

HorizonRelight: Relighting Long-horizon Videos Consistently via Diffusion Transformers

Jing Yang^{1,2}, Mayoore Jaiswal¹, Zian Wang¹, Xiao Steven Zeng¹,
Yajie Zhao², Rochelle Pereira¹, and Jianyuan Min^{1*}

¹ NVIDIA ² University of Southern California *Corresponding author

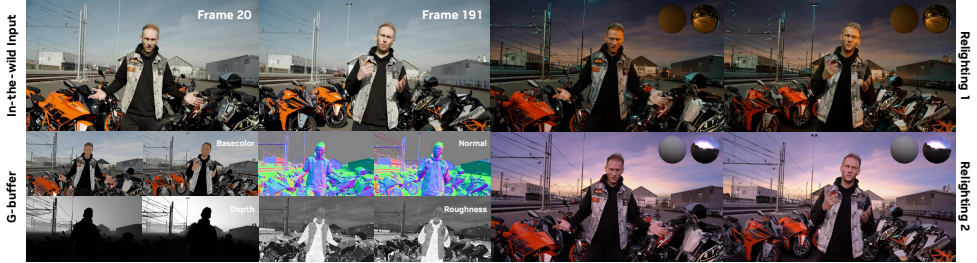


Fig. 1: HorizonRelight delivers temporally consistent long-horizon relighting under a target illumination from a single input video by coupling inverse decomposition with forward re-synthesis, with inverse-estimated G-buffer conditioning preserving content consistency over long horizons.

Abstract. Diffusion-based video relighting enables controllable relighting from a single input video, but modern video diffusion backbones are trained on short clips and applied to long-horizon videos through chunked sliding-window inference, often causing temporal discontinuities at chunk boundaries. We address this by reframing long-horizon relighting as *temporally conditioned latent domain translation*. Our framework enforces cross-chunk continuity by propagating target-domain latents across boundaries and makes this behavior learnable using *masked target-domain self-conditioning*, training the model to continue from temporally masked propagated context. We further introduce *warm-start prompting* with a relit prompt anchor from a controllable generative model, which establishes the initial target-domain state and creates a general interface for prompt-based relighting. Experiments on in-the-wild long-horizon videos show markedly improved temporal consistency, with chunk-boundary artifacts largely reduced and unwanted appearance changes across chunks greatly suppressed. Visit our webpage at <https://research.nvidia.com/labs/sil/projects/horizonrelight/>.

1 Introduction

Relighting is a core capability for modern visual production, enabling post-capture lighting edits for content creation, AR/VR, and visual effects. Recent advances combine inverse decomposition with forward re-synthesis by leveraging video diffusion priors, making controllable relighting possible from a single input video [12, 19]. In particular, DIFFUSIONRENDERER explicitly models both inverse decomposition and forward re-synthesis within a unified diffusion-based pipeline [19], while UniRelight performs end-to-end relighting in a single model [12] and already delivers strong results on short clips on the order of a few seconds (e.g., ~ 2 s). In film practice, takes/shots often span

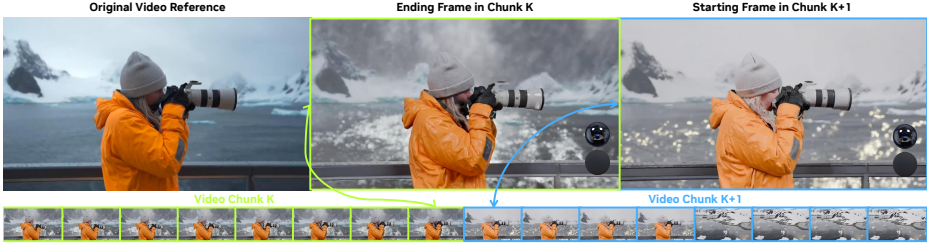


Fig. 2: Temporal inconsistency under chunked long-horizon relighting. Left to right: input frame, then two relit results produced under the same target illumination in two independent chunked runs via [19]. The mirror ball and gray ball indicate the intended illumination, yet the visible differences across runs are enough to break long-horizon consistency.

only a few to a few tens of seconds (average shot length roughly 5–25 s across eras) [8], which already pushes beyond the fixed clip lengths that current diffusion backbones can process reliably at high resolution.

Modern video diffusion backbones rely on spatiotemporal self-attention, whose memory grows rapidly with sequence length and resolution [21, 30]. As a result, training is typically performed on fixed-length clips (“chunks”) that fit GPU memory, and long-horizon video inference is implemented as a sliding-window procedure over time. This is not a modeling preference but a compute constraint. Even systems intended for long-horizon prediction often run as bounded-window continuations to efficiently use GPU resources [25]. Distributed strategies such as context parallelism further reinforce this regime by scaling effective context across GPUs while keeping per-device windows bounded [13]. In the relighting setting, public Cosmos-based diffusion-renderer models are explicitly framed around 57-frame clips, reflecting the practical fixed-window deployment mode [3, 19].

In the compute-constrained chunked regime, relighting pipelines process each chunk largely independently, sometimes with simple overlap or blending [12, 19]. While each window can look sharp in isolation, independence forces the model to re-infer latent causes such as intrinsics and appearance cues at every chunk, creating a train–test mismatch between short-clip training and long sliding-window deployment. As a result, visible differences can appear whenever the window shifts, leading to temporal discontinuities near chunk boundaries and making relighting unreliable for long-horizon videos. Figure 2 visualizes this temporal inconsistency under chunked execution in DIFFUSIONRENDERER [19].

In this work, we address chunk-boundary inconsistency by reframing long-horizon relighting as *temporally conditioned latent domain translation*. Rather than treating each sliding-window chunk as a fresh short-clip generation problem, we model long-horizon relighting as a chained continuation process in the relit target domain. At inference time, each chunk is conditioned on the previous chunk’s ending target-domain latents, so the relit state is continued rather than repeatedly re-inferred, reducing boundary discontinuities and long-horizon drift. To make this continuation behavior learnable, we train with *masked target-domain self-conditioning*, which temporally masks propagated target-domain context during training so the model learns to continue relighting from partial evidence. This learned continuation also enables a principled initialization for

the first chunk, where no propagated target-domain state exists; we therefore introduce *warm-start prompting*. We initialize the forward stage with an externally generated relit prompt anchor, which can be either a single image or a short video from a controllable generative model (e.g., ChatGPT [1], Gemini [26]). This establishes the initial target-domain state for the subsequent chunk chain and exposes a direct interface for prompt-based relighting edits. We evaluate on in-the-wild long-horizon videos and show that the proposed training and inference yield markedly improved temporal consistency, with chunk-boundary artifacts largely removed and unwanted appearance changes across chunks greatly reduced.

