Single-Shot HDR Recovery via a Video Diffusion Prior

Chinmay Talegaonkar[*], **Jinshi He, Nicholas Antipa**
University of California San Diego

Abstract

Recent generative methods for single-shot high dynamic range (HDR) image reconstruction show promising results, but often struggle with preserving fidelity to the input image. They require separate models to handle highlights and shadows, or sacrifice interpretability by directly predicting the final HDR image. We address these limitations by re-casting single-shot HDR reconstruction as conditional video generation and fusing the generated frames into an HDR image. We fine-tune a video diffusion model to generate an exposure bracket, conditioned on a low dynamic range (LDR) input. We fuse this image bracket using per-pixel weights predicted by a light-weight UNet. This formulation is simple, interpretable, and effective. Rather than directly hallucinating an HDR image, it explicitly reconstructs the intermediate exposure stack and fuses it into the final output. Our method eliminates the need for separate models across exposure regimes and produces HDR reconstructions with high input fidelity. On quantitative benchmarks, we outperform state-of-the-art generative baselines with comparable model capacity on several reconstruction metrics. Human evaluators further prefer our results in 72% of pairwise comparisons against existing methods. Finally, we show that this input-conditioned sequence generation and fusion framework extends beyond HDR to other image reconstruction tasks, such as all-in-focus image recovery from a single defocus-blurred input.
`https://github.com/chinmay0301ucsd/Vid2HDRImg`

1 Introduction

High dynamic range (HDR) imaging mitigates the gap between digital imagery and human perception of the real world McCann and Rizzi (2011). It powers richer visual experiences in photography and displays, and supports high-fidelity imaging in scientific, medical, and industrial settings Cao et al. (2018). Capturing an HDR image usually involves acquiring multiple low dynamic range (LDR) images at different exposures and fusing them into a single HDR output Mantiuk et al. (2015); Debevec and Malik (1997); Kalantari et al. (2017); Wu et al. (2018); Jiang et al. (2023); Kong et al. (2024); Yan et al. (2019). While effective, this method introduces practical challenges due to motion during capture. Single-shot HDR methods recover an HDR image from a single LDR capture, eliminating temporal artifacts. However, single-shot HDR is fundamentally ill-posed, as it requires hallucinating missing information in the over/under-exposed regions of the input LDR measurement. We propose a two-stage pipeline for single-shot HDR reconstruction by reframing it as a conditional video generation problem. Our central hypothesis is that an exposure bracket is structurally analogous to a video where instead of scene motion, the temporal variation reflects changes in exposure. We observe that an off-the-shelf image-to-video generation model, with an appropriate prompt, produces coherent bracketed exposures (fig. 1). This suggests that video diffusion models Wan et al. (2025); Blattmann et al. (2023) can serve as strong priors for synthesizing plausible exposure stacks from a single image. Because video models allow inter-frame motion, which is undesirable in HDR imaging, we fine-tune a video diffusion model Blattmann et al. (2023) to generate a motion-free LDR bracket. Unlike recent generative single-shot HDR methods Wang et al. (2025); Bemana et al. (2025); Wu et al. (2026), which rely only on spatial priors, our approach leverages both spatial and temporal

[*]Corresponding author: `ctalegaonkar@ucsd.edu`



(a) Input Image (b) Generated Exposure Stack

Figure 1: **Video diffusion models contain implicit priors to generate exposure bursts.** Given an input LDR image (a) captured at a high exposure, we prompt an off-the-shelf video model (Runway Gen 4): "Reduce the brightness of the image over the course of the video, generating an exposure burst stack from high to low exposure". The resulting video frames are plotted in (b), demonstrating that the video model contains strong lighting priors to simulate an exposure bracket.

priors by using a video generation model to generate a plausible image sequence conditioned on the input LDR image. We then fuse the synthesized LDR bracket using per-pixel blend weights predicted by a small network. On the standard SI-HDR benchmark Hanji et al. (2022), our method outperforms generative baselines with comparable model capacity on several reconstruction metrics and remains competitive on the remaining metrics. Human evaluators further prefer our results in 72% of pairwise comparisons against baseline methods. Our framework of input-conditioned image sequence generation followed by light-weight fusion naturally extends to other image recovery tasks. We demonstrate this with all-in-focus recovery from a single defocus-blurred image.

2 Related Work

Multi-shot HDR Imaging: Classical HDR methods capture successive LDR images (exposure bracket) of a scene at different exposure settings. Debevec Merging Debevec and Malik (1997) recovers an HDR radiance map from an exposure bracket with known relative exposure times. Mertens Fusion Mertens et al. (2007) blends the exposure stack using per-pixel weights based on contrast, well-exposedness, and saturation. These methods require multiple aligned captures, making them less suitable for dynamic scenes and handheld capture. Subsequent learning-based methods address motion artifacts using learned alignment, attention, or fusion networks Kalantari et al. (2017); Wu et al. (2018); Yan et al. (2019); Niu et al. (2021). However, they still require capturing an exposure burst. In contrast, we infer the exposure bracket from a single LDR image.

Single-Image HDR Reconstruction: Single-shot HDR methods predict HDR from a single LDR capture. HDRCNN Eilertsen et al. (2017) and ExpandNet Marnerides et al. (2018) trained convolutional networks to hallucinate missing information in over-exposed regions. SingleHDR Liu et al. (2020) decomposed the camera imaging pipeline and learned to invert each stage, while Le et al. (2023) proposed generating multiple virtual exposures from a single input in a weakly supervised manner. Several works Dai et al. (2025); Nayar and Mitsunaga (2000); Metzler et al. (2020); Sun et al. (2020); Shi et al. (2024); So et al. (2022) have explored hardware modifications for single-shot HDR capture, using specialized optics and sensors. In contrast, our method can be applied to raw images captured by standard cameras. Prior methods infer exposure brackets from a single LDR image and merge them into HDR images using GAN or CNN-based approaches (Endo et al., 2017; Lee et al., 2018). We recast multi-exposure bracket inference as a video generation problem, leveraging the temporal consistency of a video diffusion model.

More recently, diffusion models have shown strong potential for single-image HDR generation. LEDiff Wang et al. (2025) uses latent diffusion to separately generate highlight and shadow exposures, which are then fused into an HDR image. UltraFusion Chen et al. (2025b) formulates exposure fusion as a guided inpainting problem using a diffusion prior, enabling fusion across large exposure gaps. X2HDR Wu et al. (2026) adapts pre-trained diffusion models to HDR generation by operating in a perceptually uniform color space, while UltraLED Meng et al. (2025) works in the raw domain for ultra-high dynamic range recovery. Notably, these generative approaches rely on *image* diffusion models. In contrast, our work leverages *video* diffusion models. This is motivated by the observation that HDR recovery can be viewed as generating an ordered sequence of images with monotonically increasing exposures, which aligns with the temporal structure modeled by video priors.

Video Diffusion Models for Computational Photography: Recent work shows that video generation models can be repurposed for computational photography. VDM-EVFI Chen et al. (2025a) adapts video diffusion priors for event-based video frame interpolation. Blur2Vid Tedla et al. (2025b) and Learning to Refocus Tedla et al. (2025a) fine-tune video models to recover sharp videos from a single motion-blurred image and to refocus images, respectively. These works suggest that video diffusion models encode useful priors over visual attributes that vary across frames. We build on this idea by modeling an HDR exposure stack as a video-like sequence with monotonically increasing exposure. Concurrent works Yu et al. (2026); Korem et al. (2026) fine-tune video models to generate HDR videos from LDR videos. In contrast, we use video diffusion for single-image HDR reconstruction by emulating the multi-exposure capture process.

3 Preliminaries

HDR Imaging: Camera sensors capture only a limited portion of the full dynamic range of a scene. Consequently, images acquired at very low or very high exposure times exhibit underexposure (lose shadow details) or oversaturation artifacts (lose highlight details), respectively. Formally, an image (I) captured with an exposure time t is related to the underlying scene radiance I_{HDR} as

$$I = g(\text{clamp}(t \cdot I_{HDR} + \eta, I_-, I_+)), \tag{1}$$

where I_- and I_+ denote the lower and upper saturation thresholds of the sensor, respectively. The function $g(\cdot)$ denotes a combination of the camera response function (CRF) and sensor quantization, and η approximates the sensor read noise and photon shot noise. As a result, different regions of the captured image may become under-exposed or over-exposed depending on the chosen exposure time. High exposures can lead to saturation, while low exposures produce noisy quantized measurements. We approximate the CRF as a gamma mapping $g(x) = x^{1/\gamma}$, where γ denotes the gamma parameter of the linear-to-sRGB transformation Chang et al. (2025).

In addition to capture-time limitations, the limited dynamic range of conventional displays requires HDR images to be *tone-mapped* for visualization Reinhard (2002). Tone-mapping applies a contraction mapping to the HDR image to reduce its dynamic range to match the display.

Recovering the full dynamic range of a scene, therefore, typically requires capturing and fusing an exposure bracket $\{I_{t_1}, I_{t_2}, I_{t_3}, \ldots, I_{t_n}\}$, to reconstruct the HDR image I_{HDR}:

$$I_{HDR} = f(I_{t_1}, I_{t_2}, I_{t_3} \ldots I_{t_n}) \tag{2}$$

The fusion function $f(\cdot)$ can be a classical Mertens et al. (2007); Debevec and Malik (1997) or a learning-based approach Kalantari et al. (2017); Yan et al. (2019); Niu et al. (2021). Single-shot HDR methods are an extreme version of this problem, where $n = 1$.

Latent Video Diffusion Models: Latent video diffusion models Blattmann et al. (2023); Wan et al. (2025) consist of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, which map an RGB video $\mathbf{x}$ to and from a lower-dimensional latent space, respectively. A diffusion model $\phi(\cdot)$ is then trained to operate in this latent space. Given an input condition c, such as a text prompt or an image, inference starts with a randomly sampled noise vector $\mathbf{z}_K \sim \mathcal{N}(0, I)$. The diffusion model $\phi(\cdot)$ then iteratively denoises $\mathbf{z}_K$ over K steps to produce a clean latent $\mathbf{z}_0$. The clean latent is decoded into an RGB video as $\mathcal{D}(\mathbf{z}_0)$. In this work, we use Stable Video Diffusion Blattmann et al. (2023), which is an image-conditioned latent video diffusion model. Training follows the EDM framework Karras et al. (2022). Given a clean latent $\mathbf{z} = \mathcal{E}(\mathbf{x})$, Gaussian noise is added at a denoising-step-dependent noise level σ to obtain

$$\mathbf{z}_\sigma = \mathbf{z} + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I). \tag{3}$$

The denoised latent is predicted using the EDM parameterization

$$\hat{\mathbf{z}} = c_{\text{skip}}(\sigma)\, \mathbf{z}_\sigma + c_{\text{out}}(\sigma)\, f_\theta(\mathbf{z}_\sigma, \sigma, c), \tag{4}$$

where f_θ denotes the spatio-temporal denoiser and c the conditioning signal. The coefficients $c_{\text{skip}}(\sigma)$ and $c_{\text{out}}(\sigma)$ are noise-dependent preconditioning terms that modulate the contributions of the noisy latent and the denoiser output. The training process minimizes

$$\mathcal{L} = \mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon},\sigma}\left[w(\sigma)\, \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2\right], \tag{5}$$

with $w(\sigma)$ being a noise-dependent weighting function.

