

DiffHDR: Re-Exposing LDR Videos with Video Diffusion Models

Zhengming Yu^{1,2†}, Li Ma², Mingming He², Leo Isikdogan³,
Yuancheng Xu^{2,3}, Dmitriy Smirnov³, Pablo Salamanca^{2,3}, Dao Mi³,
Pablo Delgado³, Ning Yu^{2,3}, Julien Philip², Xin Li¹, Wenping Wang¹,
and Paul Debevec^{2,3}

¹ Texas A&M University

² Eyeline Labs

³ Netflix

Abstract. Most digital videos are stored in 8-bit low dynamic range (LDR) formats, where much of the original high dynamic range (HDR) scene radiance is lost due to saturation and quantization. This loss of highlight and shadow detail precludes mapping accurate luminance to HDR displays and limits meaningful re-exposure in post-production workflows. Although techniques have been proposed to convert LDR images to HDR through dynamic range expansion, they struggle to restore realistic detail in the over- and underexposed regions. To address this, we present DiffHDR, a framework that formulates LDR-to-HDR conversion as a generative radiance inpainting task within the latent space of a video diffusion model. By operating in Log-Gamma color space, DiffHDR leverages spatio-temporal generative priors from a pretrained video diffusion model to synthesize plausible HDR radiance in over- and underexposed regions while recovering the continuous scene radiance of the quantized pixels. Our framework further enables controllable LDR-to-HDR video conversion guided by text prompts or reference images. To address the scarcity of paired HDR video data, we develop a pipeline that synthesizes high-quality HDR video training data from static HDRI maps. Extensive experiments demonstrate that DiffHDR significantly outperforms state-of-the-art approaches in radiance fidelity and temporal stability, producing realistic HDR videos with considerable latitude for re-exposure.

Keywords: HDR video generation · Video diffusion · LDR-to-HDR

1 Introduction

High dynamic range (HDR) video captures a wide range of scene luminance, preserving intricate details across both deep shadows and extreme highlights. This capability not only enables more faithful visual reproduction on HDR displays,

[†] Work done during an internship at Eyeline Labs.

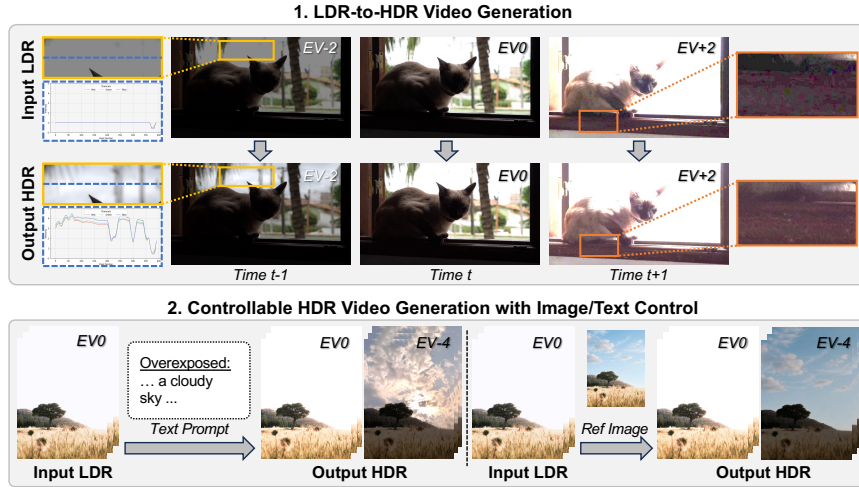


Fig. 1: DiffHDR reconstructs lost radiance to convert LDR videos into faithful HDR while maintaining temporal coherence (Top). DiffHDR further enables controllable HDR synthesis guided by text prompts or reference images, facilitating realistic hallucination of saturated regions (Bottom).

but also provides crucial flexibility in post-production workflows such as color grading, tone mapping, and re-exposure. Despite these benefits, the vast majority of digital video is confined to low dynamic range (LDR) formats, including almost the entirety of video produced using generative models. This LDR-centric ecosystem persists because LDR remains the most portable format for consumer hardware, while true HDR acquisition typically requires high-end cameras or complex multi-exposure techniques that are often impractical for everyday use. Furthermore, recent advanced video generative models [7, 84] are mostly trained on large-scale 8-bit LDR datasets, further entrenching these dynamic range limitations. Therefore, there is a critical need for effective LDR-to-HDR conversion methods which can hallucinate missing scene radiance and unlock the inherent potential of HDR within existing LDR content.

Existing LDR-to-HDR approaches can be broadly categorized into two groups. The first class reconstructs HDR content in a multi-exposure fusion setting [42–45], which requires a sophisticated capture set-up and is not practical for the single LDR video setting. The other generates HDR images from a single LDR input, typically using a feed-forward deep neural network [17, 26, 27, 59, 66, 76, 103, 111]. Due to the limited model capacity and their deterministic pixel-to-pixel translation formulation, these methods often struggle to synthesize photorealistic content in clipped regions. Fundamentally, LDR-to-HDR conversion is a one-to-many problem because of the appearance ambiguity in over- and underexposed regions. This naturally motivates the use of generative models. Training such a generative model remains challenging due to the lack of large-scale, high-quality HDR video datasets. Therefore, a more practical solution is to

leverage the strong priors of video models pretrained on large-scale LDR video. However, this is nontrivial, as models trained on LDR videos do not natively support HDR content due to the fundamental distribution mismatch between LDR and HDR videos.

To address these challenges, we propose DiffHDR, the first video diffusion-based framework for generative reconstruction of HDR videos from a single LDR video. The key enabler of our approach is a deceptively simple observation that HDR videos, when processed with carefully designed tone-mapping curves, can be aligned with the manifold of a video VAE trained on LDR videos. Specifically, we introduce a Log-Gamma color mapping which compresses high dynamic range content into the operational range of the pretrained video VAE, enabling HDR videos to be encoded and decoded without any finetuning. To overcome data scarcity, we develop a curated generation pipeline which leverages high-quality panoramic HDRI maps from Polyhaven [1] to synthesize a diverse HDR video dataset. Despite finetuning solely on synthetic videos derived from static HDRIs, our framework generalizes robustly to real-world videos by leveraging the strong priors of the pretrained video model. To address information loss in clipped regions, we employ luminance-based masks to guide both the generative process and a context-focused cross-attention module. By incorporating context-focused prompting or reference images, this module facilitates controllable reconstruction in over- and underexposed areas, utilizing spatio-temporal cues to hallucinate physically plausible details (Fig. 1). Our main contributions are as follows:

1. We introduce DiffHDR, the first video diffusion framework for LDR-to-HDR reconstruction, along with a curation pipeline which synthesizes high-quality HDR video training data from static HDRIs.
2. We introduce a Log-Gamma color mapping, enabling HDR generation within pretrained latent spaces while preserving the backbone’s generative priors and temporal consistency without any VAE finetuning.
3. We design exposure-aware control mechanisms with luminance-based mask detection, context-focused prompting, and context-focused cross-attention to enhance controllable generation in over- and underexposed regions.
4. DiffHDR achieves state-of-the-art performance on synthetic and in-the-wild benchmarks, significantly outperforming prior methods in radiance fidelity and temporal stability while enabling downstream applications such as text- and image-guided HDR video editing.