2 Related Work

2.1 Video Relighting

Relighting aims to modify scene illumination while preserving scene content and material appearance. A classical line of work approaches this problem through explicit inverse rendering and scene reconstruction, where geometry, reflectance, and lighting are estimated and then combined with physically-based rendering under novel illumination. This paradigm includes calibrated capture systems such as light stages [9] as well as more recent neural inverse-rendering and neural rendering pipelines built on explicit scene representations. While these methods provide strong physical interpretability and control, they often rely on multi-view observations, static-scene assumptions, or per-scene optimization, which limits their applicability to in-the-wild dynamic videos.

Recent learning-based methods instead seek to learn relighting across scenes from data. A representative direction uses intrinsic buffers or latent scene attributes as intermediate structure priors for controllable image synthesis. For example, Deep Shading [22] demonstrated that learned deferred shading can map G-buffers to realistic appearance, while RGB $\leftrightarrow$ X [28] unified intrinsic decomposition and image synthesis in a diffusion-based framework. Building on this line, DiffusionRenderer [19] extends diffusion-based inverse and forward rendering to video, using an explicit two-stage pipeline that first predicts editable scene attributes and then synthesizes relit outputs conditioned on those attributes and the target lighting. This explicit decomposition improves controllability, but it also makes the forward stage sensitive to errors in the intermediate buffers.

Another recent direction uses diffusion models to perform relighting more directly. Prior works such as LightIt [16], DiLightNet [27], Neural Gaffer [14], GenLit [4], IC Light [29], and RelitLRM [31] show that modern generative models can synthesize challenging illumination effects such as cast shadows, specular highlights, and reflections. However, many of these methods are designed for narrower domains, such as portraits, isolated objects, or sparse-view object reconstruction. UniRelight [12] moves toward a more general video setting by jointly modeling relit appearance and albedo in a single diffusion model, avoiding explicit G-buffer bottlenecks and reducing error accumulation between inverse and forward stages.

2.2 Long-horizon Video Diffusion and Chunked Generation

Modern video diffusion models achieve strong visual quality by operating in a compressed latent space, but their spatiotemporal computation remains expensive. As a

result, practical video diffusion systems are typically trained and deployed on bounded clips rather than arbitrarily long sequences. This operating regime underlies latent video diffusion backbones such as Stable Video Diffusion [5] and newer DiT-based video foundation models such as Cosmos [2]. Related video-generation work has explored stronger multi-modal and motion-aware representations, for example by jointly modeling multiple targets in a single diffusion transformer [7, 20]. However, these methods do not directly address relighting consistency when a long video must be generated sequentially across chunk boundaries.

Existing video relighting systems inherit the same bounded-context regime. DiffusionRenderer [19] formulates relighting through inverse and forward video diffusion stages, while UniRelight [12] also builds on a clip-based latent video diffusion backbone. These methods produce strong short-clip results, but they do not explicitly address the deployment setting in which long videos must be processed as a sequence of chunks. In that regime, weak coupling across chunks can lead to repeated re-inference of target-domain appearance and visible discontinuities at chunk boundaries.

2.3 Prompting and Model Adaptivity

Beyond architecture and clip length, another relevant perspective is how pretrained models are conditioned at inference time. In long-horizon video generation under chunked execution, each chunk is produced under incomplete temporal context, making the quality of the available prefix especially important. This connects naturally to a broader literature on prompting and prefix-conditioned behavior, which studies how pretrained models can be steered by conditioning on an explicit prefix rather than by modifying the model itself.

Decoder-only autoregressive language models (e.g., GPT-style models) are widely studied for prompt-based adaptation, where a prefix can steer the model toward a downstream behavior without changing the architecture or updating parameters at inference time [6]. This view is further reinforced by later work on prompt-based and prefix-based adaptation, which shows that conditioning a frozen model with learned prompts or continuous prefixes can serve as an effective alternative to full fine-tuning [17, 18]. By contrast, encoder-only models (e.g., BERT-style models) are pretrained for bidirectional representation learning through objectives such as masked-token prediction, and are typically adapted by attaching task-specific output heads rather than by open-ended continuation from a prompt [10]. This distinction in training objective and adaptation pattern is commonly used to explain why decoder-only models exhibit strong in-context controllability, while encoder-only models are more naturally used for feature extraction and discriminative transfer.

3 Preliminaries

3.1 Video Diffusion Models

We build on a latent video diffusion model (VDM) based on the Diffusion Transformer (DiT) architecture [24], following the Cosmos formulation adopted by recent video

relighting work [3, 12]. Given an RGB video clip $I \in \mathbb{R}^{F \times H \times W \times 3}$, a pretrained VAE encoder E maps it to a latent tensor $z = E(I) \in \mathbb{R}^{F' \times h \times w \times C}$, and a pretrained decoder D maps the latent back to video space. As in UniRelight [12], we use a Cosmos tokenizer, namely *Cosmos-1.0-Tokenizer-CV8x8x8*, that compresses the video along both temporal and spatial dimensions, so typically $F' = F/8$, $h = H/8$, $w = W/8$ and $C = 16$. In our setting, the denoiser is a DiT initialized from a Cosmos video diffusion backbone and fine-tuned for relighting [3, 12].

The diffusion process is performed in latent space under the EDM parameterization [15]. Given a clean latent z_0 , a noisy latent z_t is constructed by adding Gaussian noise according to the EDM noise schedule. A conditional denoiser f_θ predicts the injected noise from z_t under task-specific conditioning inputs cond . We optimize the standard conditional denoising objective

$$\mathcal{L}_{\text{VDM}} = \mathbb{E}_{z_0, \epsilon, t} \left[\|\epsilon - f_\theta(z_t \mid \text{cond})\|_2^2 \right], \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$. At inference time, the model iteratively denoises from Gaussian noise in latent space and decodes the final latent with D .

3.2 Chunked Execution for Long Videos

Due to the high cost of modeling long spatiotemporal token sequences, DiT-based video diffusion models [3, 12, 24] are typically trained and deployed with a fixed temporal window. Given a long video $I^{1:T}$, we partition it into K clips $\{I_k\}_{k=1}^K$, where each $I_k \in \mathbb{R}^{F \times H \times W \times 3}$ denotes a bounded-length chunk processed under this fixed clip budget.