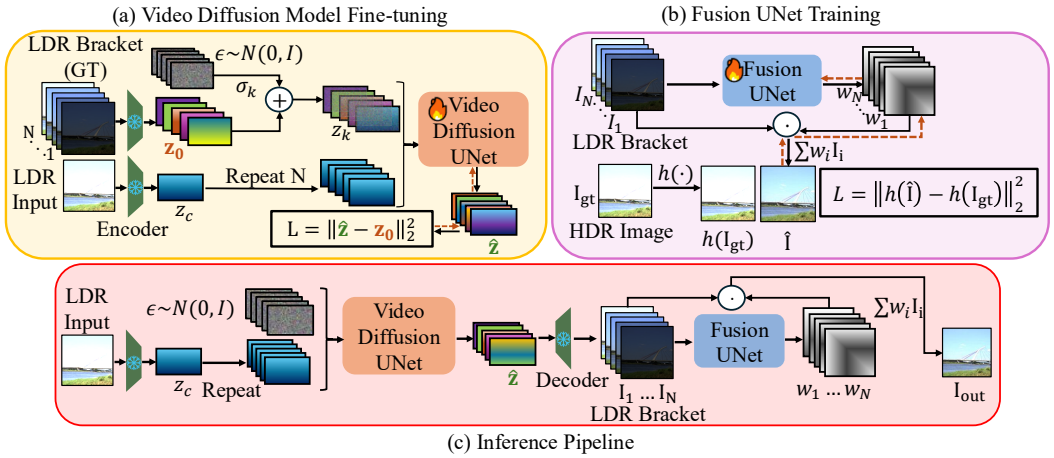

Figure 2: **Method overview.** (a) Given an input LDR image, we fine-tune a latent video diffusion model to generate an LDR exposure bracket corresponding to the scene. (b) A fusion UNet then predicts per-pixel weights to fuse this bracket into an HDR image. The fusion network is supervised in PU-21 encoded space, and at (c) inference time, we sample the video diffusion model to synthesize an LDR bracket, which is fused by the Fusion UNet to obtain the final HDR output.

4 Method

We recast single-shot HDR reconstruction as conditional exposure-bracket generation from a single LDR image. Since adjacent exposures form a coherent video-like sequence, we fine-tune a video diffusion model to synthesize the full exposure bracket. During training, each HDR image is converted into an LDR exposure bracket comprising monotonically increasing exposures. As a result, the model always learns a canonical mapping from low to high exposures along the temporal axis, even when the observed conditioning frame is not the first frame in the sequence. We then train a separate, smaller network (Fusion UNet) to predict per-pixel weights for fusing the LDR frames generated by the fine-tuned video model. Below, we outline our two-stage training process (fig. 2), which consists of fine-tuning the video model and training the Fusion UNet.

Video model fine-tuning algorithm: Given a single input LDR image at an arbitrary exposure, we fine-tune the video diffusion model to generate a plausible sequence of LDR images with monotonically increasing exposures that covers the scene's dynamic range. Training data consists of pairs of raw images and their corresponding ground truth (GT) HDR images, sourced from Fairchild (2007); Zou et al. (2023), following Wu et al. (2026). Given a GT HDR image I_{HDR}, the corresponding LDR image captured at an exposure t can be obtained as

$$I_t = (\text{clamp}(t \cdot I_{HDR}, 0, 1))^{1/\gamma} \tag{6}$$

We use $\gamma = 2.2$ as a simple approximation to the nonlinear tone response of sRGB-like images, which dominate the internet-scale pretraining data of the base video model. We generate the GT LDR bracket over an exposure range $[t_s, t_e]$ determined from the GT HDR image. To compute these exposure bounds, we first estimate the luminance map

$$Y = 0.2126\, I^r_{HDR} + 0.7152\, I^g_{HDR} + 0.0722\, I^b_{HDR},$$

where $I^{\{r,g,b\}}_{HDR}$ denote the red, green, and blue channels of I_{HDR}. Following prior work Wang et al. (2025), we set

$$t_s = \frac{v_s^\gamma}{Y_{\max}}, \qquad t_e = \frac{v_e^\gamma}{Y_{\text{median}}}, \tag{7}$$

where $Y_{\max}$ and Y_{median} are the maximum and median values of Y. The target LDR intensities $v_s, v_e \in (0, 1]$ control the dynamic range mapping, and are both set to 0.85, following Wang et al. (2025); Andersson et al. (2021) in all experiments. This choice maps the brightest pixel and the median-luminance pixel to v_s and v_e, respectively, after applying the γ curve. We then sample $N = 5$

exposure times uniformly in log-space between $\log_2 t_s$ and $\log_2 t_e$, and generate the corresponding noiseless LDR bracket using eq. (6).

To fine-tune the video model, we first encode the raw LDR input image I_{cond} into a latent representation $z_{\text{cond}} = \mathcal{E}(I_{\text{cond}})$, and encode the simulated ground-truth LDR exposure bracket into latent representations $z^1, z^2, \ldots, z^N$. Importantly, the diffusion-model input $\mathbf{z}_\sigma = [z^1_\sigma, z^2_\sigma, \ldots, z^N_\sigma]$ is constructed by adding noise only to the clean exposure-bracket latents and stacking the resulting noisy latents along the exposure dimension. The model is conditioned on z_{cond}, corresponding to the input raw image I_{cond}. We do not use classifier-free guidance Ho and Salimans (2022) (CFG) during inference and accordingly disable conditional dropout during training. This follows prior work on repurposing diffusion models for downstream vision tasks Ke et al. (2024). Unlike text-conditioned generation, where CFG is often used to improve sample quality and diversity, our task requires strong pixel-level consistency with the input condition. In this setting, we find that omitting CFG leads to better reconstruction performance.

For fine-tuning the video diffusion UNet, we concatenate z_{cond} and the noisy latent $\mathbf{z}_\sigma$ and feed the resulting tensor to the model as input, similar to prior conditioning schemes used in related work Chen et al. (2025a); Tedla et al. (2025a). Specifically, we broadcast z_{cond} along the exposure dimension and channel-wise concatenate it with $\mathbf{z}_\sigma$ along the feature dimension. The diffusion UNet is then trained to minimize eq. (5). Compared to prior HDR reconstruction methods that either rely on separate models for shadow and highlight recovery Wang et al. (2025) or directly predict HDR-space representations from a single image Wu et al. (2026), our approach retains simplicity and interpretability by using a single model to generate the full LDR exposure bracket, which is then fused into the final HDR image using per-pixel weights.

Fusion UNet: We train a separate fusion network U_θ that takes as input an LDR exposure bracket $\mathbf{I} = [I_1, I_2, \ldots, I_N]$ and predicts a corresponding stack of per-pixel fusion weights $\mathbf{W} = [W_1, W_2, \ldots, W_N]$ while also applying a softmax over the N generated exposure frames such that $\sum_{i=1}^{N} W_i(\mathbf{p}) = 1 \quad \forall \mathbf{p}$, where $\mathbf{p}$ indexes pixel location. The fusion network is trained independently using LDR brackets synthesized from clean ground-truth HDR images I_{HDR} via the HDR-to-LDR mapping in eq. (6) and the exposure sampling strategy in eq. (7). As a result, it is trained on the same target distribution of LDR brackets as the video diffusion model. The fused prediction is then given by

$$\hat{I} = \sum_{i=1}^{N} W_i \odot I_i, \tag{8}$$

where $\odot$ denotes element-wise multiplication. The input LDR frames are linearized during both training and inference using the inverse gamma mapping, $x \to x^\gamma$, so that the fused output is a linear radiance image. We scale the ground-truth HDR I_{HDR} by the inverse of its peak luminance, resulting in a normalized $\tilde{I}_{HDR}$. As a result, our model predicts HDR only up to scale, following prior work Wang et al. (2025); Wu et al. (2026). We train the fusion network by minimizing

$$\mathcal{L}_{\text{fusion}}(\theta) = \left\| h(\hat{I}) - h(\tilde{I}_{\text{HDR}}) \right\|_2^2, \tag{9}$$

where $h(\cdot)$ denotes the PU-21 transformation Wu et al. (2026); Ke et al. (2025)

$$h(L) = a\big(\log_2 L - L_{\min}\big)^2 + b\big(\log_2 L - L_{\min}\big), \tag{10}$$

which maps HDR intensities to a normalized perceptual space with greater precision in low-luminance regions. We use the same hyperparameter values $(a, b, L_{\min})$ as prior work Wu et al. (2026). Other log-based HDR transformations Kalantari et al. (2017); Yu et al. (2026) could also be used in place of PU-21. This loss encourages the fusion network to predict weights that produce perceptually and quantitatively accurate HDR reconstructions up to a scaling factor.

Why a light-weight fusion network? We intentionally choose a light-weight fusion network that predicts only per-pixel fusion weights, restricting fusion to a linear aggregation of the generated exposure bracket. We show the benefit of this design over directly estimating the HDR image by the fusion network in Appendix C. Strong HDR reconstruction under this constrained design provides indirect evidence that the synthesized brackets contain useful HDR information. We also include additional analysis and visualizations of these synthesized exposure brackets by the fine-tuned video model in Appendix E.

Implementation Details: All models are trained at an image resolution of 512×512. We use Stable Video Diffusion (SVD) as the video diffusion backbone with $N = 5$ frames per sequence. We fine-tune the diffusion UNet (1.5B parameters) of SVD for 16K iterations using AdamW with a learning rate of 5×10^{-5} and 100 warm-up steps. Training uses full FP32 precision with gradient clipping at 1.0. We use 8 AMD MI300X GPUs with a batch size of 1 per GPU and gradient accumulation over 8 steps. The Fusion UNet is a small UNet-based model Ronneberger et al. (2015) with ~ 0.5M parameters, trained for 25K iterations on 4 MI250 GPUs with total batch size 64 and learning rate 1×10^{-5}. Total training time is approximately 36 hours for the video diffusion UNet and 12 hours for the Fusion UNet. During inference, our method requires approximately 13.5 GB of GPU VRAM.