2 Related Work

2.1 Multi-Exposure Fusion

HDR videos can be captured directly using specialized hardware such as beam splitters or multi-sensor camera systems [52, 82]. While effective, these solutions are typically expensive and impractical for widespread deployment. As a result, recovering HDR content from standard LDR images has become an attractive alternative.

The classical paradigm reconstructs HDR images from multiple photographs captured with different exposure settings [14], commonly known as multi-exposure fusion. However, capturing multi-exposure image sequences often leads to spatial misalignment due to camera motion or dynamic scene content, making naive fusion prone to artifacts. Early methods addressed this issue by explicitly aligning multi-exposure images using global or local registration techniques [20, 35, 46, 77]. Learning-based approaches have since shown advantages over explicit alignment pipelines. Convolutional neural network (CNN)-based methods generate HDR images directly from misaligned multi-exposure inputs by implicitly handling alignment or fixing misalignment artifacts during reconstruction [43, 51, 71, 93, 94, 98]. Transformer-based architectures further improve performance by modeling long-range dependencies [11, 60, 78, 81, 97, 99, 100, 102].

The multi-exposure paradigm has also been extended to HDR video reconstruction, where different exposure settings are temporally interleaved across frames to provide complementary information [42–45]. In addition to compensating for inter-frame motion, HDR video methods must also enforce temporal consistency to avoid flickering and other temporal artifacts [10, 13, 95]. Despite their success, multi-exposure fusion methods typically require specialized acquisition setups and are inapplicable to single-exposure LDR inputs.

2.2 HDR from a Single Image

While multi-exposure fusion focuses on combining information from multiple LDR images, a complementary line of work aims to generate HDR content from a single LDR input [17]. This can be achieved by explicitly estimating an inverse tone-mapping function [59]. Another class of methods directly regresses HDR outputs from LDR images using neural networks [17, 26, 27, 66, 76, 103, 111]. Some increase dynamic range in intermediate representations, such as gain maps [56, 69] or intrinsic components like shading maps [16]. An alternative strategy predicts multiple virtual LDR images at different exposure levels from a single input, which can then be fused to produce HDR content [18, 53, 54, 69, 110].

Single-image HDR generation relaxes the capture requirements but introduces a fundamental challenge, where the lost information in overexposed or underexposed regions needs to be re-synthesized. Several methods explicitly incorporate inpainting modules to hallucinate missing details in saturated regions [22, 59, 111]. However, when using limited-capacity generative models, the synthesized content often lacks realism or fine details.

2.3 Generative HDR

Advances in generative modeling, including GANs [4, 8, 9, 21, 39, 47–49, 79, 83, 105] and diffusion models [15, 74], have shown strong priors for generation applications [3, 29, 37, 38, 84, 88–90, 96, 101, 104, 106, 112]. Some approaches learn the mapping from LDR images to HDR using only LDR videos, without requiring HDR supervision [5]. Similarly, GlowGAN [85] enables GAN-based HDR image generation by learning from the distribution of LDR content.

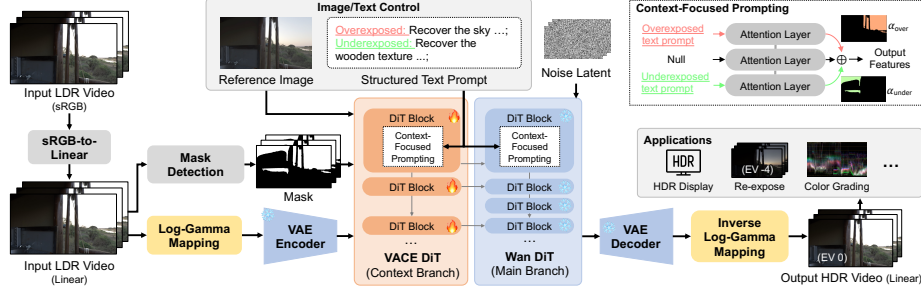


Fig. 2: Framework of DiffHDR. Given an input LDR video, we first detect its clipped regions and map it into the proposed Log-Gamma color space. A finetuned video diffusion model reconstructs missing radiance in over- and underexposed regions. A mask detector and context-focused prompting module support controllable detail synthesis. The final output HDR video supports faithful re-exposure, accurate reproduction on HDR displays, and flexible post-production workflows.

Diffusion models, in particular, have demonstrated strong capability in generating photorealistic image and video, and have been widely applied to tasks such as controllable generation [70, 84, 108], editing [40, 68, 107], inpainting [2, 41, 62], relighting [32, 63, 67, 109, 113] and restoration [55, 75]. These strengths have motivated their adoption in generative HDR creation. Hu *et al.* [36] employ diffusion models to reduce ghosting artifacts in multi-exposure fusion. UltraFusion [12] formulates exposure fusion as a guided inpainting task, using a latent diffusion model to hallucinate missing information in overexposed regions with guidance from underexposed inputs. Bracket Diffusion [6] enables pretrained diffusion models in LDR to generate HDR outputs through multiple diffusion passes under different exposure conditions. HDR-V-Diff [25] introduces a latent diffusion model specifically designed for HDR video generation. Guan *et al.* [23] fine-tune diffusion models to jointly generate gain maps and LDR images for HDR generation, while LEDiff [86] performs HDR generation via latent-space fusion. Concurrently, X2HDR [92] reuses a pretrained variational autoencoder (VAE) for LDR images by compressing HDR content into the PU21 color space, enabling HDR reconstruction within an LDR-oriented latent representation. However, leveraging pretrained video diffusion priors for controllable HDR generation remains largely unexplored.

3 Method

We adopt a latent video diffusion framework to achieve controllable LDR-to-HDR conversion. The overall pipeline is illustrated in Fig. 2. Due to the lack of existing HDR video data, we first construct a curated HDR video dataset using a data generation pipeline based on static HDRIs (Sec. 3.1). To fully leverage the pretrained video VAE, we introduce a Log-Gamma mapping that compresses HDR value into a bounded range (Sec. 3.2). The input LDR video is first mapped

to the Log-Gamma color space, encoded into latent space using the video VAE. A finetuned latent video diffusion model then reconstructs plausible radiance in saturated and noisy regions (Sec. 3.3). We adopt VACE [40], a video-to-video latent diffusion framework, as our backbone. To enhance controllability, we incorporate structured text prompts and reference-image-based control signals that explicitly guide detail synthesis in over- and underexposed regions (Sec. 3.4).

3.1 HDR Video Dataset Curation

Dataset Curation Pipeline. Training video diffusion models for HDR video reconstruction requires paired LDR–HDR video data. While LDR video can be synthesized from HDR video, publicly available HDR video data with high fidelity and adequate dynamic range remains limited. Therefore, we construct a curated HDR video dataset using a rendering-based data generation pipeline built upon 16K resolution HDRIs from Polyhaven [1].