This fixed-window regime is the practical setting for long-horizon video relighting, but it provides only limited temporal coupling across chunk boundaries. As a result, when a long video is handled as a sequence of bounded-window generations, the relit target-domain state may be re-established at each window shift, leading to visible mismatch near chunk boundaries. In the following section, we specify how the conditioning term cond in Eq. (1) is augmented to support cross-chunk continuation.

4 Method

4.1 Problem Formulation

Our objective is to achieve temporally consistent relighting for long-horizon videos. Let the input video be $I^{1:T} \in \mathbb{R}^{T \times H \times W \times 3}$ and let $\ell^{1:T}$ denote the aligned per-frame lighting condition, represented as equirectangular environment maps. We partition the sequence into K non-overlapping chunks $\{(I_k, \ell_k)\}_{k=1}^K$, where each chunk has length F . The task is to generate a relit long video $\hat{I}^{1:T}$ by applying a chunked relighting model $\mathcal{G}_\theta$ sequentially, i.e., $\hat{I}_k = \mathcal{G}_\theta(I_k, \ell_k)$ for $k = 1, \dots, K$.

A key requirement is that neighboring predictions remain consistent across chunk boundaries. We therefore view long-horizon relighting as a *consistent long-horizon latent domain translation* problem, where the last b frames of $\hat{I}_k$ should agree with the first b

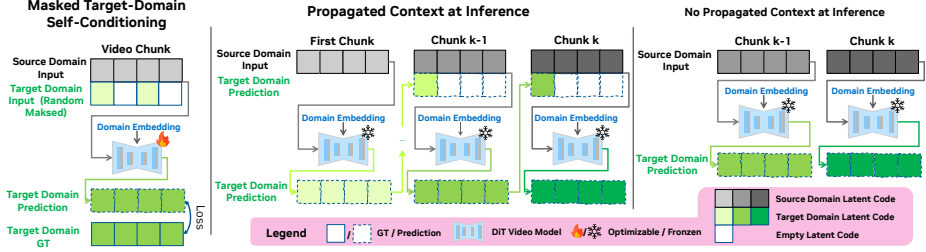


Fig. 3: Masked target-domain self-conditioning and propagated context. During training, a subset of target-domain latent frames is retained as observed anchors while the remaining frames are masked out and predicted. At inference, the observed anchors are provided by the ending latent frames from the previous chunk prediction, enabling chained generation across chunks.

frames of $\hat{I}_{k+1}$ in both content and appearance. Formally, letting $\text{Tail}(\cdot)$ and $\text{Head}(\cdot)$ extract the last and first b frames, respectively, we seek to minimize the seam discrepancy $\Delta_{\text{seam}}(\text{Tail}(\hat{I}_k), \text{Head}(\hat{I}_{k+1}))$ for all adjacent chunk pairs.

4.2 Propagated Context for Consistent long-horizon Latent Domain Translation

For a sequence of chunks, we denote the source-domain latent chunks by $\{z_k^I\}_{k=1}^K$, where $z_k^I = E(I_k) \in \mathbb{R}^{F' \times h \times w \times C}$, and the corresponding target-domain latents by $\{z_k^Y\}_{k=1}^K$. We achieve cross-chunk consistency by training the model to predict target-domain latents from masked target-domain observations, and then reusing the same conditioning structure across chunks at inference time.

Masked Target-domain Self-conditioning. During training, we condition the model on the source-domain latent and a masked target-domain latent, and train it to reconstruct the full target-domain latent, as illustrated in Fig. 3 (left). Concretely, for paired latent chunks (z_k^I, z_k^Y) , we sample an arbitrary temporal mask $m \in \{0, 1\}^{F'}$, where $m_t = 1$ indicates an observed frame and $m_t = 0$ indicates a masked-out frame, and broadcast it to the latent shape, yielding $M \in \{0, 1\}^{F' \times h \times w \times C}$. We then form the masked target-domain latent as $\bar{z}_k^Y = M \odot z_k^Y$, so that only observed target-domain entries are retained. The conditioning input is defined as $\text{cond} = \langle z_k^I, \bar{z}_k^Y \rangle$, yielding

$$\mathcal{L}_{\text{VDM}} = \mathbb{E}_{z_k^Y, \epsilon, t} \left[\left\| \epsilon - f_{\theta}(z_t \mid \langle z_k^I, \bar{z}_k^Y \rangle) \right\|_2^2 \right], \quad (2)$$

where z_t is the noised version of the target-domain latent z_k^Y .

Cross-chunk Context Propagation and Chained Inference. At inference, the same conditioning structure is applied sequentially across chunks, as illustrated in Fig. 3 (middle). For chunk k , we use the previous prediction as the target-domain anchor for the current chunk, with propagated context $s_{k-1} = \text{Tail}(\hat{z}_{k-1}^Y)$, where $\text{Tail}(\cdot)$ extracts the last b frames along the latent temporal axis. For the first chunk, the initial context s_0 can be provided by an external starter anchor; if instead $s_0 = \emptyset$, the generation is cold-started and may degrade in visual fidelity. These propagated frames serve as target-domain

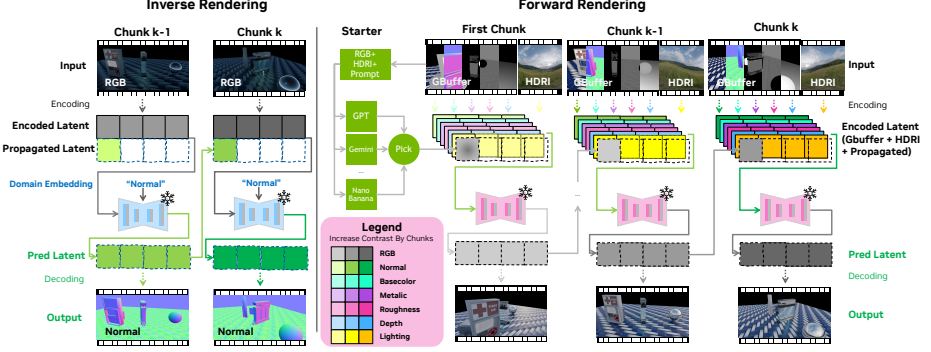


Fig. 4: Applying propagated context for consistent long-horizon relighting. The propagated context is taken as the ending latent frames from the previous chunk prediction and used to condition the next chunk. We apply the same mechanism to both stages: inverse rendering in the intrinsic latent domains and forward rendering in the relit RGB latent domain.

anchors for the next chunk, and the model predicts the remaining target-domain frames conditioned on this partial context. In practice, adjacent chunks can additionally be processed with overlapping frames to further smooth transitions near chunk boundaries. After generating chunk k , we update the propagated context as $s_k = \text{Tail}(\hat{z}_k^Y)$ and use it to condition the next chunk, enabling chained inference across the full sequence.