5 Experiments and Results

Training and evaluation data: We generate training data by combining HDR images from the same sources as X2HDR Wu et al. (2026), resulting in a total of approximately 11K training crops. Please see Appendix A for more details on the crop extraction process. We use the SI-HDR dataset Hanji et al. (2022) for evaluation, adopting the same subset as X2HDR Wu et al. (2026). We also evaluate our method on the MobileHDR Liu et al. (2023) test set. The scenes cover diverse indoor, outdoor, natural, and urban content. Notably, some of the raw images contain line noise artifacts, which pose an additional challenge for HDR reconstruction.

Quantitative HDR evaluation: We follow the evaluation criteria Cao et al. (2024) (Q^* metrics) used in prior work Wu et al. (2026). The Q^* metrics circumvent the global exposure difference between the ground truth and the predicted HDR image by optimizing exposure values that maximize performance for a given metric. We evaluate Q^* metrics for PSNR, SSIM, LPIPS, and MAE. Additionally, we compute the Q-JOD metric from HDR-VDP3 Mantiuk et al. (2023), which is known to correlate better with human preference. See Appendix B for more evaluation details. We compare our method with several kinds of baselines: diffusion-based generative methods with comparable model capacity, such as X2HDR Wu et al. (2026) and LEDiff Wang et al. (2025); BracketDiff Bemana et al. (2025), which is a diffusion-based inference-time optimization method; and learning-based feedforward methods like SingleHDR Liu et al. (2020) and HDRCNN Eilertsen et al. (2017).

We show in table 1 that our method achieves strong performance against state-of-the-art generative baselines, outperforming them on several reconstruction metrics (Q^*-PSNR, Q^*-MAE, and Q^*-LPIPS). While BracketDiff achieves the best Q^*-SSIM, it is $25\times$ slower than our method. Non-generative feedforward methods such as HDRCNN and SingleHDR are faster and perform marginally better on some metrics (Q-JOD, Q^*-SSIM), but are less robust in challenging exposure regimes. In under-exposed scenes, they often fail to suppress raw sensor noise (see fig. 3), while in over-exposed regions they can hallucinate unrealistic structure (see fig. 4). By training directly on raw inputs, our method better handles noisy under-exposed images while maintaining scene consistency across the synthesized exposure bracket. Although X2HDR is also trained on similar raw-paired data, our method better preserves plausible scene structure in saturated regions (row 4, fig. 3) and achieves stronger aggregate performance. Overall, these results highlight the strength of our method in challenging ill-exposed regimes.

For the X2HDR baseline, we adopt their direct HDR-generation formulation with Stable Diffusion 2.1 (SD 2.1) as the image-diffusion backbone. Following X2HDR, we train the model to predict PU-21-encoded HDR latents. We use SD 2.1 because it shares the same backbone family as Stable Video Diffusion, enabling a controlled comparison. Since X2HDR uses a Flux-based backbone Labs (2024), this SD 2.1 variant helps isolate the effect of the reconstruction formulation from differences in backbone capacity. We report the Flux-based X2HDR results in Appendix K.

Discussion: Beyond the quantitative gains, generating an explicit exposure bracket gives our method an interpretable intermediate representation. Visualizing the generated bracket helps diagnose failure cases, such as insufficient dynamic range across exposures (fig. 9, row 3). The bracket also enables direct control over exposure spacing through inference-time optimization. Improving ΔEV consistency between adjacent frames (table 4, section 5.1) makes the generated bracket compatible with fusion methods such as Debevec and Malik (1997), which assume known exposure spacing.

Compared to LEDiff, which requires separate models for shadow and highlight recovery, we use a single video diffusion model to synthesize the full exposure sequence, covering both low- and

Method	Q^*-PSNR ↑	Q^*-SSIM ↑	Q^*-MAE ↓	Q^*-LPIPS ↓	Q-JOD ↑	Time ↓
Ours	$\mathbf{22.36 \pm 3.48}$	0.664 ± 0.156	$\mathbf{0.060 \pm 0.027}$	$\mathbf{0.289 \pm 0.111}$	7.51 ± 0.99	∼15s
LEDiff	18.97 ± 3.07	0.552 ± 0.128	0.098 ± 0.043	0.384 ± 0.128	6.55 ± 0.98	∼15s
BracketDiff	21.41 ± 3.87	$\mathbf{0.734 \pm 0.128}$	0.077 ± 0.044	0.353 ± 0.161	7.51 ± 1.21	∼6min
X2HDR-SD2.1	21.24 ± 3.32	0.585 ± 0.150	0.068 ± 0.030	0.335 ± 0.120	7.17 ± 0.97	∼15s
X2HDR-Flux	21.37 ± 3.10	$\mathbf{0.726 \pm 0.108}$	0.067 ± 0.025	$\mathbf{0.232 \pm 0.106}$	$\mathbf{7.59 \pm 1.24}$	∼11s
HDRCNN	21.50 ± 6.27	0.601 ± 0.281	0.098 ± 0.073	0.342 ± 0.246	$\mathbf{7.66 \pm 1.68}$	<1s
SingleHDR	21.61 ± 3.93	0.698 ± 0.176	0.075 ± 0.039	0.302 ± 0.180	7.56 ± 1.40	<1s

Table 1: **Quantitative HDR comparison on SI-HDR dataset.** Our method outperforms the generative baselines on most quantitative metrics (PSNR, MAE, LPIPS), while achieving competitive performance on the remaining metrics. We compare favorably with non-generative feedforward methods (HDRCNN, SingleHDR) while avoiding some of their failure modes in severely under- and over-exposed regions. X2HDR-Flux uses a 12B parameter DiT backbone compared to our 1.5B parameter UNet backbone, hence is not a capacity-matched baseline.

Method	Q*-PSNR ↑	Q*-SSIM ↑	Q*-MAE ↓	Q*-LPIPS ↓	Q-JOD ↑
Ours	$\mathbf{25.15 \pm 2.67}$	$\mathbf{0.798 \pm 0.143}$	$\mathbf{0.0390 \pm 0.0121}$	0.221 ± 0.058	$\mathbf{8.251 \pm 0.693}$
LEDiff	18.96 ± 3.10	0.575 ± 0.134	0.1175 ± 0.1398	0.361 ± 0.101	7.042 ± 0.427
BracketDiff	24.08 ± 2.72	0.765 ± 0.132	0.0462 ± 0.0181	0.267 ± 0.074	8.052 ± 0.974
X2HDR-SD2.1	23.38 ± 2.28	0.668 ± 0.133	0.0460 ± 0.0132	0.317 ± 0.065	7.742 ± 0.445
X2HDR-Flux	21.22 ± 2.45	0.796 ± 0.090	0.0646 ± 0.0186	$\mathbf{0.209 \pm 0.063}$	6.881 ± 0.918
HDRCNN	23.62 ± 3.20	0.605 ± 0.171	0.0536 ± 0.0277	0.380 ± 0.133	7.654 ± 0.889
SingleHDR	23.63 ± 3.69	0.700 ± 0.151	0.0514 ± 0.0224	0.313 ± 0.110	7.559 ± 0.697

Table 2: **Quantitative HDR comparison on MobileHDR dataset.** We outperform all capacity-comparable baselines across all metrics. X2HDR-Flux achieves slightly lower Q*-LPIPS but uses a substantially larger backbone and is therefore not capacity-matched to our method. See Appendix B for MobileHDR evaluation details and fig. 8 for qualitative comparisons.

high-exposure LDR images. In contrast to X2HDR, which directly predicts HDR latents in a PU-21-transformed space, our method generates an explicit LDR bracket, keeping the generative process in the image space where latent diffusion models are natively trained.

User Study: We evaluate the qualitative preference rate of our method against the baselines through a user study on the SI-HDR dataset. We display our output alongside a randomly chosen baseline from table 1, with all method labels hidden. We then show the ground-truth HDR image and ask the participant to select the image that best matches it. Each participant completes 96 randomized pairwise comparisons. Across 20 participants, our method achieves a 72% preference rate in pairwise comparisons. A two-sided binomial test against chance (p_0=0.5) confirms a statistically significant preference for our method against every baseline ($p < 10^{-4}$), as shown in table 3. We use a MacBook Pro with an HDR display and a peak brightness of 1600 nits. See Appendix I for more details on the user-study setup and the user-study results on the MobileHDR test set. In Appendix L, we provide additional details on the Q-JOD metric and discuss its interpretation in context of the user-study.

	LEDiff	HDRCNN	SingleHDR	BracketDiff	X2HDR-SD2.1	Overall
Win rate (Ours) ↑	86.2%	73.6%	71.5%	67.6%	60.8%	**72.0%**
Binomial p ↓	7.2×10^{-51}	1.1×10^{-20}	1.7×10^{-17}	4.6×10^{-12}	2.7×10^{-5}	$\mathbf{2.8\times10^{-85}}$

Table 3: **User-study results on SI-HDR test set.** Participants compared our reconstructions against each baseline using the GT HDR image as a reference. We report the percentage of trials preferring ours, with *Overall* aggregated over 1920 pairwise comparisons. Two-sided exact binomial tests against chance (p_0=0.5) show a significant preference rate for our method against the baselines.

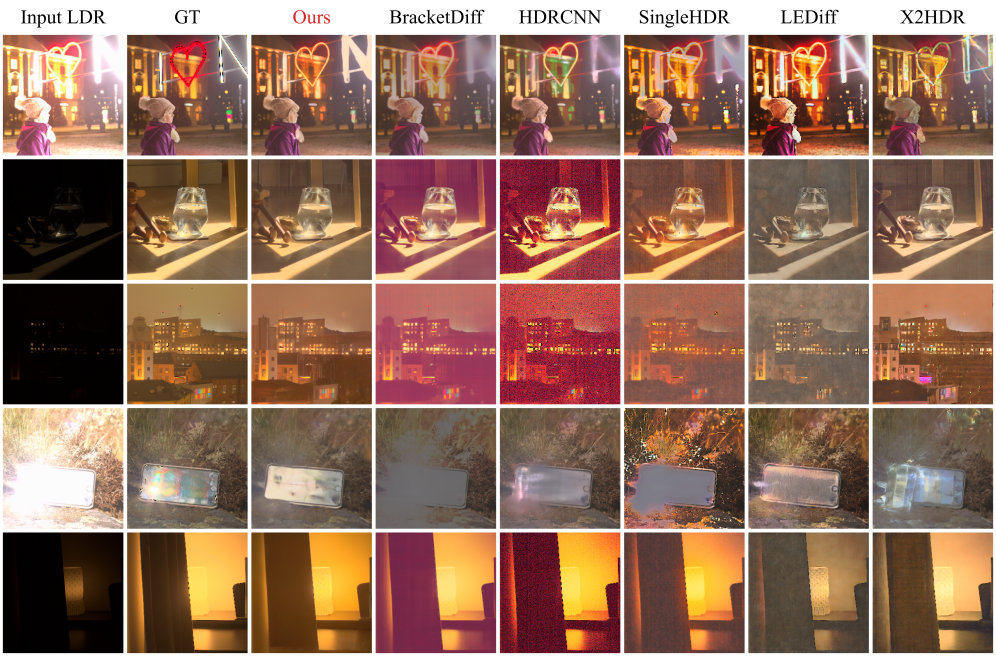

Figure 3: **Qualitative HDR reconstruction results.** We show challenging ill-exposed examples from the SI-HDR test set where our method produces higher-fidelity HDR reconstructions than the baselines. Our method better denoises underexposed regions (rows 2, 3, 5) and generates more plausible content in overexposed regions (rows 1, 4). All results are normalized to the same median luminance and visualized with Reinhard tone mapping Reinhard (2002).