For each HDRI, we place the camera at the origin and set the HDRI as the skybox. We render short video sequences in Blender using multiple predefined camera configurations. Specifically, we design three motion patterns for diverse dynamic range distributions: (1) highlight-focused zoom-in/out sequences emphasizing saturated regions, (2) shadow-focused zoom-in/out sequences emphasizing underexposed areas, and (3) camera rotation sequences introducing pseudo dynamics. For highlight-focused and shadow-focused sequences, we first identify the brightest and darkest pixels in the HDRI, respectively, and orient the camera toward the corresponding areas. For zoom-in/out, the start and end focal lengths are randomly sampled between (18, 30) mm and (50, 70) mm. For rotation sequences, we rotate the camera around the vertical axis by 120° per segment, obtaining 3 segments to cover the full 360° from each HDRI.

We render videos in linear color space (i.e. Rec.709) from about 800 HDRIs. The resulting dataset includes approximately 5400 HDR video sequences, each containing 81 frames, across diverse illumination environments. These sequences provide the temporally consistent HDR supervision that is essential for learning radiance reconstruction and re-exposure within the video diffusion framework.

Data Augmentation Strategies. Given a rendered HDR video, we synthesize its LDR video by simulating the LDR video formation process, including exposure shift, heteroscedastic camera noise, quantization, and clipping.

Exposure shift. We randomly sample an exposure offset $\Delta \in [-2, 2]$ stops and scale the video in linear space by a factor of 2^Δ .

Camera noise. To simulate realistic sensor noise, we follow CBDNet [28] and model camera noise as a heteroscedastic Gaussian process whose variance depends on the signal intensity. Specifically, the noise is formulated as:

$$\mathbf{n}_t(L_t) = \sqrt{L_t\sigma_s^2 + \sigma_c^2} \boldsymbol{\epsilon}_t, \quad (1)$$

where L_t denotes the input pixel intensity in linear space. σ_s and σ_c are the signal-dependent component and stationary noise, sampled from $(0, 8.5 \times 10^{-4})$ and $(0, 1.5 \times 10^{-5})$, respectively. $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ is a standard Gaussian noise field.

Unlike CBDNet, which samples noise independently for each image, we extend the model to videos by introducing temporal correlation in the underlying Gaussian noise. Specifically, we share (σ_s, σ_c) across all frames and model ϵ_t using an AR(1) (first-order autoregressive) process [30]:

$$\epsilon_t = \rho \epsilon_{t-1} + \sqrt{1 - \rho^2} \mathbf{u}_t, \quad (2)$$

where $\mathbf{u}_t \sim \mathcal{N}(0, \mathbf{I})$ and ρ controls the temporal correlation strength. When $\rho = 0$, the noise reduces to independent sampling per frame. In our experiments, we set $\rho = 0.5$.

Quantization and clipping. To produce the final LDR inputs, we convert the HDR video to sRGB, clip values to $[0, 1]$, and quantize to 8-bit precision.

3.2 Log-Gamma Color Mapping

The video VAE is a core component in latent video diffusion models [84]. However, as shown in Fig. 5, a VAE pretrained on LDR data fails to accurately encode and decode HDR content, as the pixel values can far exceed the standard $[0, 1]$ range of LDR signals. While it is possible to finetune the VAE to support HDR content [25, 86], this approach is hindered by the lack of large-scale, high-quality HDR video datasets. Furthermore, modifying the VAE architecture or weights shifts the learned latent space, potentially disrupting the generative priors of the pretrained model.

Instead of adapting the VAE itself, we introduce a transformation that maps HDR radiance into a representation compatible with the VAE’s pretrained domain. Specifically, we formulate this as a color-mapping function. Inspired by μ -law tone mapping [24, 97] and perceptual gamma compression in imaging pipelines, we propose a Log-Gamma color mapping defined as:

$$\mathcal{T}(x) = \left(\frac{\log(1 + \gamma x)}{\log(1 + \gamma M)} \right)^{\frac{1}{\gamma}}, \quad (3)$$

where x denotes the linear HDR radiance, M is the maximum representable radiance, and γ regulates compression. The logarithmic component compresses high dynamic range radiance, while aligning the radiance distribution with natural LDR statistics, ensuring direct compatibility with the pretrained VAE.

3.3 Diffusion-Based LDR-to-HDR Conversion

Preliminary. Our model builds upon VACE [40], a diffusion-based video-to-video framework for video editing. VACE introduces a Video Condition Unit (VCU) that integrates text prompts, context frames, and masks into a unified conditioning interface. These inputs are encoded into latent tokens via a video VAE and processed by a DiT-based backbone to model spatiotemporal dependencies under the flow matching framework [57, 58].

Model Architecture of DiffHDR. As shown in Fig. 2, the input LDR video is first linearized and mapped to the Log-Gamma color space, and then encoded into a latent representation using the VAE encoder. This LDR latent is fed into the context branch to condition the denoising process of the main branch. We additionally compute an exposure mask indicating the over- and underexposed regions, guiding the model toward areas that require detail hallucination. Starting from a random noise latent, the main branch iteratively denoises to produce the final HDR latent, which is subsequently decoded and inverse Log-Gamma mapped to linear space, yielding the final HDR video suitable for downstream applications. To enable controllable hallucination in clipped regions, we condition the model on both text prompts and reference images, as detailed in Sec. 3.4.

Fine-tuning Strategy. To preserve the pretrained generative prior of VACE, we freeze the backbone parameters and fine-tune only the DiT blocks via LoRA adapters. Specifically, rank-32 LoRA layers are inserted into the attention and feed-forward layers of the DiT blocks. This parameter-efficient adaptation ensures stable training while reducing overfitting to the HDR dataset.

Training Objective. The training objective follows a standard rectified flow-matching formulation [58]. Specifically, given a HDR video sample in latent representation $\mathbf{x}_1$ and a Gaussian noise $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$, we sample a timestep $t \in [0, 1]$ and construct an intermediate latent via linear interpolation:

$$\mathbf{x}_t = t\mathbf{x}_1 + (1 - t)\mathbf{x}_0. \quad (4)$$

The final objective is defined as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t} \|u_\Theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2, \quad (5)$$

where u_Θ indicates the video DiT and $\mathbf{c}$ denotes the conditioning signals including the LDR input, exposure masks, and optional text or image prompts.

3.4 Controllable HDR Video Reconstruction

Luminance-Based Mask Detection. We construct a luminance-based mask detection to identify over- and underexposed regions in LDR videos. The input sRGB frames are first linearized using the inverse sRGB transfer function, and luminance is computed following the Rec.709 standard. Over- and underexposed regions are detected by thresholding luminance values: pixels with luminance greater than τ_{high} are considered overexposed, while those below τ_{low} are treated as underexposed. We set $\tau_{high} = 0.95$ and $\tau_{low} = 0.05$.

To further improve temporal stability, we perform per-pixel exponential moving average (EMA) smoothing:

$$\tilde{M}_t = \alpha M_t + (1 - \alpha)\tilde{M}_{t-1}, \quad (6)$$

where α controls the smoothing strength, M_t denotes the mask detected at time t , and $\tilde{M}_t$ is the temporally smoothed mask. We set $\alpha = 0.7$. This temporal aggregation suppresses frame-wise fluctuations and improves mask consistency for video diffusion conditioning.

