TC-Light: Temporally Coherent Generative Rendering for Realistic World Transfer

Yang Liu[1,2] **Chuanchen Luo**[3] **Zimo Tang**[6] **Yingyan Li**[1,2] **Yuran Yang**[5]
Yuanyong Ning[5] **Lue Fan**[1,2] **Zhaoxiang Zhang**[1,2*] **Junran Peng**[4*]

[1]NLPR, MAIS, Institute of Automation, Chinese Academy of Sciences,
[2]University of Chinese Academy of Sciences [3]Shandong University
[4]University of Science and Technology Beijing [5]Tencent
[6]Huazhong University of Science and Technology

{liuyang2022, liyingyan2021, lue.fan, zhaoxiang.zhang}@ia.ac.cn
u202315173@hust.edu.cn yangyuran@bupt.edu.cn yyning@tencent.com
chuanchen.luo@sdu.edu.cn jrpeng4ever@126.com

Abstract

Illumination and texture rerendering are critical dimensions for world-to-world transfer, which is valuable for applications including sim2real and real2real visual data scaling up for embodied AI. Existing techniques generatively re-render the input video to realize the transfer, such as video relighting models and conditioned world generation models. Nevertheless, these models are predominantly limited to the domain of training data (e.g., portrait) or fall into the bottleneck of temporal consistency and computation efficiency, especially when the input video involves complex dynamics and long durations. In this paper, we propose **TC-Light**, a novel paradigm characterized by the proposed two-stage post optimization mechanism. Starting from the video preliminarily relighted by an inflated video relighting model, it optimizes appearance embedding in the first stage to align global illumination. Then it optimizes the proposed canonical video representation, i.e., **Unique Video Tensor (UVT)**, to align fine-grained texture and lighting in the second stage. To comprehensively evaluate performance, we also establish a long and highly dynamic video benchmark. Extensive experiments show that our method enables physically plausible re-rendering results with superior temporal coherence and low computation cost. The code and video demos are available at our Project Page.

1 Introduction

Lighting and its interaction with both real and synthetic environments fundamentally shapes how humans—and embodied agents—perceive the world. The ability to re-render the illumination and texture (or so-called relighting) of captured image sequences, especially in complex, highly dynamic scenes, is critically valuable for various world-to-world transfer use cases like filmmaking [49] and augmented reality [35]. Crucially, by re-rendering the CG-simulated or realistic video data used to train embodied agents, it can bridge the sim-to-real gap and enable real-to-real transfer, thus unlocking access to massive high-quality data that is essential for stepping towards embodied intelligence.

Despite its importance, the video illumination and texture rerendering remains a highly challenging problem, particularly when **camera motion is highly dynamic** and **foreground objects frequently enter and exit scenes**, as shown in Fig. 1. Most existing generative relighting techniques [60, 35, 23, 48, 29, 57] are tailored for static images. As shown in Sec. 4.2, naively inflating them to

*Correponding author.

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

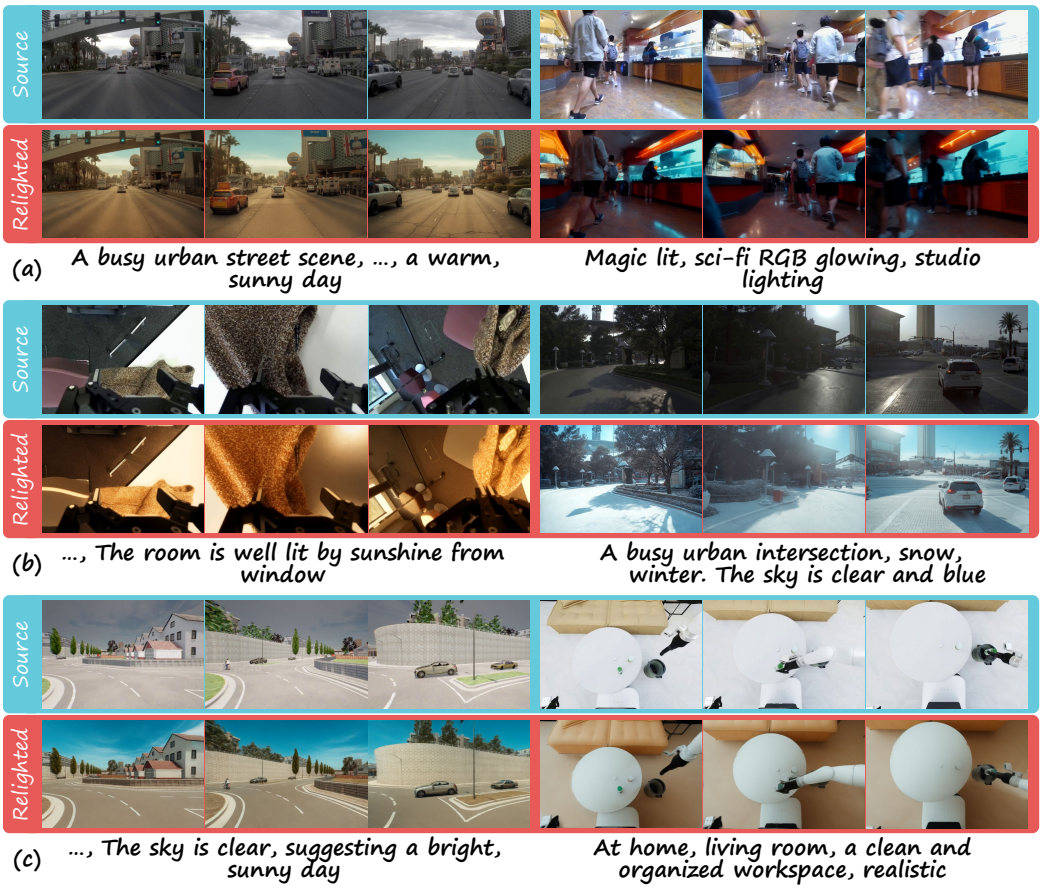

Figure 1: Relighting results on long videos under various dynamic scenes, averaging 256 frames per clip. Though the video involves frequent changes of foreground objects (row **(a)**), highly dynamic camera motions (row **(b)**), the TC-Light realizes consistent and physically plausible relighting results. Row **(c)** also shows its potential to mitigate the sim2real gap for synthetic renderings.

a video model with existing zero-shot strategies struggles to balance the consistency and quality. Moreover, the considerable training cost and scarcity of video lighting datasets hinder fine-tuning a pretrained model for this task. Besides, though generative video relighting and world generation models are emerging, they are either restricted on domain of training data [59, 10, 7, 3] or burdened by considerable computation overhead [64, 17] on long video, as validated in Sec. 4.2.

To address the limitations outlined above, we propose **TC-Light**. We utilize the SOTA image relighting model IC-Light [60] as the baseline, and inflate it to a video model in a zero-shot manner with our decayed multi-axis denoising model, which is distinguished by the proposed Decayed Noise Weighting and Noise Statistic Alignment. It provides a preliminary video relighting result. The core innovation of TC-Light lies in a two-stage post-optimization framework that substantially improves temporal consistency. The first stage introduces per-frame appearance embedding to compensate for exposure discrepancy. It is optimized with photometric loss against the preliminarily relighted video and a flow-based loss between adjacent frames. This enforces global illumination consistency and facilitates consequent optimization. The second stage compresses the output to a canonical representation, i.e., **Unique Video Tensor (UVT)**, according to priors including optical flow and depth of the source video. UVT is then optimized by minimizing the warping error across decompressed frames while aligning the content with the first stage result. As shown in Tab. 2, our optimization procedure is extremely efficient and introduces minimal VRAM overhead.

To comprehensively assess the effectiveness of our model, we introduce a challenging benchmark tailored for complex and highly dynamic scenes. It comprises 58 videos of averagely 256 frames per clip, spanning both indoor and outdoor environments, realistic and synthetic settings, and a

wide range of lighting and weather conditions. Extensive experiments demonstrate that our method achieves high-quality, temporally consistent video relighting while maintaining low computational overhead, highlighting its great potential for downstream applications such as embodied AI. Our main contributions are as follows:

- A novel optimization-based video relighting paradigm for long videos with high and complicated dynamics, significantly improving the temporal consistency of the relighting result.
- We establish a new long-video relighting benchmark characterized by high motion dynamics and broad scene diversity, covering various environments and data domains.
- Extensive experiments validate that our method achieves SOTA performance in producing temporally consistent, naturally relighted videos with minimal computational cost.

