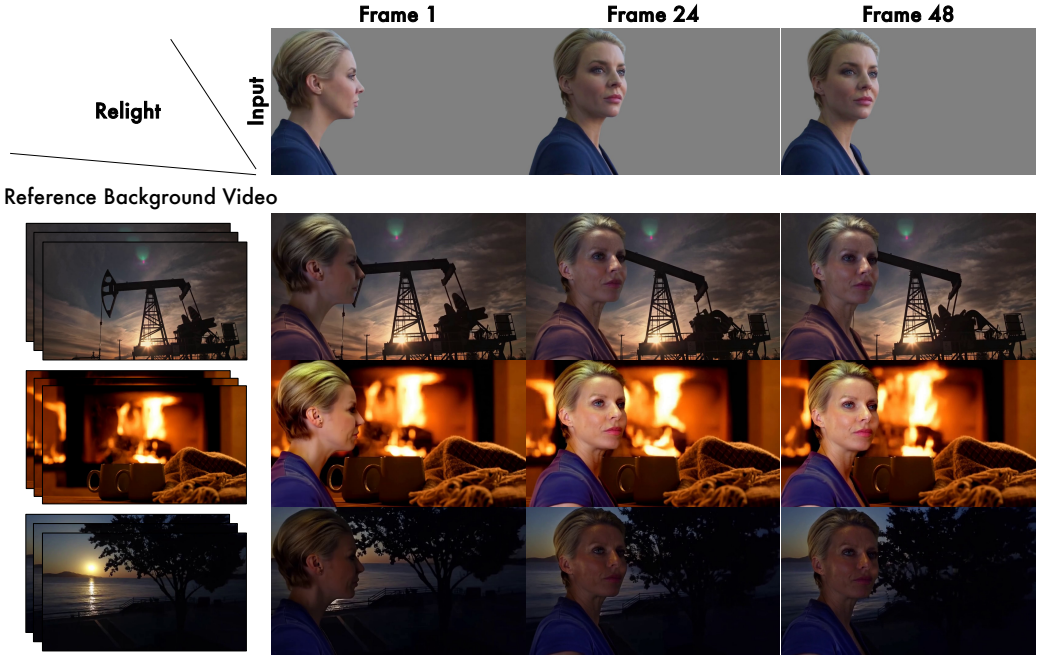

Figure 9: UniLumos performs physically plausible video relighting, conditioned on reference videos.

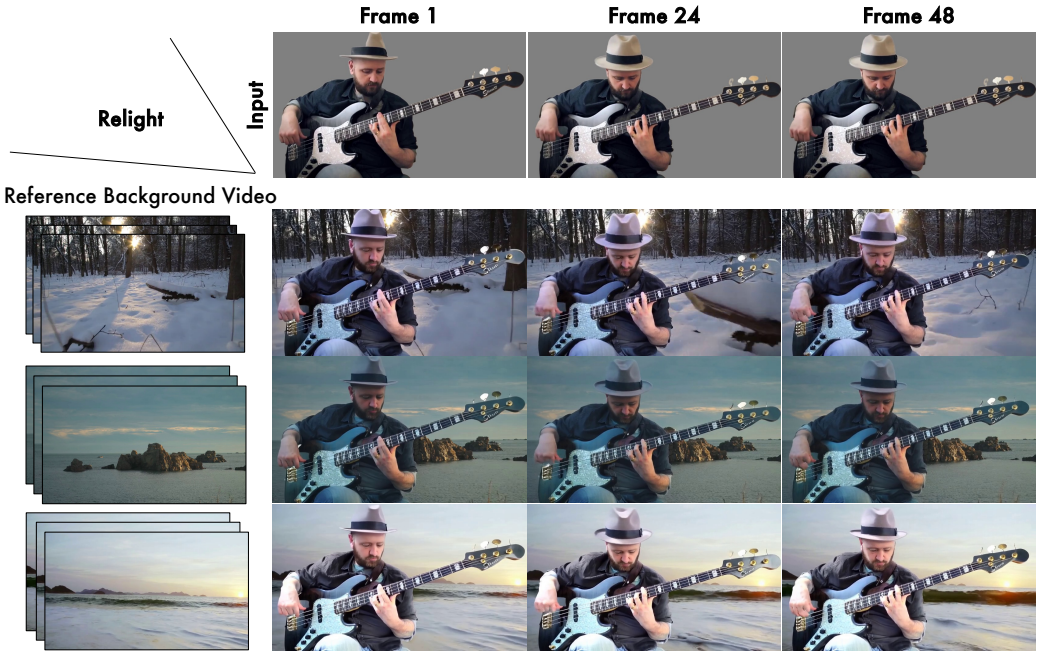

Figure 10: UniLumos performs physically plausible video relighting, conditioned on reference videos.