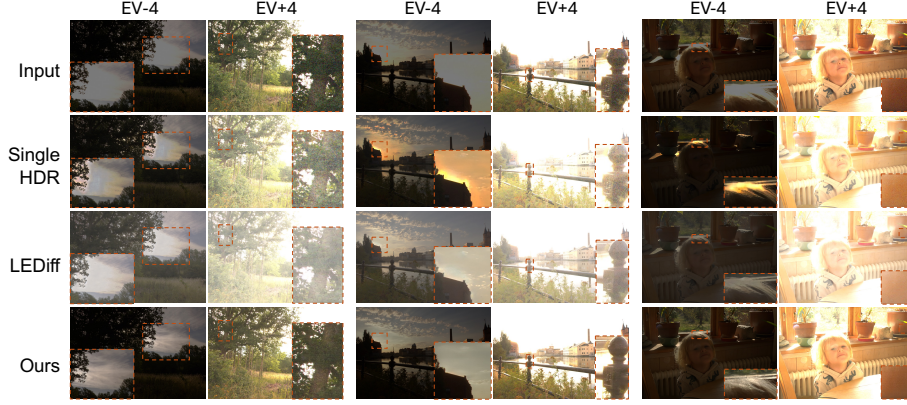


Fig. 3: Qualitative comparison on the SI-HDR dataset. Results are shown under multiple re-exposure levels to assess highlight restoration, shadow recovery, and radiance consistency across methods. Zoom in for detailed comparison.

Context-Focused Prompting. Our key idea is to design a context-focused captioning format that explicitly grounds the visual semantics of regions with distinct exposure characteristics. Unlike standard prompts that provide a single holistic description of the scene, our prompts follow a structured format: `[overexposed: <description>]; [underexposed: <description>]`. This formulation disentangles the semantic guidance for saturated highlights and shadowed regions, enabling region-aware conditioning.

Inspired by classifier-free guidance [34], which manipulates the global denoising trajectory using conditional and unconditional prompts, we introduce a context-focused cross-attention mechanism that operates locally inside the DiT cross-attention blocks. Importantly, this modification is applied exclusively at inference, preserving pretrained weights and training objectives. Specifically, our CFA module is applied at every cross-attention layer for both the VACE context branch and main branch. Let $\mathbf{x}$ denote the current token features and $\mathbf{c}$, $\mathbf{c}_{\text{over}}$, and $\mathbf{c}_{\text{under}}$ indicate the unconditional embedding, the overexposed text prompt, and the underexposed text prompt, respectively. The output of the cross-attention layer is

$$\mathbf{r}_{\text{base}} = \text{CA}(\mathbf{x}, \mathbf{c}), \quad \mathbf{r}_{\text{over}} = \text{CA}(\mathbf{x}, \mathbf{c}_{\text{over}}), \text{ and } \mathbf{r}_{\text{under}} = \text{CA}(\mathbf{x}, \mathbf{c}_{\text{under}}), \quad (7)$$

where $\text{CA}(\cdot)$ denotes the cross-attention operator. Given the corresponding spatial masks $\mathbf{M}_{\text{over}}$ and $\mathbf{M}_{\text{under}}$, we then refine the model output using a mask-guided routing mechanism:

$$\mathbf{r} = \mathbf{r}_{\text{base}} + \alpha_{\text{over}} \mathbf{M}_{\text{over}} \odot (\mathbf{r}_{\text{over}} - \mathbf{r}_{\text{base}}) + \alpha_{\text{under}} \mathbf{M}_{\text{under}} \odot (\mathbf{r}_{\text{under}} - \mathbf{r}_{\text{base}}), \quad (8)$$

where α_{over} and α_{under} control the strength of region-specific modulation, and $\odot$ denotes element-wise multiplication.

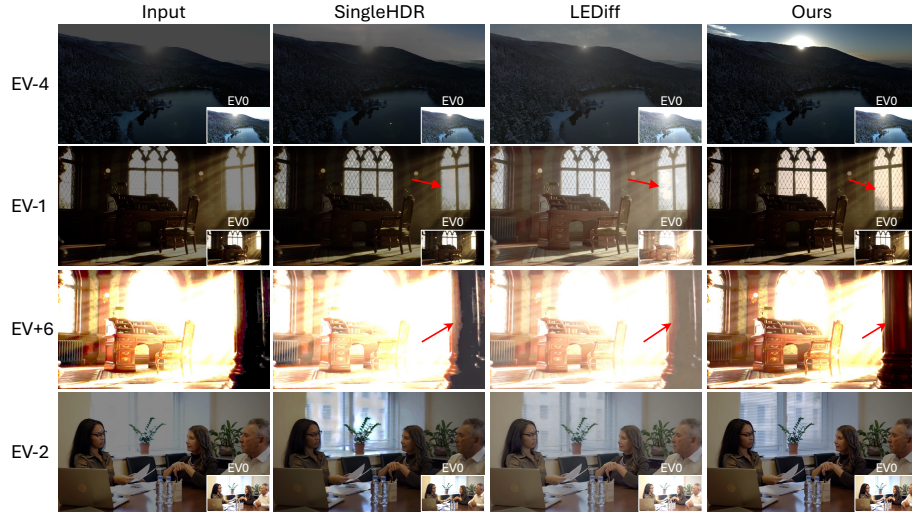


Fig. 4: Qualitative comparison on in-the-wild video dataset. Results are shown under multiple re-exposure levels across frames to assess highlight restoration, shadow recovery, and radiance consistency across methods. Zoom in for detailed comparison.

This design preserves the global semantic structure from the base prompt while selectively steering the generation in over- and underexposed regions. Because the modification only alters cross-attention residuals at inference, it is fully compatible with trained DiT models and does not require retraining.

Reference Image-Based Conditioning. Text prompts provide simple and abstract control for guiding the synthesis of new details. However, in some cases more fine-grained control is required. To this end, we allow the user to optionally provide a reference image that specifies detailed appearance cues in LDR. Such reference images can be generated using existing image editing models. To condition the generation process, we follow the VACE architecture and encode the reference image using the VAE. The encoded reference is then concatenated along the temporal dimension to inject the reference signal into the model.

4 Experiments

We evaluate our method across diverse datasets, including the SI-HDR dataset [31], the Cinematic Video dataset [19], and 50 held-out videos from our Polyhaven synthetic dataset (excluded from training). In addition, we collect 50 in-the-wild videos from Pexels [72], and 10 videos generated by Veo2 [80].

Table 1: Quantitative comparison on the SI-HDR dataset.

Method	HDR-VDP3 $\uparrow$	PU21-PIQE $\downarrow$	FID $\downarrow$
HDRCNN	6.82	24.30	19.26
MaskHDR	6.87	24.00	19.78
SingleHDR	7.37	26.64	27.55
LEDiff	6.56	22.71	25.98
Ours	6.98	19.37	18.68

Table 2: Quantitative comparison on Cinematic Video and synthetic datasets.

Method	Cinematic Video Dataset				Polyhaven Synthetic Video Dataset			
	FOVVPD $\uparrow$	DOVER $\uparrow$	MUSIQ $\uparrow$	CLIPQA $\uparrow$	FOVVPD $\uparrow$	DOVER $\uparrow$	MUSIQ $\uparrow$	CLIPQA $\uparrow$
SingleHDR	6.56	0.77	51.79	0.33	7.48	0.66	57.60	0.46
LEDiff	3.75	0.63	47.90	0.28	4.68	0.59	58.92	0.46
Ours	6.89	0.81	58.38	0.41	7.65	0.68	60.02	0.50

Table 3: Quantitative comparison on in-the-wild and Veo2 video datasets.