2 Related Work

2.1 Learning-based Illumination Editing

Over the past few years, deep neural networks have become one of the main forces behind research in the field of illumination control. Pioneering works [55, 42, 44, 13] train convolutional encoder–decoder networks on light-stage data. The learned prior knowledge enables models to relight a portrait according to the specified light conditions. More recently, large diffusion-based generators have gained popularity for illumination editing. LightIt [32] explicitly conditions the diffusion process on estimated shading and normal maps, giving fine-grained lighting control ability, while SwitchLight [29] incorporates a physics-guided architecture to simulate light-surface interactions better. [63, 4] leverage video foundation models to generate realistic lighting variations over a static image. IC-Light [60], the current state of the art, learns illumination mixture and decomposition from a large quantity of data. Building on these advances in image relighting, video relighting has started to gain traction. [59, 10] learns to disentangle light and intrinsic appearance on portrait videos. [7] represents talking faces as relightable NeRFs guided by predicted albedo and shading features. Extending IC-Light, Light-A-Video [64] introduces zero-shot cross-frame attention modification, while RelightVid [17] trains a temporally inflated IC-Light with a carefully designed video relighting dataset. However, these methods are either restricted to portrait scenarios or struggle with computational efficiency on long videos. In contrast, our model delivers high-quality relighting with strong temporal consistency and low computation cost, even in complex and highly dynamic scenes.

2.2 Diffusion-based Video Editing

The diffusion model [20] has become the go-to model for visual domain transfer and content editing. Based on training paradigms, recent advancements can be grouped into three categories: **(i) training-based** models extend pretrained image diffusion models with temporal layers and are trained on large-scale video datasets, such as [37, 9, 41, 38, 46, 58]. CCEdit [18] and FlowVid [36] further integrate depth and flow cues for improved consistency and control. **(ii) training-free** models mainly rely on cross-frame attention to enforce temporal coherence. TokenFlow [19] and FLATTEN [12] guide attention using estimated optical flow. RAVE [24] enhances latent interactions by denoising over a reorganized latent grid, while Slicedit [11] uses spatiotemporal slices to inject motion priors. VidToMe [34], on the other hand, exploits temporal redundancy through token merging and unmerging. **(iii) one-shot-tuned** models typically learn a canonical video representation in a few iterations and propagate its edits across frames. StableVideo [8] learns to represent video as a foreground and background atlas. CoDeF [43] learns a hash table and decoding MLP to map frames to a single canonical image. Video-3DGS [53] adapts deformable 3DGS [26] to model input video. Our method combines **(ii)** and **(iii)** and proposes an explicit, compact, and efficient canonical representation, i.e., Unique Video Tensor. It enables optimization to be finished within several minutes, which is much faster than 10-30 minutes cost [53] of CoDF and Video-3DGS. Our method also inherits the diffusion model design from training-free algorithms to reduce overall memory and time cost, enabling the processing of long videos.

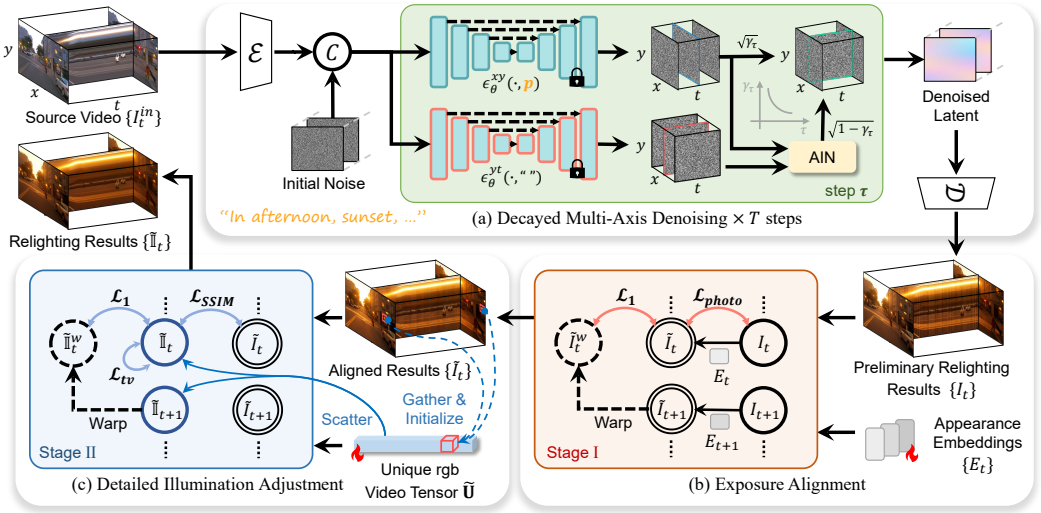

Figure 2: TC-Light overview. Given the source video and text prompt p, the model tokenizes input latents in xy plane and yt plane separately. The predicted noises are adaptively combined together for denoising (cf. Sec. 3.2). Its output then undergoes two-stage optimization to enhance temporal consistency of illumination and texture, which are respectively detailed in Sec. 3.3.1 and Sec. 3.3.2.

3 Method

In this section, we first introduce the task setting and preliminary knowledge about latent diffusion models in Sec. 3.1. Sec. 3.2 further illustrates how our proposed Decayed Noise Weighting and Noise Statistics Alignment helps effectively lift the image relighting model to video space. Sec. 3.3 details how our **key innovation**, i.e., the two-stage temporal consistency optimization strategy, helps align overall illumination and texture appearance.

3.1 Preliminaries

Task Setting. As shown in Fig. 2, we take RGB video as input. The axes of the video space-time volume are denoted by (x, y, t), where xy planes correspond to video frames and yt planes are defined as **spatiotemporal slices** [11]. Since the camera motion is highly dynamic, the target illumination can no longer be simply appointed by a static image or an HDR environment map. Due to superior flexibility and operability, we use textual prompts as the control signal and relight the entire frame.

Latent Diffusion Models (LDMs). Denoising Diffusion Probabilistic Models (DDPMs) [20] are a class of generative models that aim to recover target data distribution through an iterative denoising process. Due to the high computational cost of operating directly in pixel space, LDMs [47, 50, 52] perform diffusion in a lower-dimensional latent space. Given a clean image x_0, and a pretrained autoencoder $\{\mathcal{E}(\cdot), \mathcal{D}(\cdot)\}$, LDMs first encode the image into latent space $z_0 = \mathcal{E}(x_0)$. The forward diffusion process then gradually corrupts z_0 with Gaussian noise ϵ over time steps $\tau = 1, ..., T$

$$z_\tau = \sqrt{\alpha_\tau} z_0 + \sqrt{1 - \alpha_\tau}\epsilon, \tag{1}$$

where $\{\alpha_\tau\}$ is a monotonically decreasing noise schedule. The reverse process begins from pure noise $z_T \sim \mathcal{N}(0, \boldsymbol{I})$. With guidance from control signal (image, text, depth, etc) c, the trained UNet [51] ϵ_θ estimates the noise direction and progressively removes the noise from z_T. After the final denoising step, the estimated clean latent $\hat{z}_0$ is decoded by $\mathcal{D}(\cdot)$ to obtain the generated image $\hat{x_0} = \mathcal{D}(\hat{z}_0)$, which approximates the training distribution.

3.2 Lifting Image Diffusion Model to Video Space

Considering outstand ability in physical plausibility and intrinsic property preservation, we adapt IC-Light [60] into a zero-shot video diffusion model. Concretely, we **(i)** enhance its diffusion blocks

to capture spatiotemporal dependencies and **(ii)** introduce consistency prior from original frames. For **(i)**, we apply the token merging and unmerging technique of VidToMe [34] to self-attention blocks. It divides the video frames into chunks and applies intra-chunk local token merging and inter-chunk global token merging, enabling short- and long-term consistency. The derived model ϵ_θ serves as the basis of **(ii)**. Since it reduces the token count fed to the self-attention module, the computation cost is significantly decreased. For full details, please refer to the original VidToMe paper [34].