4.3 Relighting Long-horizon Videos with Warm-start Prompting

Following the DiffusionRenderer architecture [19], we define long-horizon relighting with an inverse stage $\mathcal{F}_\theta^{\text{inv}}$ and a forward stage $\mathcal{F}_\phi^{\text{fwd}}$:

$$\hat{X}_k = \mathcal{F}_\theta^{\text{inv}}(I_k \mid \text{cond}^{\text{inv}}), \quad \hat{I}_k^\ell = \mathcal{F}_\phi^{\text{fwd}}(\hat{X}_k, \ell_k \mid \text{cond}^{\text{fwd}}), \quad (3)$$

where $\hat{X}_k$ is the predicted intrinsic sequence for chunk k , and $\hat{I}_k^\ell$ is the relit RGB output. We use $\mathcal{D}$ for the intrinsic domains, including basecolor, normal, metallic, roughness, depth, and specular. In latent space, z_k^I denotes the input RGB latent, $z_k^{X_d}$ the intrinsic latent for domain $d \in \mathcal{D}$, and z_k^{ℓ} the relit RGB latent. Here, cond^{inv} and cond^{fwd} denote the stage-specific conditioning inputs used in each stage. We visualize the long-horizon relighting architecture in Fig. 4.

Training. During training, both stages operate on fixed-length video clips with masked target-domain self-conditioning. For each intrinsic domain $d \in \mathcal{D}$, we use the corresponding domain embedding e_d to specify the target intrinsic domain, and form a masked intrinsic latent $\bar{z}_k^{X_d} = M \odot z_k^{X_d}$, yielding $\text{cond}^{\text{inv}} = \langle z_k^I, e_d, \bar{z}_k^{X_d} \rangle$ and

$$\mathcal{L}_{\text{VDM}}^{\text{inv}} = \mathbb{E}_{z_k^{X_d}, \epsilon, t} \left[\left\| \epsilon - f_\theta^{\text{inv}}(z_t \mid \langle z_k^I, e_d, \bar{z}_k^{X_d} \rangle) \right\|_2^2 \right]. \quad (4)$$

For the forward stage, let ℓ_k denote the target lighting condition, represented in the same form as prior video relighting frameworks [12, 19]. We then form a masked relit latent

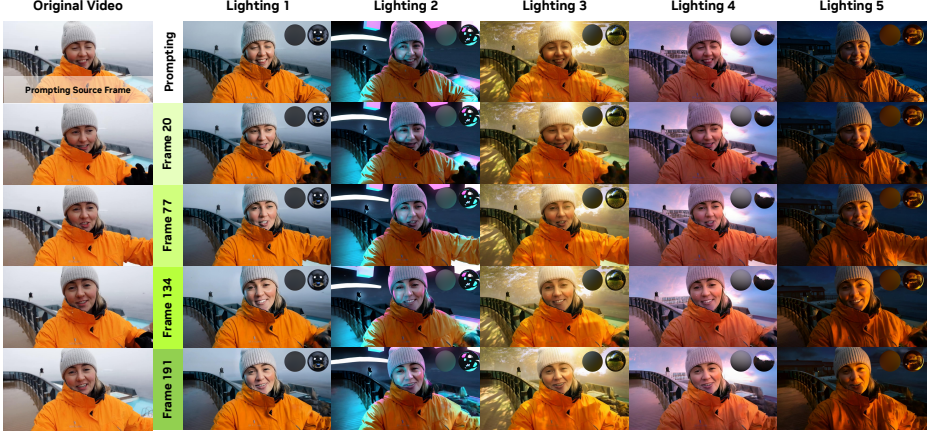


Fig. 5: Long-horizon relighting with warm-start prompting. We use Nano Banana Pro to generate a starter image that initializes the relighting under the target illumination, after which the relit sequence is produced by chained inference across chunks. We visualize frames sampled at the same relative position from consecutive chunks to highlight temporal consistency under different lighting conditions.

$\bar{z}_k^{I^\ell} = M \odot z_k^{I^\ell}$, yielding $\text{cond}^{\text{fwd}} = \langle \{z_k^{X_d}\}_{d \in \mathcal{D}}, \ell_k, \bar{z}_k^{I^\ell} \rangle$ and

$$\mathcal{L}_{\text{VDM}}^{\text{fwd}} = \mathbb{E}_{z_k^{I^\ell}, \epsilon, t} \left[\left\| \epsilon - f_\phi^{\text{fwd}}(z_t \mid \langle \{z_k^{X_d}\}_{d \in \mathcal{D}}, \ell_k, \bar{z}_k^{I^\ell} \rangle) \right\|_2^2 \right]. \quad (5)$$

Here, M is sampled by an arbitrary temporal mask and z_t denotes the noised target latent in the corresponding domain.

Warm-start Prompting. Warm-start prompting follows the same intuition as prompt-based adaptation in GPT-style models: a prefix defines the task and steers generation without fine-tuning the model itself. In the relighting setting, the prefix is a relit starter anchor, provided as a single image or a well-aligned short video (e.g., a relit clip from [19]). This starter anchor fits naturally into our continuation design: it initializes the chunk chain by defining the desired target-domain appearance that the model should start from and continue under. In the inverse stage, we estimate intrinsic factors and G-buffers for forward-stage content conditioning, with the first chunk also cold-started. In our forward stage, the relit latent anchor initializes the target-domain appearance, the predicted intrinsic latents $\{z_k^{X_d}\}_{d \in \mathcal{D}}$ preserve scene content, and the lighting condition ℓ_k specifies the target illumination. The starter anchor serves as a visual prompt that initializes the chain, after which generation proceeds under the same continuation process, and it can be produced by any controllable external model.

5 Experiments

5.1 Dataset

Synthetic Dataset. For training and controlled evaluation, we use the same synthetic data generation setup as DiffusionRenderer [19] and UniRelight [12]. The dataset is

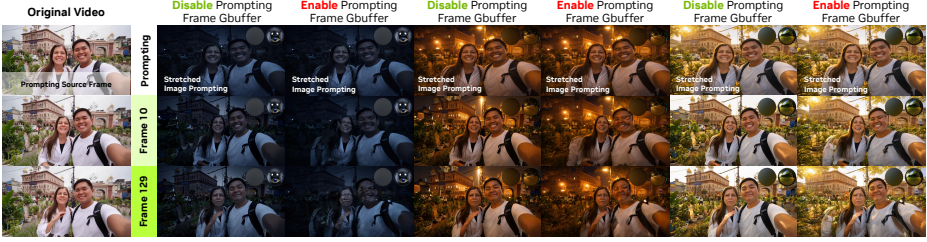


Fig. 6: Ablation on stretched warm-start prompting. When the starter image is not well aligned with the predicted structural cues, directly conditioning on the corresponding G-buffer can introduce blur or content mixing. Disabling the warm-start G-buffer for the prompted frame preserves the initialization benefit while avoiding misaligned guidance.