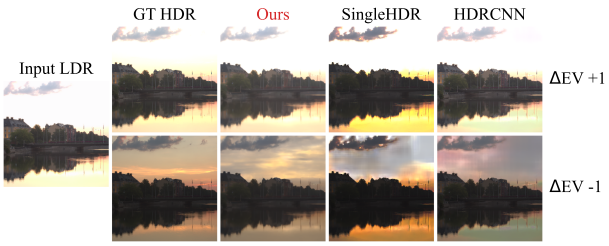

Figure 4: **Non-generative methods struggle with over-exposed regions.** Our generative prior produces plausible details for over-exposed regions as visible at a lower exposure image (row 2), while SingleHDR and HDRCNN produce artifacts (zoom in on clouds).

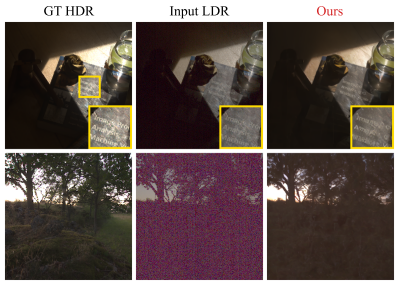

Figure 5: **Failure cases**. Our method can struggle with fine text (row 1), as the VAE encoding can lose out on small, high-frequency details. Extremely dark inputs (row 2) remain challenging for all methods due to a low signal-to-noise ratio.

5.1 Ablation Studies

Fusion strategy ablation: We ablate different strategies for fusing the generated LDR images by replacing our Fusion UNet with two widely used classical methods: Mertens fusion Mertens et al. (2007) and Debevec merging Debevec and Malik (1997). Measured over the test set, our trained Fusion UNet (Q^*-PSNR: 22.36) outperforms both Mertens fusion (Q^*-PSNR: 18.87) and Debevec merging (Q^*-PSNR: 15.23). All fusion methods receive the exposure brackets generated by the fine-tuned model as input. Debevec merging underperforms Mertens fusion because it relies on known relative exposure values, which are not predicted by our method. Our Fusion UNet is instead trained on LDR brackets synthesized from the GT exposure distribution used to fine-tune the video

Model	HDR Merge	Q^*-PSNR ↑	Q^*-SSIM ↑	Q^*-MAE ↓	Q^*-LPIPS ↓
Base Model	Debevec	17.82	0.52	0.11	0.44
Ours	**Debevec**	**21.36**	**0.62**	**0.06**	**0.34**

Table 4: **Comparison under inference-time optimization.** After enforcing exposure-consistent brackets before Debevec merging, our model outperforms the base SVD model.

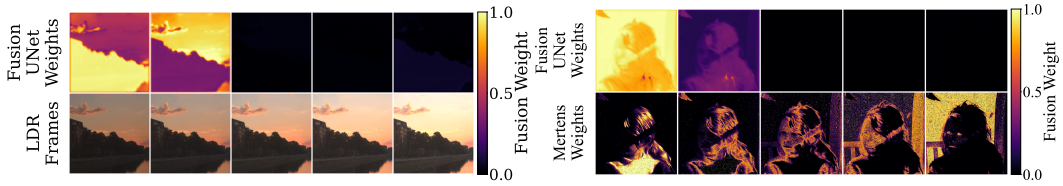

Figure 6: **Fusion method comparison.** (Left) UNet fusion weights on a test scene. (Right) Visual comparison of Mertens vs. UNet fusion on a test scene. Our Fusion UNet generates spatially smoother and less noisy fusion weights as compared to Mertens.

model, allowing it to learn a fusion strategy better suited to our predicted LDR brackets. It also produces spatially smoother weighting maps than Mertens fusion (see fig. 6). Please see Appendix C for more details and additional Fusion UNet ablations.

Bracket Consistency: While our current model is trained to generate monotonically increasing exposures, it does not explicitly enforce a fixed ΔEV between consecutive frames, where $\Delta EV = \log_2(t_i/t_{i-1})$ denotes the log-ratio of their exposure times (t_i, t_{i-1}). Known exposure spacing enables radiometrically consistent HDR recovery with methods such as Debevec merging. To achieve controlled exposure spacing, we further demonstrate that our method is compatible with inference-time optimization. Following BracketDiff, we optimize initial noise latents corresponding to the generated LDR bracket at inference time so that consecutive frames follow a monotonic 1-EV spacing, i.e., $\Delta EV = 1$. Table 4 shows that inference-time optimization is substantially more effective with our fine-tuned model than with the base Stable Video Diffusion (SVD) model.

We also achieve better bracket consistency between consecutive bracket images in comparison to the original SVD model after inference time optimization. We measure bracket consistency by comparing the relative brightness changes in the generated LDR bracket against the expected exposure spacing of a valid bracket. Since inference-time optimization enforces $\Delta EV = 1$ between adjacent LDR frames, corresponding well-exposed regions should exhibit an ideal brightness ratio of $2\times$. We therefore measure the mean absolute deviation of the observed adjacent-frame brightness ratios from the ideal value of 2, where lower is better. Averaged over the test set, the fine-tuned model achieves a lower deviation (0.233) than the base SVD model (0.326), corresponding to an approximately 29% improvement in bracket consistency. See Appendix G for details on bracket consistency evaluation.

In Appendix D, we additionally report results for LoRA-based fine-tuning and using classifier-free guidance (CFG) during both fine-tuning and inference. In Appendix F, we show that reconstruction quality is primarily limited by the generated LDR bracket, rather than the Fusion UNet.

5.2 Beyond Single-Shot HDR

We present a proof of concept that our framework can be adapted to other computational photography tasks, such as all-in-focus image recovery from a single defocus-blurred input in fig. 7. We fine-tune the video model to generate a focal stack with monotonically varying focus depth, analogous to the monotonic exposure progression in the HDR case. We then train a Fusion UNet on the ground-truth focal stacks to fuse them into an all-in-focus image. We simulate training data using RGB images and depth maps from the NYUv2 dataset Silberman et al. (2012), by rendering focal stacks with a simple depth-aware defocus blur model Gur and Wolf (2019). See Appendix H for more details.

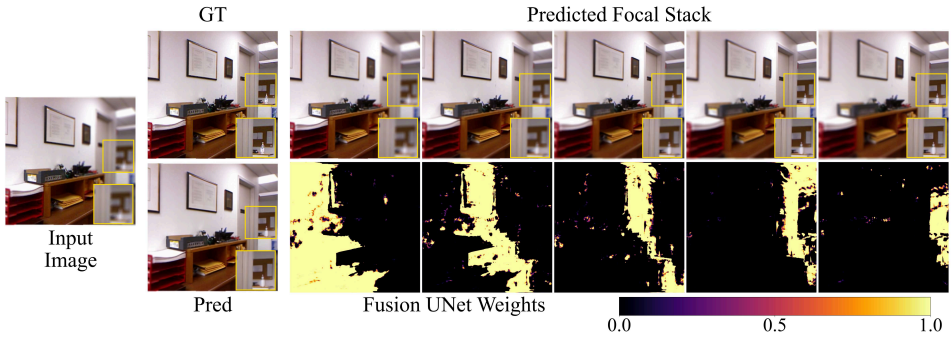

Figure 7: **Extending our method to all-in-focus (AIF) image recovery.** We fine-tune the video model to take a defocused image as input and generate a focal stack of images with varying focus distances. We fuse the generated focal stack into an AIF image using a separate Fusion UNet.

6 Limitations and Future Work

We recast single-shot HDR reconstruction as conditional video generation, where the generated frames form an exposure bracket that is fused into an HDR image using a lightweight network. This results in a simple and interpretable framework that is compatible with other computational photography tasks. However, our method is not without limitations. As with any learned generative model, empirical suppression of scene motion and dynamic objects does not provide a formal guarantee that such artifacts will never occur. Scenes containing fine text or other high-frequency details (see fig. 5) can be degraded by the latent compression introduced by the Stable Video Diffusion VAE. In the trivial case, when the shadows and highlights are well captured in the input image, our method may slightly lag behind existing baselines. Since the input already preserves most of the scene content, such cases require minimal to no dynamic range expansion. We present a thorough analysis of these failure cases in Appendix J. These failure modes could be mitigated by training on more diverse datasets and stronger video generation backbones Wan et al. (2025). Runtime can be substantially reduced by one-step distillation Yin et al. (2024), thereby bridging the gap between our approach and non-generative methods. Extending this framework to other reconstruction tasks, such as hyperspectral imaging, is a promising direction.

Acknowledgments and Disclosure of Funding

We thank Ishit Mehta and Ganesh Iyer for helpful feedback on the draft; Yash Belhe for suggesting useful experiments; Sai Tedla for very helpful feedback and discussions during the rebuttal phase; and Xiang Dai for discussions on HDR imaging. We also gratefully acknowledge the Machine Intelligence, Computing, and Security (MICS) Center at UCSD for GPU and storage support. This work was supported in part by NASA under SBIR Award 318257-00001.

References

Pontus Andersson, Jim Nilsson, Peter Shirley, and Tomas Akenine-Möller. Visualizing errors in rendered high dynamic range images. 2021.

Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Bracket diffusion: Hdr image generation by consistent ldr denoising. In *Computer Graphics Forum*, volume 44, page e70086. Wiley Online Library, 2025.

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

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237. IEEE, 2018.

Peibei Cao, Rafal K Mantiuk, and Kede Ma. Perceptual assessment and optimization of hdr image rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22433–22443, 2024.