For **(ii)**, we apply multi-axis denoising and adapt it with our Decayed Noise Weighting and Noise Statistics Alignment strategy. This modified version is named **decayed multi-axis denoising**. Similar to Slicedit [11], the denoiser has two components with shared weights $\epsilon_\theta^{xy}(\cdot, p)$ that tokenizes each frame and merges tokens from local temporal slots, while $\epsilon_\theta^{yt}(\cdot, \text{“”})$ tokenizes the yt planes (cf. Sec. 3.1) and merges tokens from local image width slot. Note that ϵ_θ^{xy} conditions on target prompt p, while ϵ_θ^{yt} takes empty prompt “” as input (making the denoiser unconditional). The noises separately predicted by two parts according to the same input latents are combined together [11]

$$\epsilon_\theta^V(\cdot, p) = \sqrt{\gamma}\epsilon_\theta^{xy}(\cdot, p) + \sqrt{1-\gamma}\epsilon_\theta^{yt}(\cdot, \text{“”}), \tag{2}$$

where hyperparameter $\gamma \in [0, 1]$ balances effect from ϵ_θ^{yt}. However, the unconditional ϵ_θ^{yt} would overly biases texture and lighting toward the source video, and therefore lead to unnatural relighting results, as validated in Fig. 3 and Fig. 4. To alleviate this problem, we introduce Decayed Noise Weighting, which replaces γ with a timestep-dependent γ_τ that exponentially decays during denoising. To further align predicted noise from ϵ_θ^{yt} to that of ϵ_θ^{xy}, we use Adaptive Instance Normalization (**AIN**) [21] to align noise statistics

$$\epsilon_\theta^V(\cdot, p) = \sqrt{\gamma_\tau}\epsilon_\theta^{xy}(\cdot, p) + \sqrt{1-\gamma_\tau}\widehat{\epsilon_\theta^{yt}}(\cdot, \text{“”}), \tag{3}$$

$$\widehat{\epsilon_\theta^{yt}}(\cdot, \text{“”}) = \sigma_{\epsilon_\theta^{xy}}\left(\frac{\epsilon_\theta^{yt}(\cdot, \text{“”}) - \mu_{\epsilon_\theta^{yt}}}{\sigma_{\epsilon_\theta^{yt}}}\right) + \mu_{\epsilon_\theta^{xy}}, \tag{4}$$

where μ_* and σ_* are the channel-wise mean and standard deviation of each frame. This design preserves motion guidance from the source video while reducing unwanted texture and lighting bias, as validated by ablation studies in Sec. 4.3. The output denoised video is denoted as $\{I_t\}$.

3.3 Post Optimization for Temporal Consistency

Although the video diffusion extension in Sec. 3.2 has introduced spatial-temporal awareness and motion prior from the source video, noticeable illumination and texture flicker persist. To efficiently remove these artifacts, we introduce a two-stage post-optimization framework, as illustrated in parts (b) and (c) of Fig. 2.

3.3.1 Stage I: Exposure Alignment

As shown in part (b) of Fig. 2, the first stage introduces a per-frame appearance embedding E_t to compensate for exposure misalignment between adjacent frames. Inspired by [27], we model E_t as a 3×4 affine transformation matrix, initialized to the identity and optimized via Adam [30]. Its supervision combines a photometric term with a flow-warp alignment term using hyperparameter λ_e

$$\mathcal{L}_{exposure} = (1-\lambda_e)\,\mathcal{L}_{photo}\left(\tilde{I}_t, I_t\right) + \lambda_e\mathcal{L}_1\left(\tilde{I}_t \odot M_t, \text{Warp}_{t+1\to t}\left(\tilde{I}_{t+1}\right) \odot M_t\right), \tag{5}$$

where the homogeneously transformed pixel color $\tilde{I}_t(x, y) = E_t\left[I_t(x, y)\,|1\right]^T$. The photometric loss $\mathcal{L}_{photo}$ is the weighted sum of L1 loss and D-SSIM loss [27], ensuring the transformed frame retains its original content and structure. The second term warps the next frame back to the current timestamp t, according to forward and backward flows $F_{fwd,t}$ and $F_{bwd,t}$ estimated through MemFlow [15] or provided by the dataset. Then it applies an L1 penalty $\mathcal{L}_1$ to align their exposures. To mask out regions with unreliable flow or occlusion, we apply a soft mask M_t

$$M_t = \text{sigmoid}\left(\beta\left(\xi_{flow} - E_{flow}\right)\right) \odot \text{sigmoid}\left(\beta\left(\xi_{rgb} - E_{rgb}\right)\right), \tag{6}$$

$$E_{flow} = \text{Norm}\left(F_{bwd,t} + \text{Warp}_{t-1\to t}\left(F_{fwd,t-1}\right)\right), \quad E_{rgb} = \left|I_t - \text{Warp}_{t+1\to t}\left(I_{t+1}\right)\right|. \tag{7}$$

Here, β is a constant scaling factor, ξ_{flow} and ξ_{rgb} are thresholds set from the statistics of error map E_{flow} and E_{rgb}. This soft mask is also applied in the second stage of optimization. As shown in Tab. 4, soft masking outperforms the hard one in both temporal consistency and prompt alignment.

3.3.2 Stage II: Optimization over Unique Video Tensor

In the second stage, we refine illumination and texture details. Compared with vanilla video, its canonical representation can incorporate spatial-temporal priors and facilitate consistency [43, 53]. But popular NeRF or 3DGS are too complex and costly for learning (cf. Sec. 2.2). Instead, we compress the video to a one-dimensional RGB vector of shape $(N, 3)$, as shown in part (c) of Fig. 2. Specifically, we define a d-dimensional index $\kappa(x, y, t)$ for each pixel based on priors extracted from the source video. An example index could be $[22, 127, 0, 255]$, where the first element is the flow ID (pixels connected by the optical flow predicted by MemFlow share the same flow ID), and the rest are 8-bit quantized RGB values. It is also allowed to extend this 4-element index to more elements with voxel coordinate (from depth projection) or any other cues that indicate spatial–temporal similarity and locality. All pixels with identical κ are gathered via averaging to form one element of the one-dimensional vector, where N is the number of unique κ. Take the source video $\{I_t^{in}\}$ as an example, the **gathering** and **scattering** operations are formulated as

$$\mathbf{U}\left(\kappa_n\right) = \text{Avg}\left(\left\{I_t^{in}\left(x, y\right) | \kappa(x, y, t) = \kappa_n\right\}\right), \quad \mathbb{I}_t^{in}\left(x, y\right) = \mathbf{U}\left(\kappa(x, y, t)\right), \tag{8}$$

where $\mathbf{U}$ is referred to as the **Unique Video Tensor (UVT)**. With an appropriate definition of κ, the scattered $\mathbb{I}_t^{in}\left(x, y\right)$ reconstructs the original $I_t^{in}\left(x, y\right)$ with minimal information loss, as validated in Tab. 5. For relighting, the ideal edited video frames must preserve consistent motion and intrinsic image details with the source; thus, they share the same index tensor κ for UVT representation. Accordingly, we compress the first-stage output $\tilde{I}_t\left(x, y\right)$ into $\tilde{\mathbf{U}}$ via Eq. (8), which then serves as the primary optimization target. This formulation not only facilitates optimization but also naturally embeds spatial-temporal similarity priors (cf. Sec. 4.3). With CUDA parallelism, the gathering and scattering process can be performed instantly. The optimization of $\tilde{\mathbf{U}}$ is supervised by

$$\begin{aligned}\mathcal{L}_{unique} = \lambda_{tv}\mathcal{L}_{tv}\left(\tilde{\mathbb{I}}_t\right) + \left(1 - \lambda_u\right)\mathcal{L}_{SSIM}\left(\tilde{\mathbb{I}}_t, \tilde{I}_t\right) + \\ + \lambda_u\mathcal{L}_1\left(\tilde{\mathbb{I}}_t \odot M_t, \text{Warp}_{t+1\to t}\left(\tilde{\mathbb{I}}_{t+1}\right) \odot M_t\right),\end{aligned} \tag{9}$$

where $\tilde{\mathbb{I}}_t\left(x, y\right) = \tilde{\mathbf{U}}\left(\kappa(x, y, t)\right)$, and λ_{tv} and $\lambda_u \in [0, 1]$ balance the loss terms. The total variation loss $\mathcal{L}_{tv}$ suppresses noise. Notably, Eq. (9) applies SSIM loss instead of photometric loss. This leaves space to fine-grained appearance and illumination adjustment without altering image structure. Finally, the optimized $\widehat{\mathbf{U}}$ is used to reconstruct $\widehat{\mathbb{I}}_t(x, y)$ according to Eq. (8) as the final output.