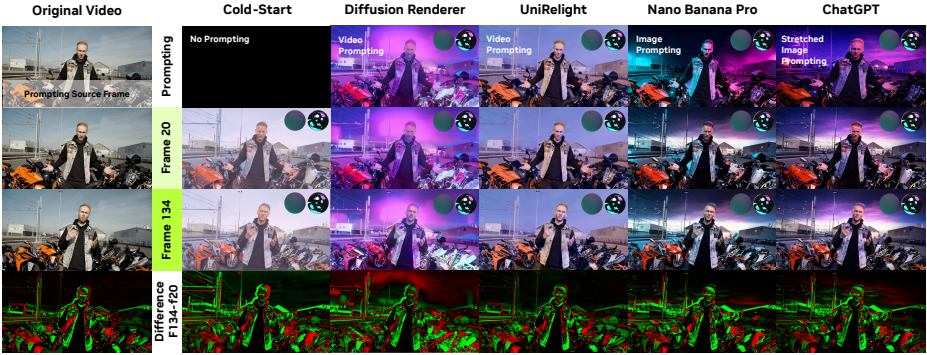


Fig. 7: Generality of warm-start prompting. Our framework supports multiple initialization modes, including cold start, video-based prompting, and image-based prompting. We compare frames across chunks to show that different prompt sources support stable long-horizon generation. Difference maps highlight true scene dynamics while remaining largely black in static regions, indicating preserved scene and lighting consistency across chunks.

procedurally rendered from scenes built with 36.5K Objaverse LVIS objects, 4,260 PBR materials, and 766 HDRI environment maps, with randomized camera/object/lighting motion and additional primitives to increase geometric and lighting diversity. In our rendering pipeline, each video clip is rendered as a 121-frame sequence. For training, we follow the DiffusionRenderer/UniRelight convention and use the first 57 frames from each clip.

In-the-wild YouTube C.C. To evaluate generalization beyond synthetic data, we additionally curate an in-the-wild benchmark from YouTube Creative Commons videos. We collect around 100 arbitrary samples using broad keywords such as “vlog,” “travel long take,” and “film long take.” The resulting set covers diverse camera motion, object motion, and scene layouts, and is used primarily for qualitative evaluation of long-horizon temporal consistency and warm-start behavior under practical deployment conditions.

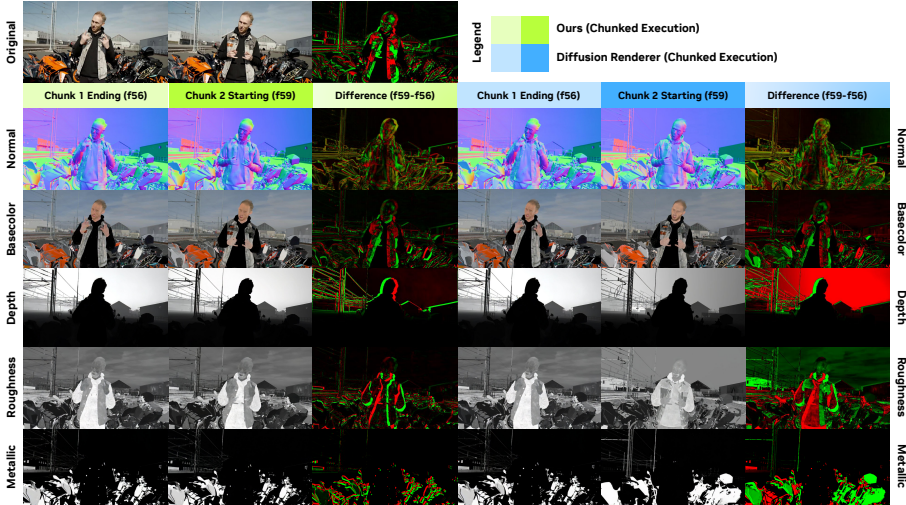


Fig. 8: Temporal consistency of long-horizon inverse decomposition. We compare predictions around chunk boundaries using frame differences to reveal cross-chunk variation. Our method maintains clean, localized changes near true scene dynamics, indicating substantially improved inverse-stage consistency.

5.2 Qualitative

Relighting Long-horizon Videos. We present relighting results in Fig. 5, where the warm-start prompting uses a high-quality image generated by Nano Banana Pro [11] that is spatially consistent with the input frame. To obtain this warm-start image, we use the first frame of the input video and the target HDRI as inputs to Nano Banana Pro, with the prompt: “*Relight the first image using the second image as the environment lighting. Update the lighting for both the foreground and background, while preserving the content, resolution, and aspect ratio of the first image. Use the second image only for lighting, and do not transfer its content.*” This warm-start image provides an explicit target-domain anchor for the first chunk, after which our model maintains the relit appearance through chained latent propagation across later chunks. As shown in Fig. 5, this produces visually stable relighting across long horizons, with consistent shading and appearance preserved across consecutive chunks under the same target illumination.

In practice, however, the warm-start image often does not exactly match the input frame layout. The external editor may slightly change the framing, or return a stretched image that no longer matches the predicted G-buffer intrinsics for that frame. We therefore study this practical case in Fig. 6. Even under this mismatch, the warm-start image can still provide a useful target-domain initialization. The main issue is the conflict between the prompted RGB image and the predicted G-buffer intrinsics used by the forward stage. If both are used together when they are spatially inconsistent, the model receives incompatible cues, which can lead to blur or content mixing. Disabling the warm-start G-buffer for that frame removes this conflict and allows the model to retain the benefit of the appearance prior. This shows that warm-start prompting remains effective even when the starter image is imperfect.