Method	In-the-wild Video Dataset			Veo2 Video Dataset		
	DOVER $\uparrow$	MUSIQ $\uparrow$	CLIPQA $\uparrow$	DOVER $\uparrow$	MUSIQ $\uparrow$	CLIPQA $\uparrow$
SingleHDR	0.71	53.21	0.46	0.59	43.78	0.30
LEDiff	0.61	53.68	0.42	0.53	41.41	0.29
Ours	0.74	55.79	0.48	0.61	46.06	0.34

For the SI-HDR dataset, following LEDiff [86], we adopt HDR-VDP3 [65], PU21-PIQE [31], and FID [33] as evaluation metrics. For video benchmarks, we use FovVideoVDP [64] as a reference-based HDR video metric, and adopt DOVER [91], CLIPQA [87], and MUSIQ [50] as non-reference perceptual quality metrics, following FlashVSR [114]. These metrics effectively evaluate the spatial and temporal quality of the reconstructed HDR.

We compare DiffHDR against state-of-the-art LDR-to-HDR methods. In addition, we conduct comprehensive ablation studies to validate the effectiveness of each proposed component and demonstrate further applications of our framework. Additional results are provided in the supplementary material.

4.1 Implementation Details

We build our framework upon the pretrained video diffusion model Wan-2.1-VACE-14B [40] and adopt the corresponding Wan-2.1-VAE [84] with a spatiotemporal compression ratio of $4 \times 8 \times 8$. Our model is trained at a spatiotemporal resolution of $33 \times 1280 \times 720$. For LoRA adaptation, we insert rank-32 LoRA modules into the DiT blocks while freezing the backbone parameters. The model is trained using the AdamW [61] optimizer with a constant learning rate of 1×10^{-4} for 10,000 steps. Training is performed on 8 NVIDIA A100 GPUs with mixed precision setting. Since BF16 precision can introduce banding artifacts in HDR decoding due to its limited precision, we use BF16 for finetuning the DiT and FP32 for the VAE to preserve tonal continuity.

4.2 Comparisons with State-of-The-Art

Quantitative Evaluations. We quantitatively compare DiffHDR with state-of-the-art LDR-to-HDR methods on both image and video benchmarks. For all LDR-based perceptual metrics (i.e., FID, PU21-PIQE, MUSIQ, CLIPQA,

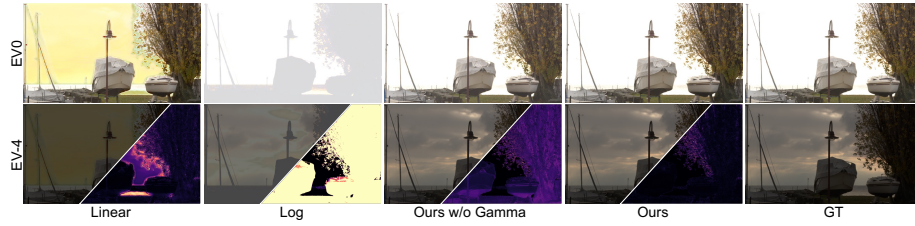


Fig. 5: Comparison of different color mappings. We compare different color mapping methods by feeding the mapped images into the VAE and evaluating the reconstruction quality. We also visualize per-pixel error maps, where brighter regions indicate larger reconstruction errors. Our method achieves the best performance.

and DOVER), we uniformly apply Reinhard tone mapping [73] to convert HDR outputs to LDR space to ensure fair comparison across methods. Although not trained on image data, our method achieves the best performance in PU21-PIQE and FID, and ranks second on HDR-VDP3 as shown in Tab. 1. The slightly lower HDR-VDP3 score is likely because our method generatively inpaints plausible details in clipped regions that may not be pixel-wise identical to the ground truth and are therefore penalized by this metric. Nevertheless, these results indicate the superior perceptual quality of our method. On video datasets, DiffHDR consistently achieves the best performance on both the Cinematic Video dataset and the Polyhaven synthetic video dataset across reference-based and non-reference metrics as shown in Tab. 2. Furthermore, on the in-the-wild and Veo2-generated video datasets, our method exhibits the strongest generalization capability, outperforming prior approaches across all reported metrics as shown in Tab 3. These results indicate that DiffHDR produces temporally coherent and visually stable HDR videos, enabling robust HDR reconstruction and re-exposure in dynamic real-world scenes.

Qualitative Evaluations. We present qualitative comparisons on the SI-HDR dataset (Fig. 3) and in-the-wild videos (Fig. 4). To avoid potential bias introduced by different tone-mapping operators, we directly compare HDR outputs under multiple exposure levels, allowing consistent evaluation of recovered radiance and dynamic range. As shown in Fig. 3, DiffHDR effectively restores fine details in severely saturated sky regions and generalizes well to challenging high-intensity structures such as overexposed hair strands. In shadow regions, our method suppresses noise while recovering structural details. In contrast, LEDiff [86] and SingleHDR [59] introduce visible artifacts in saturated areas and struggle to remove camera noise in dark regions. For in-the-wild videos (Fig. 4), DiffHDR successfully reconstructs the radiance of the sun with wider dynamic range in the first example, while preserving surrounding high-frequency details. It also restores overexposed window regions and underexposed pillars with improved dynamic range and structural fidelity. LEDiff can approximate the sun’s shape but produces limited dynamic range and flattened highlights. SingleHDR fails to recover accurate structures in saturated regions, with noticeable artifacts.



Fig. 6: Ablation study on data augmentation strategy and mask detection. We compare our methods without the data augmentation training and the mask detection module.

Moreover, both LEDiff and SingleHDR suffer from temporal inconsistencies in high-intensity areas, whereas DiffHDR maintains temporally stable reconstruction.

Ablation Studies. We conduct ablation studies on the Polyhaven synthetic dataset to validate the effectiveness of the proposed components.

Effect of Log-Gamma mapping. To evaluate the proposed Log-Gamma color mapping, we compare four encoding strategies applied before the VAE encoder and decoder: (1) directly encoding linear HDR values (Linear), (2) a pure logarithmic mapping (Log) used in LEDiff’s finetuning, (3) our mapping without gamma compression, $\mathcal{T}'(x) = \frac{\log(1+x)}{\log(1+M)}$, and (4) our full Log-Gamma mapping.

We assess reconstruction quality both quantitatively and qualitatively. As shown in Fig. 5, both Linear and Log mappings fail to faithfully recover color and structural details of the ground-truth HDR inputs. The mapping without gamma compression introduces noticeable artifacts, particularly around high-contrast edges. In contrast, our full Log-Gamma mapping accurately reconstructs fine details and preserves color consistency.

These observations are further supported by the quantitative results in Tab. 4. For metric computation, reconstructed HDR outputs are converted to sRGB space and compared against the ground-truth sRGB images using PSNR, SSIM, and LPIPS. Our Log-Gamma mapping achieves the best performance, demonstrating its compatibility with the pretrained VAE.