Zhaolou Cao, Fenping Cui, and Chunjie Zhai. Vision system with high dynamic range for optical surface defect inspection. *Appl. Opt.*, 57(34):9981–9987, Dec 2018. doi: 10.1364/AO.57.009981. URL https://opg.optica.org/ao/abstract.cfm?URI=ao-57-34-9981.

Chen-Wei Chang, Cheng-De Fan, Chia-Che Chang, Yi-Chen Lo, Yu-Chee Tseng, Jiun-Long Huang, and Yu-Lun Liu. Gcc: Generative color constancy via diffusing a color checker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10868–10878, 2025.

Jingxi Chen, Brandon Y Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan, Cornelia Fermuller, Christopher A Metzler, and Yiannis Aloimonos. Repurposing pre-trained video diffusion models for event-based video interpolation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12456–12466, 2025a.

Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. Ultrafusion: Ultra high dynamic imaging using exposure fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16111–16121, 2025b.

Xiang Dai, Kyrollos Yanny, Kristina Monakhova, and Nicholas Antipa. Single-shot hdr using conventional image sensor shutter functions and optical randomization. *ACM Transactions on Graphics*, 44(5):1–20, 2025.

Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, page 369–378, USA, 1997. ACM Press/Addison-Wesley Publishing Co. ISBN 0897918967. doi: 10.1145/258734.258884. URL https://doi.org/10.1145/258734.258884.

Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *arXiv preprint arXiv:2303.11435*, 2023.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020. URL https://arxiv.org/abs/2004.07728.

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017.

Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph*, 36(6):1–10, 2017.

Mark D Fairchild. The hdr photographic survey. In *Color and imaging conference*, volume 15, pages 233–238. Society of Imaging Science and Technology, 2007.

Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7683–7692, 2019.

Param Hanji and Rafał K Mantiuk. Robust estimation of exposure ratios in multi-exposure image stacks. *IEEE Transactions on Computational Imaging*, 9:721–731, 2023.

Param Hanji, Rafal Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. Comparison of single image hdr reconstruction methods—the caveats of quality assessment. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–8, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ting Jiang, Chuan Wang, Xinpeng Li, Ru Li, Haoqiang Fan, and Shuaicheng Liu. Meflut: Unsupervised 1d lookup tables for multi-exposure image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10542–10551, 2023.

Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Andrew Yanzhe Ke, Lei Luo, Xiaoyu Xiang, Yuchen Fan, Rakesh Ranjan, Alexandre Chapiro, and Rafal Mantiuk. Training neural networks on raw and hdr images for restoration tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 810–819, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9492–9502, 2024.

Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen. Safnet: Selective alignment fusion network for efficient hdr imaging. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Naomi Ken Korem, Mohamed Oumoumad, Harel Cain, Matan Ben Yosef, Urska Jelercic, Ofir Bibi, Yaron Inger, Or Patashnik, and Daniel Cohen-Or. Hdr video generation via latent alignment with logarithmic encoding. *arXiv preprint arXiv:2604.11788*, 2026.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. Single-image hdr reconstruction by multi-exposure generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4063–4072, 2023.

Minhaeng Lee and Yu-Wing Tai. Robust all-in-focus super-resolution for focal stack photography. *IEEE Transactions on Image Processing*, 25(4):1887–1897, 2016.

Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *European Conference on Computer Vision*, pages 613–628. Springer, 2018.

Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013.

Shuaizheng Liu, Xindong Zhang, Lingchen Sun, Zhetong Liang, Hui Zeng, and Lei Zhang. Joint hdr denoising and fusion: A real-world mobile hdr image dataset. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13966–13975. IEEE, 2023.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1651–1660, 2020.

Rafał Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. *High Dynamic Range Imaging*, pages 1–42. John Wiley & Sons, Ltd, 2015. ISBN 9780471346081. doi: https://doi.org/10.1002/047134608X.W8265. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/047134608X.W8265`.

Rafal K Mantiuk, Dounia Hammou, and Param Hanji. 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.

Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer graphics forum*, volume 37, pages 37–49. Wiley Online Library, 2018.

John J McCann and Alessandro Rizzi. *The art and science of HDR imaging*. John Wiley & Sons, 2011.

Yuang Meng, Xin Jin, Lina Lei, Chun-Le Guo, and Chongyi Li. Ultraled: Learning to see everything in ultra-high dynamic range scenes. *arXiv preprint arXiv:2510.07741*, 2025.

Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 382–390. IEEE, 2007.

Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020.

Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 472–479. IEEE, 2000.

Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021.

Michael Potmesil and Indranil Chakravarty. A lens and aperture camera model for synthetic image generation. *ACM SIGGRAPH Computer Graphics*, 15(3):297–305, 1981.

Erik Reinhard. Parameter estimation for photographic tone reproduction. *Journal of graphics tools*, 7(1):45–51, 2002.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Zheng Shi, Ilya Chugunov, Mario Bijelic, Geoffroi Côté, Jiwoon Yeom, Qiang Fu, Hadi Amata, Wolfgang Heidrich, and Felix Heide. Split-aperture 2-in-1 computational cameras. *ACM Transactions on Graphics (TOG)*, 43(4):1–19, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Haley M So, Julien NP Martel, Gordon Wetzstein, and Piotr Dudek. Mantissacam: Learning snapshot high-dynamic-range imaging with perceptually-based in-pixel irradiance encoding. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2022.

Irwin Sobel, Gary Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968.

Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1386–1396, 2020.

SaiKiran Tedla, Zhoutong Zhang, Xuaner Zhang, and Shumian Xin. Learning to refocus with video diffusion models. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025a.

SaiKiran Tedla, Kelly Zhu, Trevor Canham, Felix Taubner, Michael S Brown, Kiriakos N Kutulakos, and David B Lindell. Generating the past, present and future from a motion-blurred image. *ACM Transactions on Graphics (TOG)*, 44(6):1–15, 2025b.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, and Xuaner Zhang. Lediff: Latent exposure diffusion for hdr generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 453–464, 2025.

Ronghuan Wu, Wanchao Su, Kede Ma, Jing Liao, and Rafał K Mantiuk. X2hdr: Hdr image generation in a perceptually uniform space. *arXiv preprint arXiv:2602.04814*, 2026.

Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018.

Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1751–1760, 2019.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Zhengming Yu, Li Ma, Mingming He, Leo Isikdogan, Yuancheng Xu, Dmitriy Smirnov, Pablo Salamanca, Dao Mi, Pablo Delgado, Ning Yu, et al. Diffhdr: Re-exposing ldr videos with video diffusion models. *arXiv preprint arXiv:2604.06161*, 2026.

Yunhao Zou, Chenggang Yan, and Ying Fu. Rawhdr: High dynamic range image reconstruction from a single raw image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12334–12344, 2023.

Appendix

We organize the appendix as follows. In Appendix A, we describe the preprocessing used for the HDR training and evaluation datasets. Appendix B details our evaluation protocol, with emphasis on scale-invariant Q^* metrics and the preprocessing applied before HDR-VDP3 evaluation. Appendix C provides additional quantitative results for the fusion ablation, comparing our Fusion UNet against classical LDR bracket merging methods and other Fusion UNet designs. In Appendix D, we study the effect of classifier-free guidance and LoRA fine-tuning on the performance of our method. In Appendix E, we visualize the generated LDR brackets against the GT LDR brackets and also present a quantitative comparison. Appendix F analyzes the role of the Fusion UNet in the performance of our method, including the drop observed when applying it to LDR brackets produced after inference-time optimization of the video model. We define the bracket-consistency metric in Appendix G. Appendix H describes how our framework can be adapted to all-in-focus image estimation from a single defocus-blurred input. Appendix I reports additional details on the user-study setup. Appendix J presents a detailed failure-case analysis, showing when our method performs worse and how this relates to input edge density and high-frequency detail loss from the VAE, along with a per-exposure diagnostic (fig. 12) that localizes prediction errors to specific bracket frames. In Appendix K, we compare our method with X2HDR-Flux, which uses a substantially larger denoising backbone than our model. Finally, Appendix L outlines the Q-JOD metric computation, explaining potential disagreements between Q-JOD and user-study preference rates for some baselines in the SI-HDR test set evaluation.

A Training and Evaluation Dataset details

We collate $10{,}965$ training crops ($5{,}820$ from Fairchild HDRPS Fairchild (2007), $5{,}145$ from RawHDR Zou et al. (2023)) using a recipe similar to X2HDR Wu et al. (2026). Source images are first downsampled to a width of 2560 px while preserving the aspect ratio. From each downsampled image, we extract 15 non-overlapping 512×512 crops, where the centers of the crops are selected via greedy farthest-first traversal of the regular crop grid: starting at the top-left position, each subsequently chosen origin maximizes the minimum squared pixel-distance to all previously-chosen origins, yielding spatially diverse, non-redundant samples that cover the full image. Raw and HDR are cropped while preserving the pixel-level alignment between them.

Fairchild HDRPS Fairchild (2007): We use all 44 scenes (388 bracketed NEF exposures total). The LDR exposure bracket for each scene is fused into a single ground-truth HDR using homography alignment and a minimum-spanning-tree exposure estimator Hanji and Mantiuk (2023) for a globally consistent radiometric scale. Every NEF in the bracket is demosaiced to linear sRGB and paired with the fused HDR image, giving one `(raw, gt_hdr)` sample per exposure ($388 \times 15 = 5\,820$ crops). This is a public domain dataset listed for non-commercial research.

RawHDR Zou et al. (2023): We process 343 out of 345 packed-Bayer captures across the train and test splits (two failed to load), each providing a short-exposure raw input and a paired HDR raw. Both arrays are demosaiced and converted to linear sRGB using the per-capture white-balance gains and 3×3 camera-to-sRGB matrix exported with the dataset, yielding $343 \times 15 = 5\,145$ pairs. This dataset is MIT licensed, which allows using it for research.

Test set: Evaluation is performed on the SI-HDR Hanji et al. (2022) `raw2hdr` test set on 96 scenes, processed and provided by Wu et al. (2026) to ensure consistent evaluation between both methods. This dataset is released under a CC BY 4.0 license, and thus can be used for research purposes.

Fusion UNet training data: Unlike the video diffusion model, which is conditioned on raw LDR inputs, the Fusion UNet only requires (clean) ground-truth HDR images at training time, since the input LDR bracket to the Fusion UNet is synthesized from the GT HDR image via eq. (6). This allows us to train the Fusion UNet using both the ground-truth HDR images from the dataset described above and a subset of the public HDR datasets used to train LEDiff Wang et al. (2025), restricted to sources that overlap with our collation. The combined corpus contains approximately $21{,}000$ HDR images.