4 Experiments

4.1 Experiment Setting

Implementation Details. Following IC-Light [60], we apply $T = 25$ sampling steps and a classifier-free guidance scale of 2.0. When inflated to video model with VidToMe [34], the local and global token merging ratios are 0.6 and 0.5, respectively, to accommodate high video dynamics. In our decayed multiaxis denoising strategy, the initial γ_τ is set to 0.2 and decays exponentially to 0.002 until the final sampling step. For the post-optimization stages, we use Adam [31] as optimizer and run 35 epochs in the first stage and 70 in the second with a batch size of 16, ensuring fast yet sufficient convergence. $\kappa(x, y, t)$ mainly contains quantized RGB and estimated masked flow, and optionally depth if provided. Emperically, the weighting coefficients λ_{tv} is set to 0.01, λ_e and λ_u are set to 0.8. Following [26], the learning rate in the first stage decays from 0.01 to 0.001, while the second stage uses a fixed learning rate of 0.05. Additional details are included in the Appendix.

Table 1: Datasets [39, 16, 54, 14, 2, 28, 33, 25] contained in established benchmark. $N_{seq.}$ and $\bar{N}_{frames}$ denote number of sequence and average frames. C, F, D, S respectively denote RGB image, Optical Flow, Depth, Instance Segmentation. Notably, AgiBot here denotes AgiBot Digital World. Due to lacking of extrinsics, its depth is indeed not applicable. Only DRONE is self-collected data.

Datasets	SceneFlow	CARLA	Waymo	NavSim	AgiBot	DROID	InteriorNet	SCAND	DRONE
Agent	Vehicle	Vehicle	Vehicle	Vehicle	Robot	Robot	Robot	Robot	Drone
Synthetic	✓	✓			✓		✓		
Modality	C,F,D,S	C,D,S	C	C	C	C	C,D,S	C	C
$N_{seq.}$	4	8	5	5	8	12	5	6	5
$\bar{N}_{frames}$	300	208	198	250	305	243	300	289	213
Width	960	960	960	960	640	960	640	960	1280
Height	512	536	640	536	480	536	480	536	720

Figure 3: Qualitative comparison of results. The proposed TC-Light avoids unnatural relighting like Slicedit [11] and COSMOS-Transfer1 [3] in (a) and blurring like [3] in (b), or inconsistent illumination like per-frame IC-Light [60] and VidToMe [34] as highlighted by the red squares.

Dataset. To comprehensively evaluate the generation capability, we collect video clips with high motion dynamics and broad scene diversity. This **subjective** evaluation benchmark, as detailed in Tab. 1, covers scenarios like autonomous driving, robot manipulation, and navigation, as well as drone flight. It includes data from synthetic and realistic environments under various weather conditions. Each clip is a long video with on average 256 frames, making it extremely challenging. To provide a more accurate and robust probe on the performance, we also conduct **objective** ground-truth-based evaluation on the Virtual KITTI 2 dataset [6]. We selected five scenes and relit them to match the illumination of morning, sunset, rain, overcast, and fog settings. Each sequence averages 281 frames at a resolution of 1248×384. To obtain edit prompts, we use some prompts from [60] and generate others using COSMOS [1].

Metrics. Following prior works [45, 34, 24, 58], we assess the relighting performance along following four dimensions: (i) *Temporal consistency* is quantified via motion smoothness (**Motion-S**) [22] and structural warping error (**Warp-SSIM**). Motion-S evaluates the continuity and physical plausibility of motion in the edited sequence, whereas Warp-SSIM computes the SSIM between a frame and its warped neighbors using flow from RAFT [56]. (ii) *Textual alignment* is measured by average CLIP embedding similarity between the text prompt and all edited frames (**CLIP-T**). (iii) *User preference* is evaluated by a study on 19 randomly selected videos and 65 valid submissions collected. Participants choose their preferred relighting results among our method and established baselines, from which we derive the Bradley–Terry preference rate (**User-PF**) [5]. Additional details are included in the Appendix. (iv) *Alignment with Groundtruth* is measured using **SSIM** and **LPIPS** [26] with ground

Table 2: Comparison with existing methods. "OOM" here means the method is unable to finish the task due to an out-of-memory error. For a fair comparison, the base models of VidToMe and Slicedit are replaced with IC-Light here. **Ours-light** applies post-optimization to VidToMe, while **Ours-full** further introduces decayed multi-axis denoising. Experiments are conducted on 40G A100. The best and the second best of each metric are separately highlighted in red and blue.

Method	Motion-S↑	WarpSSIM↑	CLIP-T↑	User-PF↑	FPS↑	Time(s)↓	VRAM(G)↓
IC-Light* [60]	94.52%	71.22	0.2743	10.97%	0.123	2075	16.49
VidToMe [34]	95.38%	73.69	0.2731	6.97%	0.409	626	11.65
Slicedit [11]	96.48%	85.37	0.2653	18.39%	0.122	2101	17.87
VideoDirector [58]	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Light-A-Video [64]	OOM	OOM	OOM	OOM	OOM	OOM	OOM
RelightVid [17]	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Cosmos-T1 [3]	96.83%	83.47	0.2529	16.06%	0.101	2543	34.87
Ours-light	97.39%	88.53	0.2700	23.66%	0.359	771	14.36
Ours-full	97.80%	91.75	0.2679	23.96%	0.204	1255	14.37

Table 3: Comparison with existing methods on the Virtual KITTI 2 dataset [6]. The symbol definition aligns with Tab. 2. Experiments are conducted on 40G A100. The best and the second best of each metric are separately highlighted in red and blue.

Method	SSIM↑	LPIPS↓	Motion-S↑	Warp-SSIM↑	Time(s)↓	VRAM(G)↓
IC-Light* [60]	0.5102	0.4470	95.23	68.13	1770	10.25
VidToMe [34]	0.5359	0.4262	95.95	71.33	444	6.96
Slicedit [11]	0.5080	0.4237	96.91	80.74	2346	17.68
VideoDirector [58]	OOM	OOM	OOM	OOM	OOM	OOM
Light-A-Video [64]	OOM	OOM	OOM	OOM	OOM	OOM
RelightVid [17]	OOM	OOM	OOM	OOM	OOM	OOM
Cosmos-T1 [3]	0.4833	0.4841	97.81	84.35	3314	34.83
Ours-light	0.5855	0.4026	98.51	90.94	580	15.21
Ours-full	0.5910	0.3971	98.62	92.38	1002	15.21

truth relighted results. These two metrics replace (ii) and (iii) on Virtual KITTI 2 [6] for a more accurate performance evaluation. (v) *Computation efficiency* is reported in terms of runtime speed (FPS) and peak GPU memory consumption (VRAM) during editing. All experiments are conducted on a 40GB A100 GPU. Additionally, to appraise the reconstruction quality of UVT, we report average PSNR, SSIM, and LPIPS between the original and reconstructed frames.

Baselines. We benchmark our approach against several recent state-of-the-art techniques, whose code is publicly available at the time of writing. These include per-frame IC-Light (denoted as IC-Light*) and its video extensions, Light-A-Video [64] and RelightVid [17]. We also implement two IC-Light variants by incorporating leading zero-shot video editing methods: VidToMe [34] and Slicedit [11]. For fairness, we disable the image downsampling to 512×512 resolution before the diffusion step in Slicedit. In addition, we compare two advanced training-based methods—VideoDirector [58] and COSMOS-Transfer1 [3]. For the latter, due to out-of-memory (OOM) issues when applying full multimodal control on long videos, we employ only its edge branch, which offers a favorable balance between preserving image details and adhering to relighting prompts.

4.2 Comparison with SOTA

Quantitative and qualitative comparisons with state-of-the-art methods are reported in Tab. 2, Tab. 3, and Fig. 3. The result indicates that per-frame relighting (IC-Light*) follows prompts well and produces physically plausible illumination, but the adapted illumination suffers from severe flicker, as shown in columns (a) and (b) of Fig. 3. IC-Light would even randomly hallucinate non-existent objects in textureless regions (cf. column (c) of Fig. 3), further degrading consistency. Extending IC-Light* with VidToMe [34] yields modest gains in temporal coherence but dramatically lowers computation cost for long videos, so we adopt it as our primary baseline. Slicedit [11] significantly suppresses flicker and hallucinations, yet its computation overhead exceeds that of IC-Light*. Besides,

Table 4: Ablation over module component. The experiments here are conducted on CARLA [16] and the Interiornet [33] subset, which both provide depth and instance mask as priors. There are 13 sequences in total and 254 frames on average, covering scenes of indoor and outdoor scenarios. The gray row denotes modification that is aborted and not included in the following experiments.