Effect of data augmentation and mask guidance. As shown in Fig. 6, removing our exposure-aware data augmentation during training leads to insufficient noise suppression, resulting in visibly noisy outputs. Without mask guidance, the model struggles to correctly inpaint shadow textures. In contrast, the full

Table 4: Quantitative comparison of different mapping strategies.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Linear	22.16	0.74	0.28
Log	14.61	0.74	0.57
Ours w/o gamma	25.38	0.75	0.34
Ours	32.86	0.86	0.15

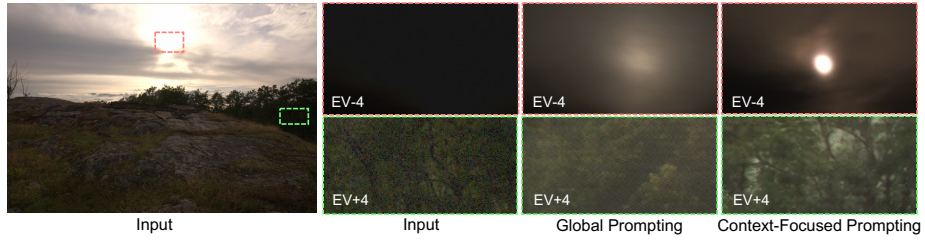


Fig. 7: Ablation study on context-focused prompting. We compare our context-focused prompting with global prompting.

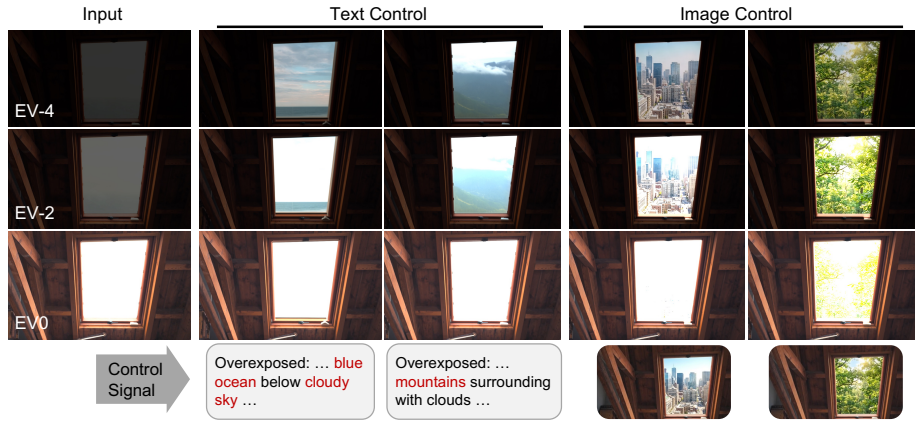


Fig. 8: Text- and image-guided generation. DiffHDR supports both text and image controls for guiding the generation in reconstructed regions.

model effectively suppresses camera noise and restores detailed textures. These improvements are quantitatively validated in Tab. 5.

Effect of context-focused prompting. We further evaluate the proposed context-focused prompting modules in Fig. 7. When using global prompts alone, the model fails to reconstruct accurate high-intensity structures, such as the sun’s shape. With context-focused prompting, the model successfully restores the correct solar structure and shadowed tree regions.

Controllable Generation. In many real-world LDR videos, severely saturated regions may correspond to multiple plausible underlying radiance configurations, leading to inherent ambiguity in HDR reconstruction. To leverage the generative capability of the video diffusion model, our framework enables controllable HDR reconstruction guided by text prompts or reference images.

Table 5: Ablation study on data augmentation and mask guidance.

Method	FOVDP ↑	DOVER ↑	MUSIQ ↑	CLIPQA ↑
Ours w/o data aug	7.57	0.67	59.86	0.49
Ours w/o mask	7.58	0.67	59.42	0.48
Ours	7.65	0.68	60.02	0.50

As shown in Fig. 8, by providing different textual descriptions or image references, DiffHDR generates diverse and semantically consistent HDR outputs within overexposed regions. The reconstructed radiance not only aligns with the conditioning inputs but also remains temporally coherent across frames. These results demonstrate that our method goes beyond deterministic restoration and supports controllable, content-aware HDR video synthesis.

5 Conclusion

We presented DiffHDR, a novel video diffusion-based framework for generative LDR-to-HDR reconstruction. By reformulating conversion as a radiance inpainting problem within the latent space of a pretrained video diffusion model, our approach leverages strong spatiotemporal priors to recover plausible HDR radiance in overexposed and underexposed regions while maintaining temporal coherence. The proposed Log-Gamma color mapping enables HDR modeling without modifying the pretrained VAE, effectively bridging the distribution gap between LDR and HDR videos. Combined with our HDR video curation pipeline and exposure-aware control mechanisms, DiffHDR achieves state-of-the-art performance across synthetic and real-world benchmarks. Our framework further supports controllable HDR reconstruction guided by text or reference images, opening new possibilities for creative post-production. This work establishes a promising direction for integrating generative video models into practical HDR reconstruction and re-exposure workflows.

Limitations. Our current model is trained and evaluated at 720p resolution, which limits its direct applicability to higher-resolution settings such as 4K. Direct 4K inference remains challenging due to increased GPU memory consumption, computational cost, and the resolution gap between training and testing. Future work could improve this through more efficient training and inference strategies, such as memory-efficient optimization and lightweight architectures. These directions may reduce the resource requirements for high-resolution training and inference while improving robustness across output resolutions.

References

1. Poly haven (2024), <https://polyhaven.com/>, accessed: 2024-10-21
2. Adiya, T., Ha, S.J.: Omnipainter: Global-local temporally consistent video inpainting diffusion model. In: International Conference on Learning Representations (ICLR) (2024)
3. 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)
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. Pmlr (2017)
5. Banterle, F., Marnerides, D., Bashford-rogers, T., Debattista, K.: Self-supervised high dynamic range imaging: What can be learned from a single 8-bit video? ACM Trans. Graph. **43**(2) (Mar 2024). <https://doi.org/10.1145/3648570>, <https://doi.org/10.1145/3648570>
6. Bemana, M., Leimkühler, T., Myszkowski, K., Seidel, H.P., Ritschel, T.: Bracket diffusion: Hdr image generation by consistent ldr denoising. arXiv preprint arXiv:2405.14304 (2024)
7. 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)
8. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16123–16133 (2022)
9. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5799–5809 (2021)
10. Chen, G., Chen, C., Guo, S., Liang, Z., Wong, K.Y.K., Zhang, L.: Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2502–2511 (October 2021)
11. Chen, R., Zheng, B., Zhang, H., Chen, Q., Yan, C., Slabaugh, G., Yuan, S.: Improving dynamic hdr imaging with fusion transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 340–349 (2023)
12. Chen, Z., Wang, Y., Cai, X., You, Z., Lu, Z., Zhang, F., Guo, S., Xue, T.: Ultra-fusion: Ultra high dynamic imaging using exposure fusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16111–16121 (2025)
13. Chung, H., Cho, N.I.: Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12760–12769 (October 2023)
14. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. p. 369–378. SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., USA (1997). <https://doi.org/10.1145/258734.258884>, <https://doi.org/10.1145/258734.258884>
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