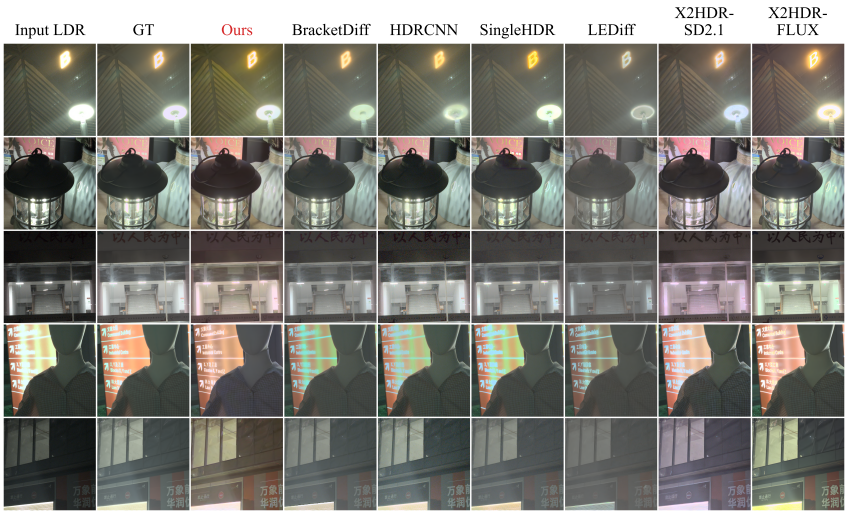

Figure 8: **Qualitative HDR reconstruction results on the MobileHDR test set.** We show the five MobileHDR test scenes with the largest margin between our method and the mean of the baselines (ranked by Q-JOD). All results are normalized to the same median luminance and visualized with Reinhard tone mapping Reinhard (2002).

B Evaluation Metrics Details

Our method, similar to existing works Wang et al. (2025); Wu et al. (2026); Bemana et al. (2025), predicts the radiance up to a global scale. We use the Q^* metrics proposed in Cao et al. (2024), and also used in Wu et al. (2026), to ensure invariance to the global scales in the quantitative evaluation. Computing Q^* metrics begins by optimizing the exposure values (using gradient descent) to match the GT and predicted HDR images based on a metric of choice, such as PSNR, SSIM, etc. Additionally, the evaluation averages spatially weighted metrics across a range of exposures, restricting each metric to image regions that are meaningful at the corresponding exposure. Applying the exposure mask is straightforward for pixel and patch-wise metrics such as SSIM, PSNR, and MAE. For LPIPS, we instead compute patch-wise perceptual distances first and apply the mask afterward, avoiding artifacts introduced by evaluating LPIPS on masked images. We omit DISTS Ding et al. (2020) as it is a global image-level metric, making a masked patch-wise evaluation nontrivial. Before computing the Q-JOD metric according to HDR-VDP3 Mantiuk et al. (2023), we normalize the ground truth and baseline images to the same median luminance value (8.0), as Q-JOD uses the full global scale for evaluation. Prior work Wu et al. (2026) also uses a similar normalization for displaying images in its user study.

MobileHDR evaluation. The MobileHDR dataset Liu et al. (2023) provides raw images without color correction matrices and white balance parameters. For fairness to the sRGB baselines (Bracket-Diff, HDRCNN, SingleHDR, LEDiff), whose LDR inputs we render with a gray-world white balance, we apply the same per-scene gray-world white balance to the GT and to the raw-input methods (Ours, X2HDR-SD2.1, X2HDR-Flux) prior to computing metrics for MobileHDR. The qualitative results in fig. 8 use the same white balance.

C Fusion Ablation Study Metrics

We report the additional metrics in table 5 to compare our Fusion UNet with the classical LDR bracket fusion methods outlined in section 5.1. We apply the Mertens algorithm on linearized input brackets and treat the weighted reconstruction as a linear-domain estimate. This ensures that the output is evaluated in the same linear HDR domain as our Fusion UNet prediction and the ground-truth HDR image. Debevec merging requires the relative exposure differences between the input images. Since our video model predicts only the bracket images and not their corresponding exposure values, we assume a fixed synthetic exposure schedule by uniformly sampling EV offsets in the range $[-5.5, 5.5]$

Fusion Method	Q*-PSNR ↑	Q*-SSIM ↑	Q*-MAE ↓	Q*-LPIPS ↓
Mertens Mertens et al. (2007)	18.87	0.513	0.096	0.352
Debevec Debevec and Malik (1997)	15.23	0.503	0.141	0.433
Ours (Fusion UNet)	**22.36**	**0.664**	**0.060**	**0.289**

Table 5: **Ablation study with different fusion methods.** Our learned Fusion UNet, trained on the same distribution of generated frames, substantially outperforms both the classical alternatives across all the metrics.

Fusion Strategy	Q-JOD	Q*-PSNR	Q*-SSIM	Q*-MAE ↓	Q*-LPIPS ↓
Softmax fusion (ours)	**7.520**	**22.39**	**0.659**	**0.060**	**0.292**
Abs-norm fusion	7.302	22.18	0.628	0.061	0.327
Direct-HDR	7.362	19.42	0.610	0.103	0.405

Table 6: Ablation of the Fusion UNet design. We compare our softmax-based weighted fusion against absolute-value normalization (Abs-norm) and direct HDR regression (Direct-HDR), while keeping the Fusion UNet architecture and training setup fixed.

across the generated bracket. This comparison, therefore, evaluates whether classical fusion methods can be used as drop-in replacements for our learned Fusion UNet under the same generated-bracket setting. The weaker performance of Debevec merging reflects its sensitivity to accurate exposure metadata, whereas our Fusion UNet does not require exposure estimates at inference time.

Fusion UNet design ablations We ablate the Fusion UNet design to determine whether constraining the final HDR reconstruction to a weighted combination of the generated exposure stack limits reconstruction quality. Our default formulation predicts per-pixel fusion logits that are normalized with a softmax across the generated exposure frames, yielding non-negative weights that sum to one. We compare this design against two alternatives: **(i) Abs-norm**, which replaces the softmax with explicit absolute-value normalization of the predicted weights, and **(ii) Direct-HDR:** the same lightweight UNet directly regresses the HDR image from the generated bracket without any fusion constraint. We keep the Fusion UNet architecture, training data, optimization settings, and training schedule the same for both design choices. As shown in table 6, softmax-based fusion consistently outperforms both alternatives across all reconstruction metrics. In particular, Direct-HDR performs substantially worse despite having greater output flexibility, suggesting that restricting the Fusion UNet to selecting and combining information already present in the generated exposure stack provides a useful inductive bias. Rather than requiring the 0.5M-parameter network to jointly learn exposure selection and unrestricted HDR synthesis, the proposed formulation reduces its task to estimating the relative reliability of the complementary exposure candidates at each pixel, while also retaining interpretable per-exposure contribution maps.

D Additional results from LoRA fine-tuning and no classifier-free guidance

In table 7 we report the metrics from a sweep of the CFG weight, showing that increasing the CFG scale slightly hurts performance. For this ablation, we used a model trained for $15{,}000$ iterations with conditional dropout enabled during training, following Chen et al. (2025a). We also observe slight performance drops on using LoRA fine-tuning (rank 128), where the LoRA layers were applied to the Q, K, V projections and the attention output blocks of the transformer layers of the Video Diffusion UNet.

E Generated exposure brackets from the video diffusion model

We visualize the LDR brackets synthesized by our method in fig. 9, alongside GT LDR brackets generated from the HDR images using eq. (6) and eq. (7). Recall that the video model is fine-tuned on the training data to reproduce the exposure brackets generated by eq. (6) and eq. (7). The synthesized brackets largely match the GT exposure progression, although in some cases the generated frames deviate in exposure, especially near the low-exposure end of the bracket.

Variant	Q*-PSNR ↑	Q*-SSIM ↑	Q*-MAE ↓	Q*-LPIPS ↓	Q-JOD ↑
CFG sweep					
CFG = 1.0 (no guidance)	22.33	0.654	0.060	0.293	7.14
CFG = 1.5	22.26	0.655	0.061	0.294	7.14
CFG = 2.5	22.13	0.656	0.062	0.295	7.13
CFG = 4.0	21.88	0.655	0.064	0.300	7.12
LoRA vs full fine-tune (no CFG)					
LoRA	22.00	0.656	0.063	0.296	7.10
Full FT	**22.36**	**0.664**	**0.060**	**0.289**	**7.51**

Table 7: **Ablations on classifier-free guidance scale and LoRA vs full fine-tuning.** Increasing CFG monotonically degrades most fidelity metrics (PSNR, MAE, LPIPS, JOD) for a model trained with condition dropout. LoRA fine-tuning (with no condition dropout) marginally underperforms full fine-tuning, suggesting that full-rank capacity benefits learning the HDR-bracket distribution.

	Original prediction			After exposure compensation			
Bracket idx	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	ΔEV
0 (darkest)	20.76	0.545	0.270	32.53	0.838	0.116	−2.45
1	18.31	0.591	0.263	27.38	0.810	0.151	−1.72
2	17.25	0.652	0.262	23.16	0.768	0.204	−0.97
3	17.60	0.681	0.260	20.74	0.713	0.251	−0.41
4 (brightest)	21.45	0.693	0.266	21.67	0.694	0.267	−0.03
All	**19.07**	**0.632**	**0.264**	**25.10**	**0.765**	**0.198**	−1.12

Table 8: Per-bracket-index reconstruction metrics between predicted and GT LDR brackets on the SI-HDR test set, before and after per-frame global-scale exposure compensation using eq. (11). $\Delta\text{EV} = \log_2 s^\star$ denotes the optimal exposure offset applied to the prediction. Exposure compensation improves PSNR by $+6.0$ dB on average and by $+11.8$ dB at the darkest index, suggesting that much of the error comes from global EV miscalibration rather than structural mismatch.

We show in table 8 that these errors are largely photometric rather than structural. Specifically, much of the discrepancy can be explained by a global per-frame exposure offset in linear light. Let $I_p, I_g \in [0,1]^{H\times W\times 3}$ denote a predicted LDR frame and its GT counterpart, both gamma-encoded with $\gamma = 2.2$. We first linearize them as

$$\tilde{I}_p = I_p^\gamma, \qquad \tilde{I}_g = I_g^\gamma.$$

We then solve for the optimal global linear-light scale

$$s^\star = \arg\min_{s>0} \left\| s\tilde{I}_p - \tilde{I}_g \right\|_2^2 = \frac{\langle \tilde{I}_p, \tilde{I}_g \rangle}{\langle \tilde{I}_p, \tilde{I}_p \rangle}, \tag{11}$$

where inner products are computed over all pixels and color channels. The exposure-aligned prediction is

$$\hat{I}_p = \text{clamp}\Big(s^\star \tilde{I}_p, 0, 1\Big)^{1/\gamma}.$$

We then recompute PSNR, SSIM, and LPIPS between $\hat{I}_p$ and I_g. After this simple exposure alignment, the metrics improve substantially, especially for the darkest bracket frames. This indicates that the generated brackets preserve much of the underlying structure, and that the dominant mismatch is often a global EV miscalibration. Importantly, such global scale offsets do not directly affect downstream HDR reconstruction, since HDR predictions are evaluated up to an unknown global scale.