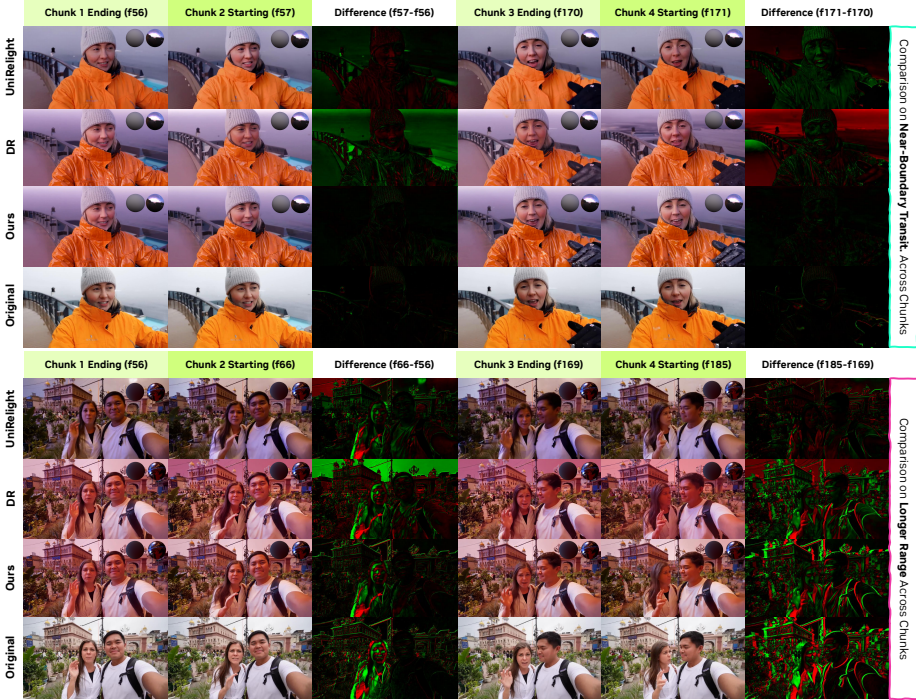


Fig. 9: Temporal consistency of long-horizon forward rendering. We compare both near-boundary transitions and longer-range frame dynamics across chunks. Near-boundary comparisons should show only thin differences around true motion, while longer-range comparisons naturally produce thicker dynamic regions due to accumulated camera and object motion. Our results follow this expected pattern and remain close to the input-video difference maps, whereas the baselines exhibit broader changes in non-moving regions, indicating blur, flicker, and cross-chunk drift.

We then examine the generality of the warm-start choice in Fig. 7. Our framework is not tied to a single initialization source: besides the standard warm-start image, it also supports cold start, video-based prompting, and alternative image-based prompts. Cold start is feasible, but without an explicit target-domain anchor it relies more heavily on the relighting prior learned from the training data, which can weaken strong relighting cues such as specular detail. As shown by the ablation in Fig. 6, misalignment between the prompt and the predicted G-buffer intrinsics can introduce conflicting cues. This makes video prompting less flexible in practice, since it requires stronger spatial alignment with the target sequence, whereas image-based prompting is more robust because any mismatch is confined to the warm-start frame and can be mitigated by disabling the corresponding G-buffer. In Fig. 7, we further show that warm-start images generated by both Nano Banana [11] and ChatGPT 4o [23] can be used within the same framework.

Cross-Chunk Consistency. We next evaluate the temporal consistency enabled by the chained design. Fig. 8 shows consistency in the inverse stage, and Fig. 9 shows consistency in the forward relighting stage, including both near-boundary transitions and longer-range comparisons across chunks.



Fig. 10: Long-horizon editing. The same warm-start mechanism can be used not only for relighting, but also for temporally consistent appearance and style edits over long-horizon videos.

We compare against DiffusionRenderer (DR) [19] and UniRelight [12] on in-the-wild YouTube samples. Because chunk-boundary discontinuities are the dominant failure mode in long-horizon chunked inference, we visualize frame differences to determine whether each method captures true scene dynamics or introduces chunk-specific appearance changes. Ideally, the difference maps should respond mainly in regions of actual motion, while non-moving regions remain largely black. Responses outside these dynamic regions indicate temporal drift or inconsistent re-inference across chunks.

In the inverse stage (Fig. 8), the baselines often show broader appearance changes across chunk boundaries, including responses in regions that should remain static. This indicates that scene factors are being re-estimated inconsistently from one chunk to the next. By contrast, our method keeps the differences concentrated near the true motion regions, while the rest of the frame remains much more stable. This indicates reduced cross-chunk drift in the inverse predictions. For the inverse stage, our method uses DiffusionRenderer to initialize the first chunk. The consistency gain observed at later chunk boundaries therefore comes from the chained propagation across subsequent chunks, rather than from a different initialization at the start.

The same pattern becomes even clearer in the forward stage (Fig. 9), where we compare both near-boundary transitions and longer-range dynamics across chunks. In the near-boundary comparisons, the motion between adjacent frames is usually very small, so the difference maps should show only thin boundaries around moving objects or slight camera motion, with the rest of the frame remaining black. Our results follow this expected pattern and closely match the input-video difference maps, while the baselines already show broader changes in non-moving regions. In the longer-range comparisons, the accumulated motion from camera movement and object motion becomes more visible, so the dynamic regions in the difference maps are naturally thicker. Again, our results remain consistent with the input-video difference maps by highlighting these true dynamic regions, while the baselines continue to respond strongly in areas that should



Fig. 11: Long-horizon limitations. Beyond sufficiently long chains, the relighting remains preserved, but scene content gradually loses fine details.

remain stable. This leads to blur, flicker, and appearance drift. Overall, these results show that our method captures the true scene dynamics while maintaining scene and lighting consistency in non-moving regions across both short- and long-range comparisons.

Long-horizon Editing. Finally, Fig. 10 shows that the same warm-start prompting mechanism supports a broader class of long-horizon generation tasks beyond relighting. When the warm-start reference is aligned with the input structural conditions, the prompted appearance can be preserved and propagated throughout the sequence. This remains effective even when the prompted appearance is not explicitly represented in the training data. Rather than reproducing only familiar object identities or styles, the model learns to extend the prompted target-domain distribution over time while maintaining temporal consistency in scene content and lighting. As a result, the same mechanism can animate a single edited image into a coherent long-horizon sequence, enabling temporally consistent style transfer, content editing, and relighting within one unified framework.

This behavior is especially relevant to practical film production. In many studio workflows, scenes are captured under controlled conditions such as green-screen or blue-screen setups, while much of the final look is determined later during visual effects and lighting design. In this setting, our warm-start prompting provides a natural way to inject a desired appearance at the beginning of the sequence and propagate it consistently across time. Figure 10 illustrates this capability with examples of both subject-level edits and global style changes that remain coherent throughout the long-horizon videos.