16. Dille, S., Careaga, C., Aksoy, Y.: Intrinsic single-image hdr reconstruction. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XI. p. 161–177. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73247-8_10, https://doi.org/10.1007/978-3-031-73247-8_10
17. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)* **36**(6), 1–15 (2017)
18. Endo, Y., Kanamori, Y., Mitani, J.: Deep reverse tone mapping. *ACM Trans. Graph.* **36**(6), 177–1 (2017)
19. Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., Brendel, H.: Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In: Digital photography X. vol. 9023, pp. 279–288. SPIE (2014)
20. Gallo, O., Troccoli, A., Hu, J., Pulli, K., Kautz, J.: Locally non-rigid registration for mobile hdr photography. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 48–55 (2015). <https://doi.org/10.1109/CVPRW.2015.7301366>
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144 (2020)
22. Goswami, A., Singh, A.R., Banterle, F., Debattista, K., Bashford-Rogers, T.: Semantic aware diffusion inverse tone mapping. *arXiv preprint arXiv:2405.15468* (2024)
23. Guan, Y., Xu, R., Liao, Y., Yao, M., Wang, L., Xiong, Z.: Hdr image generation via gain map decomposed diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17536–17545 (2025)
24. Guan, Y., Xu, R., Yao, M., Gao, R., Wang, L., Xiong, Z.: Diffusion-promoted hdr video reconstruction. In: European Conference on Computer Vision. pp. 20–38. Springer (2024)
25. Guan, Y., Xu, R., Yao, M., Gao, R., Wang, L., Xiong, Z.: Diffusion-promoted hdr video reconstruction. In: Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part IX. p. 20–38. Springer-Verlag, Berlin, Heidelberg (2025). https://doi.org/10.1007/978-3-031-91838-4_2, https://doi.org/10.1007/978-3-031-91838-4_2
26. Guo, C., Fan, L., Xue, Z., Jiang, X.: Learning a practical sdr-to-hdrtv up-conversion using new dataset and degradation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22231–22241 (June 2023)
27. Guo, C., Jiang, X.: Lhdr: Hdr reconstruction for legacy content using a lightweight dnn. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 3155–3171 (December 2022)
28. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1712–1722 (2019)
29. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103* (2024)
30. Hamilton, J.D.: Time series analysis. Princeton university press (2020)

31. Hanji, P., Mantiuk, R., Eilertsen, G., Hajisharif, S., Unger, J.: Comparison of single image hdr reconstruction methods—the caveats of quality assessment. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–8 (2022)
32. He, M., Clausen, P., Taşel, A.L., Ma, L., Pilarski, O., Xian, W., Rikker, L., Yu, X., Burgert, R., Yu, N., et al.: Diffrelight: Diffusion-based facial performance relighting. In: SIGGRAPH Asia 2024 conference papers. pp. 1–12 (2024)
33. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
34. Ho, J., Salimans, T.: Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022)
35. Hu, J., Gallo, O., Pulli, K., Sun, X.: Hdr deghosting: How to deal with saturation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2013)
36. Hu, T., Yan, Q., Qi, Y., Zhang, Y.: Generating content for hdr deghosting from frequency view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25732–25741 (June 2024)
37. Huang, Z., Chen, J., Lin, C., Hu, C., Huang, H., Yu, Z., Li, M., Liu, Y., Gu, Z., Zhao, Z., et al.: Unirecgen: unifying multi-view 3d reconstruction and generation. *arXiv preprint arXiv:2604.01479* (2026)
38. Huang, Z., Yu, N., Chen, G., Qiu, H., Debevec, P., Liu, Z.: Vchain: Chain-of-visual-thought for reasoning in video generation. *arXiv preprint arXiv:2510.05094* (2025)
39. Jiang, K., Chen, S.Y., Fu, H., Gao, L.: Nerffacelighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM Transactions on Graphics* **42**(3), 1–18 (2023)
40. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598* (2025)
41. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: European Conference on Computer Vision. pp. 150–168. Springer (2024)
42. Kalantari, N.K., Ramamoorthi, R.: Deep hdr video from sequences with alternating exposures. *Computer Graphics Forum* **38**(2), 193–205 (2019). <https://doi.org/https://doi.org/10.1111/cgf.13630>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13630>
43. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* **36**(4), 144–1 (2017)
44. Kalantari, N.K., Shechtman, E., Barnes, C., Darabi, S., Goldman, D.B., Sen, P.: Patch-based high dynamic range video. *ACM Trans. Graph.* **32**(6) (Nov 2013). <https://doi.org/10.1145/2508363.2508402>, <https://doi.org/10.1145/2508363.2508402>
45. Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. *ACM Trans. Graph.* **22**(3), 319–325 (Jul 2003). <https://doi.org/10.1145/882262.882270>, <https://doi.org/10.1145/882262.882270>
46. Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. *ACM Trans. Graph.* **22**(3), 319–325 (Jul 2003). <https://doi.org/10.1145/882262.882270>, <https://doi.org/10.1145/882262.882270>
47. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. *Advances in neural information processing systems* **34**, 852–863 (2021)

48. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
49. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
50. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
51. Kong, L., Li, B., Xiong, Y., Zhang, H., Gu, H., Chen, J.: Safnet: Selective alignment fusion network for efficient hdr imaging. arXiv preprint arXiv:2407.16308 (2024)
52. Kronander, J., Gustavson, S., Bonnet, G., Ynnerman, A., Unger, J.: A unified framework for multi-sensor hdr video reconstruction. *Image Commun.* **29**(2), 203–215 (Feb 2014). <https://doi.org/10.1016/j.image.2013.08.018>, <https://doi.org/10.1016/j.image.2013.08.018>
53. Le, P.H., Le, Q., Nguyen, R., Hua, B.S.: Single-image hdr reconstruction by multi-exposure generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4063–4072 (January 2023)
54. Lee, S., An, G.H., Kang, S.J.: Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In: proceedings of the European Conference on Computer Vision (ECCV). pp. 596–611 (2018)
55. Li, H., Yang, Y., Chang, M., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **479**, 47–59 (2022)
56. Liao, Y., Guan, Y., Xu, R., Li, J., Sun, S., Xiong, Z.: Learning gain map for inverse tone mapping. In: The Thirteenth International Conference on Learning Representations (2025)
57. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
58. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
59. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image hdr reconstruction by learning to reverse the camera pipeline. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1651–1660 (2020)
60. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: European Conference on Computer Vision. pp. 344–360. Springer (2022)
61. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
62. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
63. Ma, L., He, M., Yu, X., George, D.M., Taşel, A.L., Debevec, P., Philip, J.: Bodyrelux: Temporally consistent full-body video relighting. arXiv preprint arXiv:2605.21766 (2026)
64. Mantiuk, R.K., Denes, G., Chapiro, A., Kaplanyan, A., Rufo, G., Bachy, R., Lian, T., Patney, A.: Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)* **40**(4), 1–19 (2021)