F Performance ceiling from Fusion UNet

The performance of our pipeline depends on both the video diffusion model and the Fusion UNet. To isolate the effect of the Fusion UNet, we feed it ground-truth LDR brackets at test time. As shown in table 9, this oracle setting substantially improves performance (Q^*-PSNR 44.84 vs. 22.36),

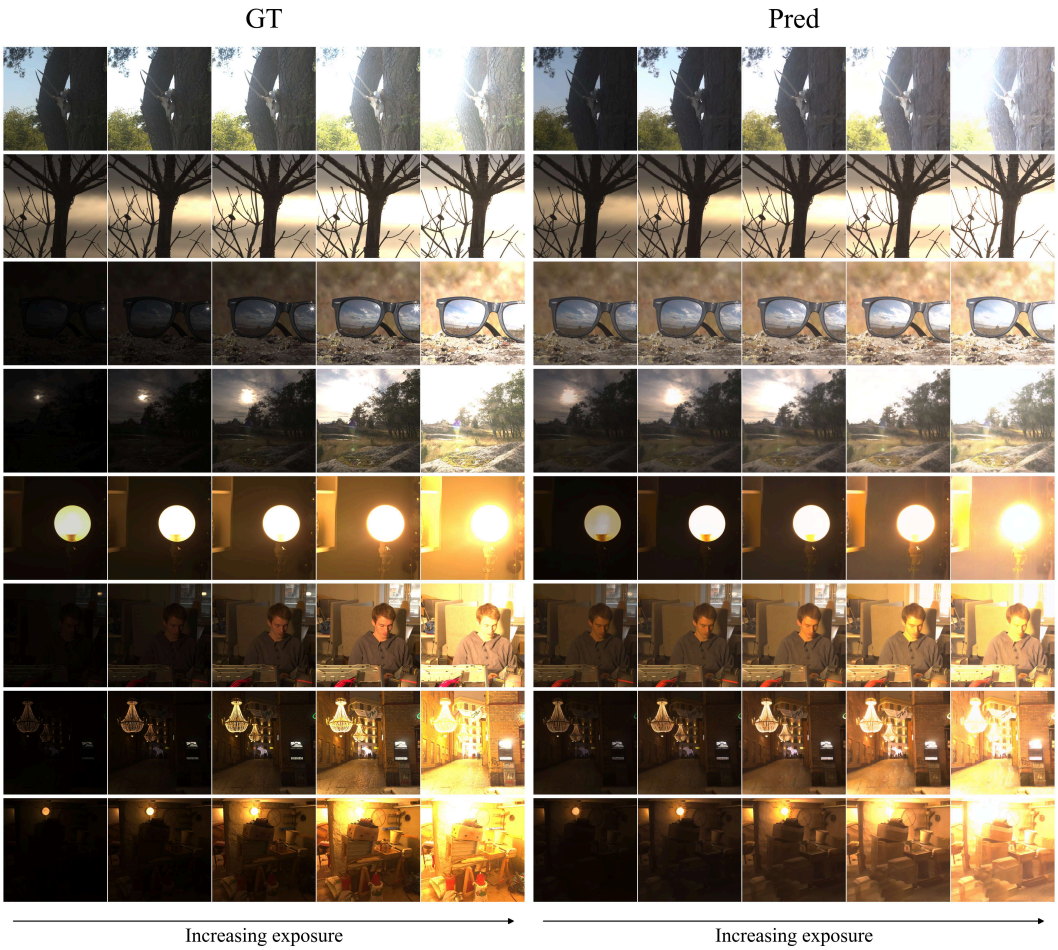

Figure 9: **Exposure brackets generated by our fine-tuned model.** Each row corresponds to one scene from the SI-HDR test set. We show the ground-truth LDR exposure bracket under the header *GT* and our predictions under *Pred*. Exposure increases from left to right under each header. Our predictions match the relative exposure progression of the GT brackets with high fidelity in most cases. In rows 3, 4, and 6, we observe an exposure mismatch at the low-EV (darkest) end of the bracket between our method and GT. However, this behavior is not consistent across the visualization, indicating output variance of the model rather than a systematic bias.

Method	**Q*-PSNR ↑**	**Q*-SSIM ↑**	**Q*-MAE ↓**	**Q*-LPIPS ↓**
Ours (Pred LDR brackets + Fusion UNet)	22.36	0.664	0.060	0.289
Ours (Pred LDR brackets + exp comp + Fusion UNet)	22.28	0.657	0.061	0.294
Fusion UNet + GT LDRs (Oracle)	**44.84**	**0.989**	**0.007**	**0.013**

Table 9: **Fusion UNet oracle analysis.** Ground-truth LDR brackets yield much higher scores (row 3) than our method (row 1), showing that the Fusion UNet is not the primary bottleneck. Exposure-compensating predicted brackets (row 2) has a negligible effect, indicating that global EV drift is not the limiting factor; performance is mainly limited by artifacts and inconsistencies in the generated LDR brackets.

HDR Merge	Q^*-PSNR ↑	Q^*-SSIM ↑	Q^*-MAE ↓	Q^*-LPIPS ↓
Debevec	**21.36**	**0.623**	**0.064**	**0.337**
Fusion UNet	16.20	0.570	0.127	0.431
Fusion UNet trained on $\Delta EV = 1$	20.98	0.622	0.068	0.333

Table 10: **Effect of distribution mismatch on Fusion UNet.** After inference-time optimization, Debevec merging outperforms the Fusion UNet. The optimized LDR brackets induce a distribution shift relative to the brackets used during Fusion UNet training and negatively affect its performance. However, training the Fusion UNet on exposure stacks simulated with $\Delta EV = 1$ closes the gap with Debevec fusion, showing that bracket consistency and Fusion UNet can be unified.

indicating that the Fusion UNet is not the primary bottleneck. We also test whether EV drift towards lower exposures in our predicted brackets affects downstream reconstruction by applying the scale alignment from eq. (11) before fusion. The negligible change in row 2 of table 9 shows that exposure drift is not the limiting factor. Instead, performance is mainly limited by artifacts and inconsistencies in the generated LDR brackets, suggesting that improvements to the video model should directly improve HDR reconstruction. However, the Fusion UNet underperforms on brackets produced after inference-time optimization (section 5.1), where the video model is constrained to generate fixed $\Delta EV{=}1$ exposure spacing. In this setting, Debevec merging performs better. We attribute this drop to distribution shift: the Fusion UNet is trained on GT brackets from eq. (6) and eq. (7), whose ΔEV spacing is input-dependent, but is tested on brackets modified to have constant spacing. This performance drop is hence expected as the Fusion UNet is intentionally lightweight and not designed to handle arbitrary bracket distributions. Its role is to show that generated LDR brackets can support HDR reconstruction through simple learned per-pixel fusion. Quantitative results are provided in table 10.

G Bracket Consistency Experiment Details

To evaluate whether the generated LDR frames form a physically plausible exposure bracket, we measure bracket consistency on 96 scenes from the SI-HDR test set. A bracket-consistent sequence should exhibit monotonically increasing linear brightness with adjacent frames separated by one exposure stop ($\Delta EV = 1$), i.e., a $2\times$ increase in linear intensity. For each generated sequence, we first linearize the frames using the inverse sRGB gamma approximation, $x^{2.2}$, and compute the luminance image. To avoid confounding the measurement with saturation or clipping, we compute a shared valid mask for each generated bracket using the linearized luminance values. Specifically, we retain only pixels p that remain unsaturated in all five generated frames:

$$0.04 < Y_i(p) < 0.96, \qquad \forall i \in \{1, \dots, 5\},$$

where $Y_i(p)$ denotes the linearized luminance at pixel p in frame i. Bracket consistency is then measured only over this shared valid region.

Here, we outline the quantitative computation of bracket consistency as reported in section 5.1. We evaluate bracket consistency locally over non-overlapping 64×64 patches. A patch is retained only if at least 50% of its pixels are valid under the shared mask. For scene s, valid patch p, and exposure index k, let $\ell_{s,p,k}$ denote the mean linear luminance of the k-th generated frame over the valid pixels in that patch. We compute the brightness ratio between adjacent exposure steps as

$$r_{s,p,k} = \frac{\ell_{s,p,k+1}}{\ell_{s,p,k} + \epsilon},$$

where k indexes the four adjacent transitions in the five-frame bracket. We define the ratio deviation as

$$\delta = \frac{1}{|\mathcal{R}|} \sum_{(s,p,k)\in\mathcal{R}} |r_{s,p,k} - 2.0|\,,$$

where $\mathcal{R}$ is the set of all valid scene–patch–transition tuples. Thus, lower values indicate more consistent exposure spacing, with $\delta = 0$ corresponding to perfect $2\times$ spacing at every measured transition. The bracket-consistency numbers reported in section 5.1 are computed using the δ metric defined here.

H Extending our approach to all-in-focus image estimation

We provide details for adapting our approach to all-in-focus (AIF) image estimation, as discussed in section 5.2.

Focal stacks and all-in-focus estimation: Many computer vision pipelines assume a pinhole camera model, where the captured image is sharp everywhere. Such an image is commonly referred to as an all-in-focus (AIF) image. In contrast, real cameras have finite apertures, which introduce depth-dependent defocus blur. This blur can be modeled using geometric optics Potmesil and Chakravarty (1981). A common approximation is to synthesize a defocused image I_b from an AIF image I using a spatially varying blur kernel:

$$I_b(x, y) = \int\int I(u, v)\, h(x, y, u, v \mid D(u, v), F)\, du\, dv, \tag{12}$$

where (x, y) denotes the output pixel location, (u, v) denotes coordinates over the image domain, $D(u, v)$ is the scene depth, and F is the focus distance. For a fixed camera aperture, the blur kernel depends on the mismatch between the scene depth and the focus distance. The blur magnitude increases with $|D(u, v) - F|$, according to the circle-of-confusion equation (Eq.1 of Gur and Wolf (2019)). Please refer to Gur and Wolf (2019) for a detailed overview of the defocus blur forward model.