Limitations. As visualized in Fig. 11, a limitation is fine-detail degradation at extremely long horizons. While chained propagation improves relighting consistency over long-horizon videos, the output quality remains bounded by the learned appearance prior. In our experiments, beyond roughly five chunks (about 300 frames), relighting stays stable, but scene content gradually loses high-frequency details. Extending the usable horizon will likely require training data and supervision targets that better cover the long-horizon appearance space encountered in practice.

5.3 Quantitative

We quantitatively evaluate cross-chunk temporal consistency for both inverse decomposition and forward relighting on the synthetic dataset. For this evaluation, we use the last 57 frames of each rendered 121-frame video in the testing set. To construct a controlled cross-chunk setting, we form two consecutive 57-frame chunks by taking the selected 57-frame sequence and concatenating it with its duplicated reversed copy. Under this

Table 1: Quantitative temporal consistency on the synthetic dataset. We report MSE under the controlled repeated-chunk setting. *Boundary* is computed only on the overlapping frames across the chunk boundary, while *Sequence* is computed over the full predicted chunk. Lower is better.

Method	Scope	Normal	BaseColor	Depth	Roughness	Metallic	Specular	Relighting
DR [19]	Boundary	0.1379	0.1613	0.1576	0.2614	0.2050	0.1205	0.1048
Ours	Boundary	0.0971	0.1173	0.1181	0.1042	0.0542	0.0935	0.0729
DR [19]	Sequence	0.1425	0.1670	0.2208	0.2664	0.2409	0.1242	0.1063
Ours	Sequence	0.1007	0.1325	0.1383	0.1199	0.1088	0.1137	0.0906

construction, the overlapping frames across the chunk boundary are exactly identical in content, so any prediction difference across the boundary directly reflects temporal inconsistency introduced by chunked inference rather than true scene motion.

We report mean squared error (MSE) in two ways to complement the qualitative analysis. *Boundary* MSE is computed on the overlapping frames across a chunk transition, directly measuring boundary consistency. *Sequence* MSE is computed over the full predicted chunk, measuring how inconsistency propagates beyond the overlap. We evaluate both inverse decomposition outputs and the final relighting output. Table 1 shows that our method consistently improves both metrics across all inverse G-buffer prediction and the final relighting output.

5.4 Implementation Details

Our models are initialized from *Cosmos-Predict1-7B-Video2World* [2] pretrained weights and fine-tuned on our relighting training data using a Cosmos-based DiT video diffusion backbone. We train the forward rendering (relighting) model and the inverse decomposition model separately under the same hardware setup: the forward model is trained for approximately one month on 16 NVIDIA H100 GPUs, and the inverse model is trained for approximately two weeks on 16 NVIDIA H100 GPUs. Training is performed in mixed precision (BF16) with AdamW and latent-space conditioning, where all inputs (video, intrinsics, and lighting cues) are encoded into tokens and concatenated with type embeddings. We employ *masked target-domain self-conditioning*: during training, we randomly sample temporal masks over the target-domain latents to expose the model to arbitrary partial-context continuation, while at inference we keep only the first frame unmasked and predict the remaining frames conditioned on this initial anchor.

6 Conclusion

We presented a long-horizon video relighting framework that improves cross-chunk consistency by reframing chunked execution as temporally conditioned latent domain translation, combining chained propagation with warm-start prompting. Experiments on both synthetic and in-the-wild videos show temporally consistent inverse decomposition for forward relighting. The main limitation is long-horizon detail degradation: while relighting remains stable, fine details weaken as the chain grows longer. Future work includes reducing this quality drop and exploring long-horizon free-form creative editing via prompt anchors.

Supplementary

We thank the reviewers for constructive feedback and denote jk2v/2LMR/xb7 as **R1/R2/R3**.

R1,R2,R3 Does HorizonRelight improve temporal stability over UniRelight? We extend main-paper Table 1 to 100 in-the-wild YouTube clips (Table 1). With no target-light ground truth, boundary drift is the temporal signal: *continuous* scores adjacent windows; *duplicate-reverse* matches boundary source frames to isolate generation drift. We stitch out-puts from each method’s native window and score the same boundaries.

Table 1. Relit RGB boundary metrics on 100 in-the-wild YouTube clips.

Method	Window	Continuous test		Duplicate-reverse test	
		MSE↓	SSIM↑	MSE↓	SSIM↑
DR	57	0.0327	0.670	0.0209	0.718
UniRelight	57	0.0131	0.783	0.00912	0.837
TC-Light	57	0.00127	0.937	0.00148	0.911
Light-A-Video	32	0.00623	0.853	0.00149	0.926
RelightVid	16	0.00763	0.760	0.00336	0.861
Ours-no-anc.	57	0.0254	0.679	0.0160	0.734
Ours	57	0.00080	0.934	0.00048	0.954

Ours-no-anc. removes the target-domain anchor but keeps source/G-buffer conditioning. Adding the anchor cuts boundary MSE by 94.0/94.7% vs. UniRelight, 97.6/97.7% vs. DR, and also improves over TC-Light/LAV/RelightVid by 37.0/87.2/89.5% (continuous) and 67.6/67.8/85.7% (duplicate-reverse).

R1,R2,R3 Direct RGB relighting baseline comparison. Because YouTube clips lack target-light ground truth, we measure fidelity on paired Multi-Illumination: 16-frame non-target inputs reconstruct held-out target lighting. Warm-start uses a UniRelight first frame as appearance cue, not target truth.

Table 2. Held-out target-light reconstruction on paired Multi-Illumination.

Method	PSNR↑	SSIM↑	RGB MAE↓
Ours-cold	12.80	0.594	0.197
Ours-warm	14.39	0.679	0.163
DR	12.39	0.611	0.197
UniRelight	14.72	0.708	0.157
TC-Light	7.58	0.255	0.378
Light-A-Video	9.42	0.469	0.282
RelightVid	14.42	0.666	0.145

Table 2 addresses the direct-baseline request: Ours-warm is within 0.33 dB/0.029 SSIM/0.006 MAE of UniRelight, within 0.03 dB of RelightVid while exceeding its SSIM, and substantially outperforms TC-Light/LAV. Table 1 and Fig. 1 and the supplement cover the central long-video boundary claim.

R1,R2,R3 HorizonRelight is not autoregressive video generation. Autoregressive generation uses previous outputs as content. HorizonRelight re-controls each chunk with source video/G-buffers; previous relit output is only a boundary color cue, so geometry, motion, and texture do not roll out. Thus scheduled sampling/self-generated histories would train recursive content rollout, unlike our observed-source translation; Table 3 summarizes the contrast.