65. Mantiuk, R.K., Hammou, D., Hanji, P.: Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625* (2023)
66. Marnierides, D., Bashford-Rogers, T., Hatchett, J., Debattista, K.: Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In: *Computer Graphics Forum*. vol. 37, pp. 37–49. Wiley Online Library (2018)
67. Mei, Y., He, M., Ma, L., Philip, J., Xian, W., George, D.M., Yu, X., Dedic, G., Taşel, A.L., Yu, N., et al.: Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5510–5522 (2025)
68. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: *International Conference on Learning Representations (ICLR)* (2022)
69. Meng, Y., Jin, X., Lei, L., Guo, C.L., Li, C.: UltraLED: Learning to see everything in ultra-high dynamic range scenes. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025), <https://openreview.net/forum?id=zZLfHw4Erp>
70. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 4296–4304 (2024)
71. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing* **30**, 3885–3896 (2021)
72. Pexels: Pexels: Free stock photos & videos. <https://www.pexels.com> (2026), accessed: 2026-03-03
73. Reinhard, E., Devlin, K.: Dynamic range reduction inspired by photoreceptor physiology. *IEEE transactions on visualization and computer graphics* **11**(1), 13–24 (2005)
74. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
75. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(4), 4713–4726 (2023)
76. Santos, M.S., Ren, T.I., Kalantari, N.K.: Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Transactions on Graphics (TOG)* **39**(4), 80–1 (2020)
77. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.* **31**(6), 203–1 (2012)
78. Song, J.W., Park, Y.I., Kong, K., Kwak, J., Kang, S.J.: Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In: *European Conference on Computer Vision*. pp. 288–304. Springer (2022)
79. Sun, J., Wang, X., Wang, L., Li, X., Zhang, Y., Zhang, H., Liu, Y.: Next3d: Generative neural texture rasterization for 3d-aware head avatars. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 20991–21002 (2023)

80. 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)
81. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12790–12799. IEEE (2023)
82. Tocci, M.D., Kiser, C., Tocci, N., Sen, P.: A versatile hdr video production system. *ACM Trans. Graph.* **30**(4) (Jul 2011). <https://doi.org/10.1145/2010324.1964936>, <https://doi.org/10.1145/2010324.1964936>
83. Trevithick, A., Chan, M., Stengel, M., Chan, E., Liu, C., Yu, Z., Khamis, S., Ramamoorthi, R., Nagano, K.: Real-time radiance fields for single-image portrait view synthesis
84. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
85. Wang, C., Serrano, A., Pan, X., Chen, B., Myszkowski, K., Seidel, H.P., Theobalt, C., Leimkühler, T.: Glowgan: Unsupervised learning of hdr images from ldr images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10509–10519 (2023)
86. Wang, C., Xia, Z., Leimkühler, T., Myszkowski, K., Zhang, X.: Lediff: Latent exposure diffusion for hdr generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 453–464 (2025)
87. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
88. Wang, J., Chen, Z., Ling, J., Xie, R., Song, L.: 360-degree panorama generation from few unregistered nfov images. arXiv preprint arXiv:2308.14686 (2023)
89. Wang, J., Lin, C., Liu, Y., Xu, R., Dou, Z., Long, X., Guo, H., Komura, T., Wang, W., Li, X.: Pdt: Point distribution transformation with diffusion models. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
90. Wang, J., Liu, Y., Dou, Z., Yu, Z., Liang, Y., Lin, C., Xie, R., Song, L., Li, X., Wang, W.: Disentangled clothed avatar generation from text descriptions. In: European Conference on Computer Vision. pp. 381–401. Springer (2024)
91. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20144–20154 (2023)
92. Wu, R., Su, W., Ma, K., Liao, J., Mantiuk, R.K.: X2hdr: Hdr image generation in a perceptually uniform space. arXiv preprint arXiv:2602.04814 (2026)
93. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 117–132 (2018)
94. Xiong, P., Chen, Y.: Hierarchical fusion for practical ghost-free high dynamic range imaging. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4025–4033 (2021)
95. Xu, G., Wang, Y., Gu, J., Xue, T., Yang, X.: Hdrflow: Real-time hdr video reconstruction with large motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24851–24860 (2024)

96. Xu, Y., Xian, W., Ma, L., Philip, J., Tasel, A.L., Zhao, Y., Burgert, R., He, M., Hermann, O., Pilarski, O., et al.: Virtually being: Customizing camera-controllable video diffusion models with multi-view performance captures. arXiv preprint arXiv:2510.14179 (2025)
97. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2019)
98. Yan, Q., Gong, D., Zhang, P., Shi, Q., Sun, J., Reid, I., Zhang, Y.: Multi-scale dense networks for deep high dynamic range imaging. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 41–50 (2019). <https://doi.org/10.1109/WACV.2019.00012>
99. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing* **29**, 4308–4322 (2020)
100. Yan, Q., Zhang, S., Chen, W., Tang, H., Zhu, Y., Sun, J., Van Gool, L., Zhang, Y.: Smae: Few-shot learning for hdr deghosting with saturation-aware masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5775–5784 (2023)
101. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
102. Ye, Q., Xiao, J., Lam, K.m., Okatani, T.: Progressive and selective fusion network for high dynamic range imaging. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 5290–5297 (2021)
103. Yu, H., Liu, W., Long, C., Dong, B., Zou, Q., Xiao, C.: Luminance attentive networks for hdr image and panorama reconstruction. In: *Computer Graphics Forum*. vol. 40, pp. 181–192. Wiley Online Library (2021)
104. Yu, Z., Dou, Z., Long, X., Lin, C., Li, Z., Liu, Y., Müller, N., Komura, T., Habermann, M., Theobalt, C., et al.: Surf-d: Generating high-quality surfaces of arbitrary topologies using diffusion models. In: *European Conference on Computer Vision*. pp. 419–438. Springer (2024)
105. Yu, Z., Li, T., Sun, J., Shapira, O., Park, S., Stengel, M., Chan, M., Li, X., Wang, W., Nagano, K., et al.: Gaia: Generative animatable interactive avatars with expression-conditioned gaussians. In: *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. pp. 1–10 (2025)
106. Zhang, J., Chen, W., Liu, Y., Wang, J., Yu, Z., Shen, Z., Yang, B., Wang, W., Li, X.: Spgen: Spherical projection as consistent and flexible representation for single image 3d shape generation. In: *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. pp. 1–12 (2025)
107. Zhang, J., Zhang, L., Liu, Q., Chiu, M.T., Barnes, C., Wang, Y., You, H., Liu, X., Zhou, Y., Lin, Z., et al.: Uniser: A foundation model for unified soft effects removal. arXiv preprint arXiv:2511.14183 (2025)
108. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 3836–3847 (2023)
109. 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)

110. Zhang, N., Ye, Y., Zhao, Y., Wang, R.: Revisiting the stack-based inverse tone mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9162–9171 (2023)
111. Zhang, Y., Aydın, T.O.: Deep hdr estimation with generative detail reconstruction. *Computer Graphics Forum* **40**(2), 179–190 (2021). <https://doi.org/https://doi.org/10.1111/cgf.142624>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.142624>
112. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404* (2024)
113. Zhou, Y., Bu, J., Ling, P., Zhang, P., Wu, T., Huang, Q., Li, J., Dong, X., Zang, Y., Cao, Y., et al.: Light-a-video: Training-free video relighting via progressive light fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13315–13325 (2025)
114. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747* (2025)