Method	Motion-S↑	WarpSSIM↑	CLIP-T↑	FPS↑	Time(s)↓	VRAM(G)↓
Baseline	94.51%	77.60	0.2871	0.693	364	10.63
+1st Stage	95.71%	81.29	0.2868	0.651	388	11.33
+2nd Stage(video)	96.40%	90.58	0.2876	0.552	460	13.53
+2nd Stage(UVT)	96.44%	91.04	0.2866	0.563	449	11.81
+soft mask	96.44%	91.05	0.2868	0.559	452	11.81
from scratch(UVT)	96.30%	90.65	0.2866	0.552	458	12.40
+depth	96.56%	91.12	0.2863	0.569	444	11.57
+instance	96.50%	91.01	0.2851	0.545	462	11.67
+multi-axis	98.41%	95.52	0.2813	0.310	805	11.57
+AIN	98.38%	95.44	0.2832	0.310	805	11.57
+weight decay	97.75%	93.74	0.2865	0.310	805	11.57

Table 5: Ablation over Unique Video Tensor (UVT). Here, $\%\mathrm{Cmpr}$ is the compression rate after applying UVT on the source video. The subscripts "f" and "f+d" indicate that, besides color cues, the UVT representation incorporates optical flow cues and both flow and depth cues, respectively.

Scene	$\%\mathrm{Cmpr}_f\downarrow$	$\mathrm{SSIM}_f\uparrow$	$\mathrm{PSNR}_f\uparrow$	$\mathrm{LPIPS}_f\downarrow$	$\#\mathrm{Cmpr}_{f+d}\downarrow$	$\mathrm{SSIM}_{f+d}\uparrow$	$\mathrm{PSNR}_{f+d}\uparrow$	$\mathrm{LPIPS}_{f+d}\downarrow$
CARLA	%39.2	0.9940	50.71	0.025	%29.2	0.9925	48.98	0.028
InteriorNet	%49.0	0.9908	46.17	0.021	%12.8	0.9755	40.86	0.047

its output remains overly biased by the original appearance of the source video. As a result, it produces unnatural relighting in many cases, as shown in column (a) of Fig. 3.

We also evaluated the T2V-model-based video editing approach [58] and concurrent video relighting techniques [64, 17]. Unfortunately, they all failed on long clips due to OOM errors caused by high computation resource demands. For the same reason, Cosmos-Transfer1 [3] can only operate in single-modality mode under GPU constraints, yet still requires over 30 GB GPU memory and more than 30 minutes per clip. Moreover, on video with high dynamics, it suffers from more severe blur and loss of details, as shown in columns (a) and (b) of Fig. 3. These failures are likely because Cosmos-Transfer1 is limited to the data domain of its training data, which contains less varied, moderately dynamic videos.

In contrast, our TC-Light first enables physically plausible relighting on long videos with high dynamics, while outperforming all baselines in temporal consistency and preference rate by a large margin, as shown in Tab. 2. Tab. 3 also demonstrates that our model outperforms all baselines in perceptual and structural similarity with ground truth relighting results. Considering computation cost, the light version adds only 2.4 minutes and 2.7 GB of VRAM overhead compared to the VidToMe baseline (cf. Tab. 2), while faithfully preserving object identity, albedo, and adherence to text prompts, as shown in Fig. 3. Incorporating our decayed multi-axis denoising further enhances temporal coherence, with a modest trade-off in efficiency and quality. Limited by page, we provide additional visualization and performance of different scenarios types in the Appendix. And the video demos and comparison can be found on our project page.

4.3 Ablation

This section analyzes the contribution of each component in our model. The **first stage optimization**, as shown in Tab. 4 and Fig. 4, markedly boosts consistency by aligning cross-frame exposure. The 6-7th rows of Tab. 4 also illustrate that, initializing UVT optimization from the first-stage results converges more efficiently than directly optimizing UVT for the same overall epochs. The **second stage optimization**, as shown in Tab. 4, further reinforces temporal coherence. Tab. 5 confirms that UVT can compress the source video with near-zero loss, which underpins our design in Sec. 3.3.2. Using UVT as the second-stage target not only boosts consistency but also cuts computational overhead. Additionally, replacing a hard mask with a soft mask consistently improves both Warp-

Figure 4: Ablation on main module components. The experiment is conducted on one sequence of the InteriorNet [33] subset, where the text prompt is "This video showcases a modern interior space, which is dimly lit". The baseline here denotes VidToMe [34] in Tab. 2.

SSIM and CLIP-T metrics, demonstrating its importance. Incorporating the depth cues alongside UVT yields a more compact representation (also in Tab. 5), which aids illumination alignment and release computation burden. In contrast, instance segmentation masks provide no clear benefit and are thus omitted from the final implementation.

For the **diffusion module**, multi-axis denoising notably enhances temporal consistency. However, it tends to inherit appearance distribution from the source video, causing drift from the target prompt and sometimes unnatural lighting, as shown in Fig. 4. The introduced AIN and weight decay mitigate these issues, achieving a promising balance between consistency and faithful prompt alignment.

4.4 Limitation and Discussion

Despite achieving impressive results, our method is still limited by its base models. For instance, the current version of IC-Light [60] still struggles to relight hard shadows or make large modifications to low-light images. Similarly, since IC-Light is pretrained on 512 resolution and fine-tuned on 1024 resolution, our model struggles to preserve image details if the resolution is lower than 512. For instance, downscaling the NavSim subset from 960×536 to 480×264 causes Warp-SSIM to drop from the average value 90.46 to 88.36, although CLIP-T remains at 0.304. Besides, since the optimization process relies on the optical flow estimation model, artifacts sometimes occur in textureless areas or under very fast motion, where the flow becomes unreliable. For instance, when downsampling videos from the NavSim subset by 4 times to simulate very fast motion, the CLIP-T and Motion Smoothness fluctuate less than 2%, while Warp-SSIM declines by 5%. Furthermore, the temporal consistency loss has the tendency to smooth the texture of flickering areas, and therefore might sacrifice some details. It can be observed by comparing our model with IC-Light* in part (c) of Fig. 3. Though the proposed decayed multiaxis denoising alleviates the problem, developing a temporally more consistent and computationally more efficient denoising strategy is desired in future work.

5 Conclusion

In summary, we present TC-Light, a one-shot-tuned framework that delivers temporally consistent and physically plausible relighting on long, highly dynamic videos. The optimization-based illumination alignment provides a new paradigm for video relighting. Central to our approach is the Unique Video Tensor—an explicit, canonical, and differentiable video representation that enables highly efficient optimization. Over the established long video relighting benchmark, TC-Light achieves state-of-the-art performance in both consistency and efficiency, endowing it with value and potential for broader application areas such as sim2real and real-world video scaling in embodied AI training and validation pipelines.

Acknowledgments

This work was supported in part by the Guangdong Provincial Foshan Joint Funds (No. 2024A1515110065) and the National Natural Science Foundation of China (No. 62320106010, No. U21B2042).

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[3] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

[4] Shrisha Bharadwaj, Haiwen Feng, Victoria Abrevaya, and Michael J. Black. Genlit: Reformulating single-image relighting as video generation, 2024.

[5] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[6] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

[7] 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.

[8] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023.

[9] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models, 2023.

[10] Jun Myeong Choi, Max Christman, and Roni Sengupta. Personalized video relighting with an at-home light stage. In *European Conference on Computer Vision*, pages 394–410. Springer, 2024.

[11] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9109–9137. PMLR, 21–27 Jul 2024.

[12] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

[13] Sourya Dipta Das, Nisarg A Shah, Saikat Dutta, and Himanshu Kumar. Dsrn: an efficient deep network for image relighting. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2788–2792. IEEE, 2021.

[14] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[15] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19068–19078, 2024.

[16] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[17] 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.

[18] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6712–6722, 2024.

[19] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[22] 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.

[23] 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. In *Advances in Neural Information Processing Systems*, 2024.

[24] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024.

[25] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 2022.

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[27] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024.

[28] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

[29] 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.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR (Poster)*, 2015.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

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

[33] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference (BMVC)*, 2018.

[34] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7495, 2024.

[35] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *European Conference on Computer Vision*, pages 555–572. Springer, 2022.

[36] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2024.

[37] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing, 2023.