R2,R3 Long-horizon scope and main-paper Fig. 11. Our claim is practical long-take relighting, not unrestricted 800-frame future generation. As motivated in the introduction, common film shots last roughly 5–25s, while current relighting backbones use short fixed windows. Figure 1 supports this scope: boundary error stays low across stitched chunks, and source-structure correlation (source-output gradient correlation) shows gradual quality decay rather than abrupt drift. Main-paper Fig. 11 reflects a data-coverage limitation from synthetic plus auto-labeled real training; in practical shot-based editing, the chain resets at shot changes, and periodic re-anchoring handles unusually long continuous takes.

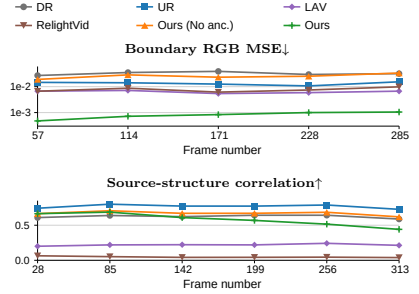


Figure 1. YouTube long-horizon boundary MSE and source-structure correlation.

R1,R2,R3 Target-domain self-conditioning and propagation. Main-paper Fig. 3 ties training masks to inference: source/G-buffer control plus sparse relit cues, similar in spirit to MAE. The inference mask is a trained subcase, not a new mode, so it does not introduce a train-test mismatch. Removing this target-domain bridge breaks translation; removing propagation leaves DR/UniRelight-style boundary jumps, as shown in main-paper Figs. 8–9 and the supplement.

R2,R3 Warm start, masks, and overlap. Warm-start only supplies the first appearance cue because chunk 1 has no previous relit boundary; main-paper Figs. 6–7 compare cold/video/image starters and G-buffer conflict. After chunk 1, content remains source/G-buffer controlled; the propagated relit frame stabilizes appearance rather than geometry. The starter is interchangeable and can use open-source relighters/starters; style editing is a byproduct, not the core claim. The mask ratio follows Cosmos convention, and tail-overlap/blending cannot fix the discontinuities in main-paper Figs. 2/9.

R2 Training asymmetry. Forward relighting has a larger target space: RGB→intrinsic predicts fixed maps, while relighting must model continuously varying target-light appearance, explaining longer training. Both stages use the same propagation design; we will report $b = 1$, schedules, window sizes, and the clip/metric protocol.

Table 3. Baseline Difference: cross-chunk control.

Method	Content source	Consistency cue	AR?
Rolling Forcing	generated history	rolling noise/sink	✓
Resampling Forcing	self-resampled history	history routing	✓
B-Agger	generated rollouts	backward correction	✓
Diffusion Forcing	generated tokens/frames	causal noise schedule	✓
StreamingT2V	previous features	short/long memory	✓
FreeNoise	prompt/noise latents	noise rescheduling	×
ChronoDepth	predicted depth	per-frame noise levels	✓
Ours	source video + G-buffers	relit color cue	×

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
3. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025)
4. Bharadwaj, S., Feng, H., Becherini, G., Fernandez Abrevaya, V., Black, M.J.: Genlit: Reformulating single-image relighting as video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
7. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. arXiv preprint arXiv:2502.02492 (2025)
8. Cutting, J.E., Brunick, K.L., DeLong, J.E., Iricinschi, C., Candan, A.: Quicker, faster, darker: Changes in hollywood film over 75 years. *i-Perception* **2**(6), 569–576 (2011)
9. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 145–156 (2000)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
11. Google: Nano banana image generation — gemini api documentation. <https://ai.google.dev/gemini-api/docs/image-generation>, accessed: 2026-03-04
12. He, K., Liang, R., Munkberg, J., Hasselgren, J., Vijaykumar, N., Keller, A., Fidler, S., Gilitschenski, I., Gojcic, Z., Wang, Z.: Unirelight: Learning joint decomposition and synthesis for video relighting. arXiv preprint arXiv:2506.15673 (2025)
13. Jiang, C., Cai, Z., Tian, Y., Jia, Z., Wang, Y., Wu, C.: Dcp: Addressing input dynamism in long-context training via dynamic context parallelism. In: Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles. pp. 221–236 (2025)
14. Jin, H., Li, Y., Luan, F., Xiangli, Y., Bi, S., Zhang, K., Xu, Z., Sun, J., Snavely, N.: Neural gaffer: Relighting any object via diffusion. *Advances in Neural Information Processing Systems* **37**, 141129–141152 (2024)
15. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems* **35**, 26565–26577 (2022)
16. Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9359–9369 (2024)

17. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 3045–3059 (2021)
18. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597 (2021)
19. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, Z.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., Wang, Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2025)
20. Lu, Y., Zhang, J., Fang, T., Nahmias, J.D., Tsin, Y., Quan, L., Cao, X., Yao, Y., Li, S.: Matrix3d: Large photogrammetry model all-in-one. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11250–11263 (2025)
21. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. Transactions on Machine Learning Research (2025), <https://openreview.net/forum?id=ntGPYNUF3t>
22. Nalbach, O., Arabadzhyska, E., Mehta, D., Seidel, H.P., Ritschel, T.: Deep shading: convolutional neural networks for screen space shading. In: Computer graphics forum. vol. 36, pp. 65–78. Wiley Online Library (2017)
23. OpenAI: Introducing 4o image generation. <https://openai.com/index/introducing-4o-image-generation/> (2025), accessed: 2026-03-04
24. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023)
25. Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., Huang, X.: Long-context state-space video world models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8733–8744 (2025)
26. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
27. Zeng, C., Dong, Y., Peers, P., Kong, Y., Wu, H., Tong, X.: Dilightnet: Fine-grained lighting control for diffusion-based image generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
28. Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.Q., Hašan, M.: Rgb $\leftrightarrow$ x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In: ACM SIGGRAPH 2024 Conference Papers. SIGGRAPH '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3641519.3657445>, <https://doi.org/10.1145/3641519.3657445>
29. Zhang, L., Rao, A., Agrawala, M.: 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), <https://openreview.net/forum?id=ulcQYxRI1H>
30. Zhang, P., Chen, Y., Huang, H., Lin, W., Liu, Z., Stoica, I., Xing, E.P., Zhang, H.: Faster video diffusion with trainable sparse attention. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), <https://openreview.net/forum?id=VrYCLQ5inI>
31. Zhang, T., Kuang, Z., Jin, H., Xu, Z., Bi, S., Tan, H., Zhang, H., Hu, Y., Hasan, M., Freeman, W.T., et al.: Relitlm: Generative relightable radiance for large reconstruction models. arXiv preprint arXiv:2410.06231 (2024)