[38] Haoyu Ma, Shahin Mahdizadehaghdam, Bichen Wu, Zhipeng Fan, Yuchao Gu, Wenliang Zhao, Lior Shapira, and Xiaohui Xie. Maskint: Video editing via interpolative non-autoregressive masked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7403–7412, 2024.

[39] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[40] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[41] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors, 2023.

[42] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020.

[43] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8089–8099, 2024.

[44] 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.

[45] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[46] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[48] 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.

[49] Christian Richardt, Carsten Stoll, Neil A. Dodgson, Hans-Peter Seidel, and Christian Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of rgbz videos. *Computer Graphics Forum*, 31(2pt1):247–256, May 2012.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[53] Inkyu Shin, Qihang Yu, Xiaohui Shen, In So Kweon, Kuk-Jin Yoon, and Liang-Chieh Chen. Enhancing temporal consistency in video editing by reconstructing videos with 3d gaussian splatting, 2024.

[54] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

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

[56] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[57] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Zhang. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20792–20802, 2023.

[58] Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector: Precise video editing via text-to-video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[59] 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.

[60] 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.

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

[62] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, 2022.

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

[64] 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.

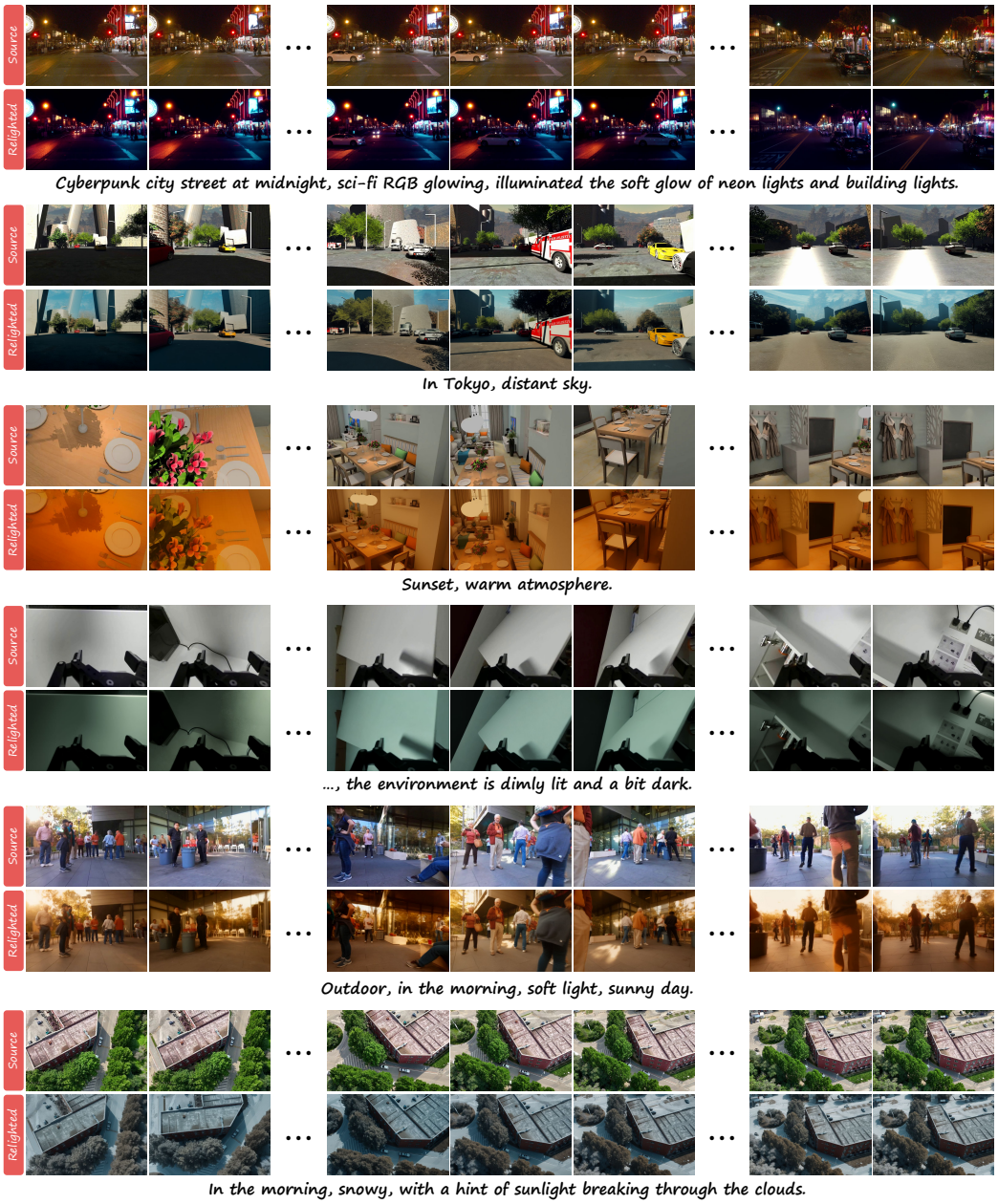

Figure 5: Qualitative results on additional long highly dynamic videos.

Appendix

A Additional Experimental Results

Fig. 5 presents additional visualizations of our relighting results across a diverse range of scenarios. Whether under nighttime or daytime conditions, in outdoor or indoor environments, or from aerial or ground-level viewpoints, the proposed TC-Light method consistently produces temporally coherent and physically plausible illumination edits, demonstrating strong generalization capabilities. It is also worth noticing that the top row demonstrates our model's ability to handle **spatially varying lighting**. In the middle three images, a white car drives from left to right. Initially, it is illuminated by orange street lamps, reflecting an orange hue. As it moves right, its rear remains orange-lit, while the front becomes blue due to a nearby advertising screen. Eventually, the car is fully bathed in blue

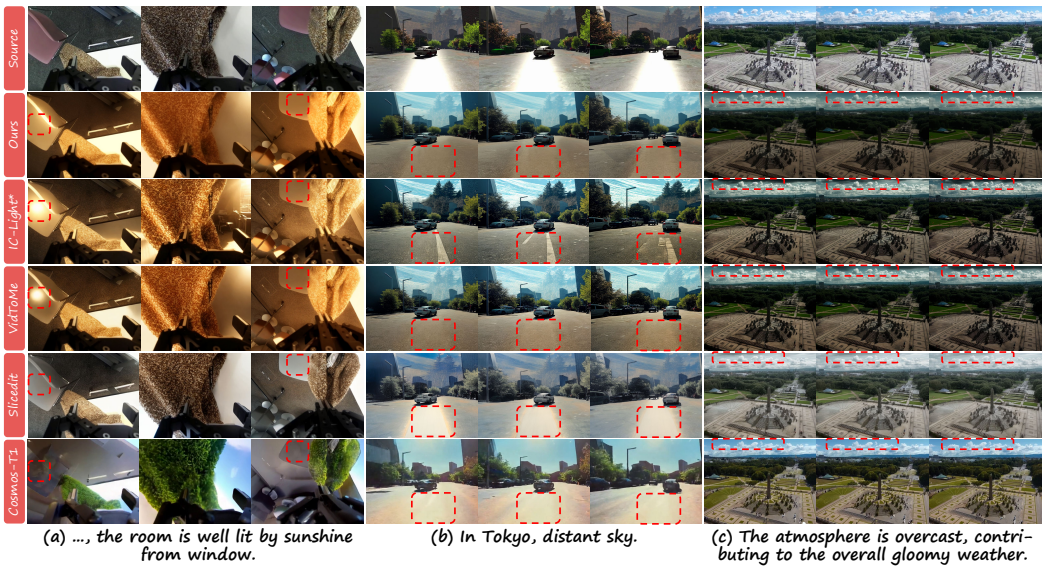

Figure 6: Additional qualitative comparison of results. The proposed TC-Light avoids unnatural relighting like Slicedit [11] and COSMOS-Transfer1 [3] in (a) and (b), or temporal inconsistency like per-frame IC-Light [60] and VidToMe [34] as highlighted by the red squares.

Table 6: Comparison on synthetic [39, 16, 2, 33] and realistic scenarios [54, 14, 28, 25]. The average resolutions are respectively 794×503 and 960×555, while the frame numbers are 272 and 246. "OOM" here means the method is unable to finish the task due to an out-of-memory error. **Ours-light** applies post-optimization to VidToMe, while **Ours-full** further introduces decayed multi-axis denoising. The best and the second best of each metric are separately highlighted in red and blue.

Method	Synthetic			Realistic		
	Motion-S↑	WarpSSIM↑	CLIP-T↑	Motion-S↑	WarpSSIM↑	CLIP-T↑
IC-Light* [60]	93.43%	66.02	0.2779	95.14%	77.13	0.2837
VidToMe [34]	94.61%	69.45	0.2776	95.82%	79.33	0.2815
Slicedit [11]	96.28%	84.90	0.2717	96.38%	88.89	0.2715
VideoDirector [58]	OOM	OOM	OOM	OOM	OOM	OOM
Light-A-Video [64]	OOM	OOM	OOM	OOM	OOM	OOM
RelightVid [17]	OOM	OOM	OOM	OOM	OOM	OOM
Cosmos-T1 [3]	96.31%	80.87	0.2537	96.78%	83.57	0.2659
Ours-light	97.02%	88.63	0.2707	97.46%	89.42	0.2816
Ours-full	97.36%	91.07	0.2695	97.90%	92.67	0.2792

light. This dynamic lighting response indicates our model can correctly handle spatially varying light. Fig. 6 provides qualitative comparisons against state-of-the-art methods across additional scenarios. As shown, our model effectively adheres to textual instructions while generating relighting results that are both natural and temporally consistent.

We also provide corresponding quantitative evaluations on synthetic and real-world scenarios. As reported in Tab. 6, performance on real-world scenes consistently exceeds that on synthetic ones. This discrepancy likely arises from the training data of the video model Cosmos-Transfer1 [3] and the foundational image model IC-Light [60], which are biased towards realistic scenes. Furthermore, the higher resolution and richer textures of real-world data mitigate hallucinations in textureless regions and help better preserve the intrinsic details of source frames for IC-Light. Such attributes are particularly critical for the consistency of methods with comparatively limited temporal modeling, namely, IC-Light* and VidToMe, which exhibit substantially higher Motion-S and WarpSSIM metrics on real-world videos than on synthetic ones. In contrast, our approach attains state-of-the-art temporal consistency across both scenario types while maintaining a favorable balance with prompt adherence.

Table 7: Licenses and video resolution of datasets [39, 16, 54, 14, 2, 28, 33, 25] contained in established benchmark. Notably, AgiBot here denotes AgiBot Digital World. DRONE is our self-collected subset. Sceneflow has no license, but is only allowed for research purposes.

Datasets	SceneFlow	CARLA	Waymo	NavSim	AgiBot	DROID	InteriorNet	SCAND	DRONE
Width	960	960	960	960	640	960	640	960	1280
Height	512	536	640	536	480	536	480	536	720
License	N/A	CC-BY	Custom[2]	CC BY-NC-SA 4.0	CC BY-NC-SA 4.0	Apache -2.0	Custom[3]	CC0 1.0	N/A

B Details of Assets

In Tab. 7, we summarize the license and **resolution** for each subset. All source videos are resized and center-cropped to their designated resolutions. Considering the computation source limitation, we keep all frames if the sequence length is shorter than 300, and randomly sample 300 consecutive frames otherwise. Statistics are provided in Table 1 of the main paper. The DRONE subset includes three clips captured using our DJI Mini4 Pro and two additional clips obtained from DroneStock[4], which are released under the CC0 1.0 License. For AgiBot Digital World [2], where the robot's head moves in coordination with its body while performing tasks, relighting is performed from the head-mounted camera view. For each scene of DROID [28], we apply relighting to both the static side camera and the dynamic left wrist camera views. For Waymo [54] and NavSim [14], relighting is conducted using the front-facing camera view. **Domain balance** is maintained between synthetic and real environments (25 vs. 28 videos), also balanced within sub-domains: autonomous driving (12 synthetic, 10 real), robotic manipulation (8 synthetic, 12 real), indoor navigation (5 synthetic, 6 real). The aerial subset is excluded from balance due to limited long dynamic drone videos in the simulation environment and serves mainly for generalization validation

This paper also benefits from the code of IC-Light [60] (Apache-2.0 License), VidToMe [34] (MIT License), Slicedit [11] (MIT License), VideoDirector [58] (MIT License), Light-A-Video [64] (Apache-2.0 License), RelightVid [17] (CC BY-NC-SA 4.0 License), and Cosmos-Transfer1 [3] (Apache-2.0 License).

C Additional Implementation Details

For competing methods, we adopt the hyperparameters from their official implementations for VideoDirector [58], Light-A-Video [64], RelightVid [17], and Cosmos-T1 [3]. We replace base models of VidToMe [34] and Slicedit [11] with IC-Light [60], and therefore we align their classifier-free guidance scale and diffusion sampling steps with those in [60]. Additionally, we set VidToMe's local and global token-merging ratios to 0.6 and 0.5, respectively, mirroring the setting of our approach. For Slicedit, we adjust the weighting factor γ in Eq. (2) from the default 0.2 to 0.05 to better balance temporal coherence and instruction adherence. All other hyperparameters remain at their default values. The modified VidToMe serves as the baseline of our model design.

During implementation, each $\kappa(x, y, t)$ comprises three components: **(1)** a per-pixel flow ID, **(2)** a quantized RGB color, and, optionally, **(3)** a world-frame voxel coordinate. For (1), flow IDs are derived from the optical flow estimated by the state-of-the-art MemFlow method [15] and the binary mask obtained by thresholding the soft mask in Eq. (6) of the main paper (values > 0.5 are set to 1; otherwise 0). In the initial frame, pixels receive unique flow IDs from 0 to $HW - 1$, where H and W denote image height and width. In subsequent frames, a pixel inherits the flow ID of its predecessor if connected by an unmasked flow; otherwise, it is assigned a new ID. This injects motion priors into the UVT representation. For **(2)**, we quantize RGB values to 7 bits, ensuring that all pixels sharing the same UVT element differ by less than $2/255$ in any channel. This constraint mitigates erroneous flows that escape the mask and reinforces representation in regions exhibiting view-dependent effects. For **(3)**, when per-frame depth maps are available, they are reprojected into a point cloud using the camera intrinsics and extrinsics to determine world-frame coordinates. This point cloud is then voxelized at a specified voxel size, and each pixel's voxel coordinate is appended

[2] `https://waymo.com/open/terms/`

[3] `https://interiornet.org/`

[4] `https://dronestock.com/`

to $\kappa(x, y, t)$, yielding a more compact representation of static regions. For the CARLA [16] and InteriorNet [33] datasets, voxel sizes are set to 0.05 m and 0.02 m, respectively. Notably, dynamic objects at different timestamps may spatially overlap in 3D, but they remain distinguishable by their flow IDs and quantized RGB colors. Consequently, each object at each timestep is represented by a distinct set of UVT elements, while the $\mathcal{L}_1$ temporal consistency loss preserves object identity across frames.

D User Study

We conducted an online user study with 78 anonymous participants, evaluating 19 randomly selected video–text pairs from our datasets. The compared methods were IC-Light* [60], VidToMe [34], Slicedit [11], Cosmos-Transfer1 [3], **Ours-light**, and **Ours-full**. A screenshot of the questionnaire interface is shown in Fig. 7. For each question, methods were anonymized and relighted videos were presented in random order; participants selected the two most preferred results. In compliance with the NeurIPS Code of Ethics, each participant received a compensation of $0.70. Besides, we ensured that all collected data remained confidential and was not disclosed to any institutions or individuals.

Since each video spanned 10–20 seconds, completing the questionnaire took on average 13.5 minutes. Submissions requiring less than four minutes were deemed unreliable and excluded, yielding 65 valid responses. Fig. 8 reports the frequency with which each method was chosen among the top two. Our full model achieved the highest preference rate, while the light variant ranked second. Although IC-Light* and VidToMe follow instructions well (cf. Tab. 2 of the main paper), their inferior temporal consistency make them much less preferred by users. Finally, we computed Bradley–Terry preference scores [5] as a comprehensive metric of user preference, as presented in the Tab. 2 of the main paper.

E Physical Plausibility

This section illustrates how our model maintains physical plausibility. The physical plausibility of our method is inherited from IC-Light [60], which is pre-trained on high-quality Light Stage data and has learned a physically grounded relighting process. Our main contribution lies in improving temporal consistency without altering the illumination priors embedded in the base model.

As detailed in Sec. 3.2, we introduce a video model inflation mechanism based on token merging/unmerging and decayed multi-axis denoising to enable temporal feature-level information exchange. Since these components do not alter the prior knowledge encoded in the base model, the distribution of edited illumination aligns with that of IC-Light while enhancing temporal consistency.

In Section Sec. 3.3, we propose a two-stage post-optimization strategy. The first stage smooths global exposure transitions using an appearance embedding, following practices in physically-based rendering methods like NeRF-W [40] and 3DGS [26]. The second stage refines local fluctuations without altering the overall lighting. Thus, our final results maintain the physically plausible qualities of IC-Light, while significantly improving temporal coherence. As shown in Fig. 3 of the main paper and the video results, our illumination remains qualitatively aligned with IC-Light and VidToMe, but with fewer artifacts and greater temporal stability.

Thanks to the strong priors of IC-Light, our method focuses on temporal coherence and computation efficiency. Compared to optimization-heavy approaches such as Nerfactor [61] and InvRender [62], our post-optimization stage takes only around 2 minutes, with the entire pipeline completing in 10 minutes—substantially faster than training NeRF or 3DGS models, as discussed in Sec. 2.2 of the main paper

F Social Impact

Positive Impacts. The proposed **TC-Light** framework for long video relighting stands to benefit a wide range of applications in both industry and research. First, by enabling consistent and physically plausible illumination editing at low computational cost, it can substantially lower the barrier to high-quality visual content creation, empowering independent filmmakers, educators, and artists to produce compelling video narratives without access to specialized hardware. Second, the capability to scale illumination-diverse training data through sim2real and real2real transfer can accelerate

Questionnaire Survey on Video Relighting Quality

Hello! Thank you for taking your precious time to participate in this video relighting quality questionnaire. This questionnaire aims to collect your preferences for the relighting effects of different algorithms so that we can understand the advantages and disadvantages among different methods. We guarantee that all the collected personal information will not be disclosed.

If your computer is available, it is recommended to browse and answer questions on it to have a better experience.

If you feel some videos take too much time to load, you can first go back to the previous page and then return. We also recommend logging into WeChat to automatically record the form filling progress.

I. Introduction
The questionnaire contains 19 questions. Each question is equipped with one source video, the relighting textual prompt, and six relighted videos from different algorithms. According to the following two dimensions, please select the top two relighting results that best match your preferences (from high to low) :
1. Relighting Effect
o Whether the relighting effect meets the requirements of the textual prompt
o Whether the relighting effect is physically plausible
2. Video Quality
o Visual Quality (no blurring, no distortion)
o Content Alignment (the essence of the original video is not changed)
o Temporal Consistency (no flickering or unnatural transitions)

II. Sorting Instructions
o Select the video clips in order of the best preferred and the second preferred.
o This questionnaire only involves video sorting; no additional text filling is required.

* **1. The following are respectively the original video and the results obtained by editing the lighting according to this prompts "Modern kitchen, matte sink and faucet, shadows cast from the window". Please select the two clips that you think have the highest quality in sequence.**

Original Video:

Videos Relighted by Different Algorithms:

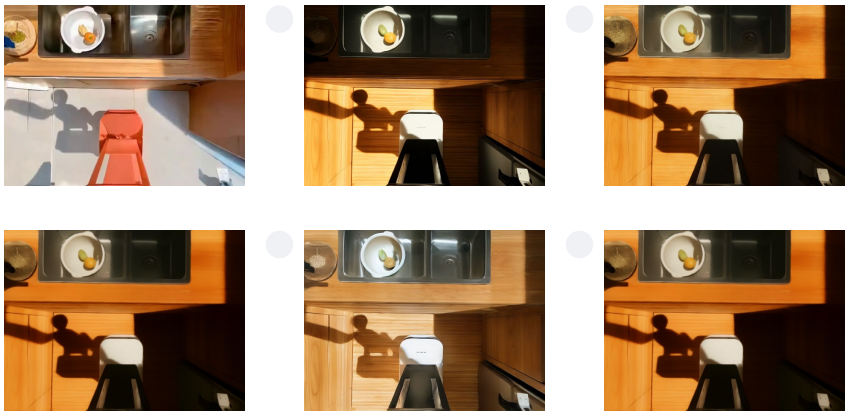

Figure 7: A screenshot of the user study.

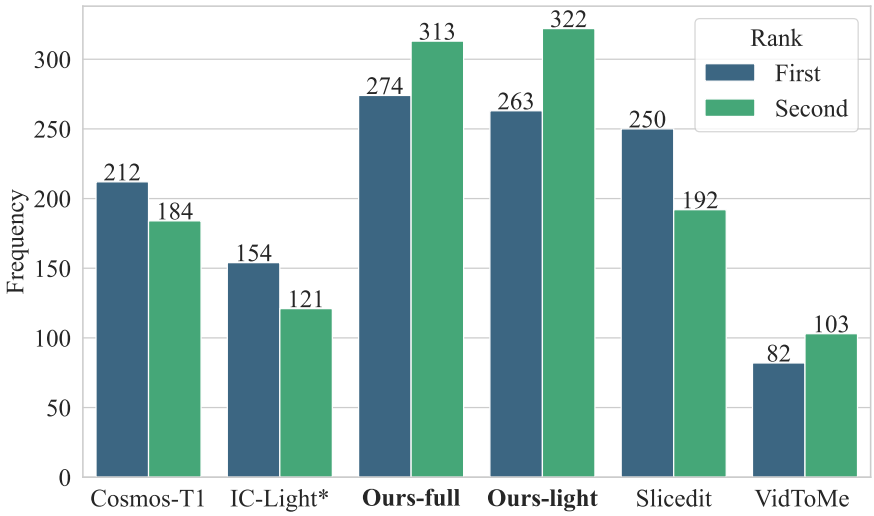

Figure 8: Results from user study with 65 valid submissions. The methods are arranged in alphabetical order. This figure reports the frequency that each method is chosen as the first- and second-most preferred video.

progress in embodied AI—robots and autonomous agents exposed to rich, temporally coherent visual environments may learn more robust perception and planning behaviors, thereby advancing safety and reliability in human–robot interaction. Finally, by fostering more efficient video synthesis pipelines, **TC-Light** may encourage energy-aware design practices in large-scale media processing systems, contributing to reduced resource consumption and attendant carbon emissions.

Negative Impacts. Despite these benefits, improved video relighting carries potential risks if misused. Enhanced realism in dynamic relighting could facilitate the creation of deceptive multimedia, including deepfake videos that manipulate shadows and highlights to conceal tampering or impersonate individuals, thereby eroding trust in digital media. Moreover, large-scale deployment of relighting tools raises privacy concerns: adversarial actors might relight surveillance footage to obscure identities or fabricate altered event sequences. To mitigate these harms, we advocate for gated access to pretrained models, integration of provenance metadata to flag relit content, and collaboration with platform providers to monitor and throttle suspicious bulk relighting requests.

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: The claims are supported by experiments in Sec. 4 both quantitatively and qualitatively. Since the dataset also covers the main application scenarios of embodied agents, we trust the potential of applying our work in embodied AI.

   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: Please refer to Sec. 4.4

   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: [NA]

Justification: No theoretical result is involved.

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: The method is detailedly illustrated in Sec. 3, and implementation details are provided in Sec. 4.1 and Appendix.

   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: [No]

Justification: The full code and dataset is likely to be open-sourced upon acceptance. But the anonymous link to the partial dataset would be provided in the supplementary.

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: Please refer to Sec. 4.1.

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: [No]

Justification: Since both the quantitative experiments in comparison and ablation reports mean metrics over 10 video sequences, we trust that the fluctuation caused by noise is sufficiently suppressed.

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: All the quantitative results are accompanied by the execution time and memory cost. Since GPU is the main compute workers, we provide its details in Sec. 4.1.

   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: As far as we know, there is no break with NeurIPS Code of Ethics.

   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: Due to the page limitation of the main paper, the discussion about positive and negative societal impacts of the work is included 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: [NA]

    Justification: This technique poses no such risks.

    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: Due to the page limitation of the main paper, we list the license of used assets in the Appendix.

    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 include self-collected drone data and introduce it in Sec. 4.1. Due to the page limitation of the main paper, we put details and the anonymized URL of the asset in the Appendix.

    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: [Yes]

    Justification: This paper involves user study. Due to the page limitation of the main paper, the related details are provided in the Appendix.

    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: [Yes]

    Justification: Yes, the potential risks are discussed in the Appendix and are disclosed to the 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.