Recovering AIF images from a focal stack: Different parts of a scene are in focus (blur-free) at different focus distances. Let $\{I_b^1, \ldots, I_b^N\}$ denote a focal stack captured with monotonically increasing focus distances $\{F^1, \ldots, F^N\}$. Conventional AIF recovery methods Li et al. (2013); Lee and Tai (2016) estimate per-pixel weight maps $\{W^1, \ldots, W^N\}$ over the stack and recover the AIF image as

$$I_{\text{AIF}} = \sum_i W^i \odot I_b^i. \tag{13}$$

However, these methods require capturing a multi-image focal stack as input. Under our proposed paradigm, we instead fine-tune the video model to generate a focal stack with monotonically increasing focus distances, conditioned on a single input image. A light-weight fusion network then predicts per-pixel fusion weights and combines the generated stack into an AIF estimate.

Experiment details: We evaluate this adaptation on the NYUv2 dataset Silberman et al. (2012). We use 1250 RGB-D images for training. For each RGB image and its corresponding depth map, we simulate a focal stack of five defocused images using the differentiable defocus blur approximation from prior work Gur and Wolf (2019). To choose the focus distances for each scene, we run K-means clustering on the ground-truth depth map with $K = 5$, selecting focus distances that cover the dominant scene depths.

During fine-tuning, a conditioning input image is randomly sampled from the simulated focal stack. The video diffusion model is fine-tuned conditioned on the input to generate the full focal stack, ordered monotonically by focus distance. In a separate training stage, we then train the Fusion UNet on generated GT focal stacks to predict per-pixel fusion weights, using an ℓ_1 reconstruction loss against the original sharp RGB images, which we treat as the ground-truth AIF targets. Thus, the overall pipeline is identical to our HDR reconstruction method in section 4, but applied to a different image formation model.

Averaged over 50 randomly selected scenes from the NYUv2 test set, our method achieves a PSNR of 31.26, SSIM of 0.90, and LPIPS of 0.096 against the original sharp RGB images. These results suggest that our formulation is not specific to exposure bracketing and can extend to other inverse problems where the desired image can be recovered by synthesizing and fusing a physically meaningful image stack.

I Details on User Study Setup

Apparatus: All sessions were run on a 14" MacBook Pro with the built-in Liquid Retina XDR display (peak 1600 nits, EDR-capable), under standard office lighting that was kept consistent across

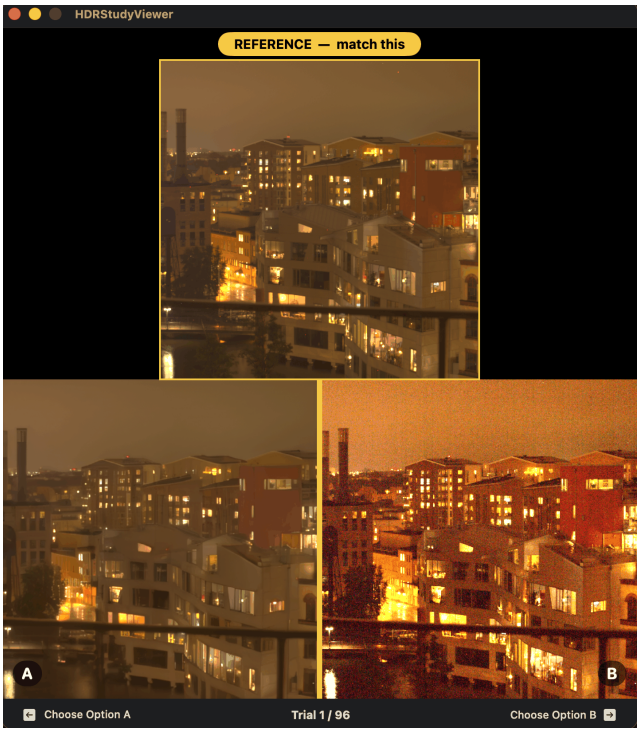

Figure 10: **User-study trial interface.** The ground-truth HDR image is displayed on the top half (centered) and the two candidates (ours vs. a randomly chosen baseline) are displayed side-by-side on the bottom half. The participant uses the left/right arrow keys to pick the candidate that best matches the ground truth.

participants. The viewer was a custom Swift/Metal application that operates in a linear color space and writes into an extended-dynamic-range Float16 layer, so values above SDR reference white are preserved through to the HDR headroom of the display, where HDR headroom is defined as the ratio of the peak luminance to the SDR reference white value.

Stimuli and pre-display pipeline: Each trial draws on a single scene from the test split, where the same scene was reconstructed by every method (ours and the five baselines: LEDiff, HDRCNN, SingleHDR, BracketDiff, and X2HDR-SD2.1) plus a ground-truth HDR image. To prevent the choice of normalization from biasing preferences, every image is sent through a fixed pipeline before display: (i) we use the linear HDR images directly; (ii) we normalize each image by dividing by its own 20.05^{th}-percentile luminance and multiplying by an exposure constant of 0.0037, anchoring the dark end of the histogram to a common level; (iii) the result is scaled by a fixed EDR gain of $12\times$ to push the working range into the HDR headroom of the display. The ground-truth image is rendered through the same pipeline, with no preferential treatment.

Task and protocol: We use a three-image layout with the ground truth on the top half (centered) and the two candidates (ours vs. a randomly chosen baseline) side-by-side on the bottom half; left/right placement is randomized per trial. The participant chooses the bottom image that best matches the ground truth using the left/right arrow keys; fig. 10 shows the on-screen arrangement. There is no time limit. Each participant completes 96 trials (for SI-HDR) drawn such that each baseline appears with approximately equal frequency, with scene order randomized per participant. The user study was uncompensated.

Participants: We collected data from $N{=}20$ participants (14 male, 4 female, 2 non-binary), ages 18–58 (median 22.5, mean 26.2); 10 wore glasses or contacts. Sessions lasted approximately 10–30 minutes (median $\approx$ 22 min). Demographics were recorded with consent. For the second user study on MobileHDR, we recruited $N{=}20$ participants spanning a broader age distribution: 3 participants

Participant	LEDiff	HDRCNN	SingleHDR	BracketDiff	X2HDR-SD2.1	Overall
P01	100%	77%	81%	86%	65%	81%
P02	95%	72%	94%	81%	60%	80%
P03	95%	75%	80%	76%	71%	80%
P04	79%	92%	67%	89%	61%	78%
P05	86%	74%	87%	65%	67%	77%
P06	86%	85%	60%	68%	80%	77%
P07	100%	73%	71%	54%	67%	74%
P08	82%	76%	86%	74%	60%	74%
P09	78%	94%	78%	59%	57%	73%
P10	89%	59%	69%	74%	75%	73%
P11	86%	71%	79%	82%	40%	73%
P12	89%	63%	82%	57%	67%	71%
P13	78%	74%	79%	62%	63%	71%
P14	83%	61%	77%	84%	50%	71%
P15	87%	54%	62%	65%	63%	68%
P16	100%	65%	46%	79%	62%	67%
P17	100%	67%	80%	54%	46%	66%
P18	59%	85%	65%	69%	36%	66%
P19	83%	70%	60%	37%	68%	62%
P20	74%	75%	45%	52%	43%	58%
# >50%	**20/20**	**20/20**	**18/20**	**19/20**	**15/20**	**20/20**

Table 11: **Per-participant aggregation of the user study for SI-HDR dataset.** Each cell is the fraction of trials (out of approximately 20 per cell, varying with random scene assignment) in which that participant preferred our reconstruction over the indicated baseline; the *Overall* column is the same fraction pooled across all five baselines (96 trials per participant). The bottom row counts how many of the 20 participants exceeded 50% in each column.

(15%) were aged 16–19, 9 (45%) were aged 20–29, 5 (25%) were aged 30–39, 1 (5%) was aged 40–49, and 2 (10%) were aged 50–59.

Aggregated results: table 3 reports the per-baseline ours-preferred rate for the SI-HDR dataset; the overall rate is 72.0% ($1382/1920$ judgements). We observe a minor difference in preference rates when our method was displayed in the left pane (70.7%) vs. the right pane (73.3%), indicating that the study was not affected by any image-placement bias. Over all pairwise comparisons, our method was displayed in the left and right panes with almost uniform probability (49.0% vs. 51.0%). Per-participant rates ranged from 58.3% to 81.2%, with every participant preferring our reconstruction in aggregate. We show a similar

Statistical significance: Following Wang et al. (2025), we test each baseline with a two-sided binomial test against chance ($H_0 : p{=}0.5$, where p is the probability that a random participant prefers our reconstruction over the baseline). All five comparisons are significant well below the $\alpha{=}0.01$ level (table 3); the smallest margin is X2HDR-SD2.1 at $p = 2.7\times10^{-5}$, and the pooled overall test gives $p = 2.8\times10^{-85}$ on $1382/1920$ aggregate comparisons.

Per-participant analysis for SI-HDR: The trial-level binomial test in table 3 treats all 1920 judgements as independent, but the observations are clustered within 20 participants (96 trials each), which can inflate apparent statistical confidence. We therefore report the raw per-participant preference rate against each baseline in table 11 on the SI-HDR dataset. Every one of the 20 participants prefers our reconstruction in aggregate (range 58%–81%), and a majority of participants also prefers our reconstruction against each baseline ($20/20$ for LEDiff and HDRCNN, $19/20$ for BracketDiff, $18/20$ for SingleHDR, and $15/20$ for X2HDR-SD2.1). The unanimity in the *Overall* column yields a two-sided sign test of $p = 2(1/2)^{20} \approx 1.9\times10^{-6}$, which requires no distributional assumptions on the per-participant rates. In table 12, we report the per-participant preference table for the MobileHDR user study.

Qualitative worst cases (user study): fig. 11 shows the five test-set scenes (SI-HDR dataset) on which our reconstructions were preferred least often by participants (preference rates ranging from 25% to 40% out of 20 judgements per scene). These tend to be scenes with already-balanced exposure and salient mid-frequency texture (e.g. the potted plants in row 3), where dynamic-range

Participant	LEDiff	HDRCNN	SingleHDR	BracketDiff	X2HDR-SD2.1	X2HDR-Flux	Overall
P01	6/6 100%	5/5 100%	4/4 100%	7/7 100%	3/3 100%	3/3 100%	28/28 100%
P02	8/8 100%	3/4 75%	1/1 100%	6/6 100%	2/2 100%	7/7 100%	27/28 96%
P03	3/3 100%	6/8 75%	4/4 100%	5/5 100%	3/3 100%	5/5 100%	26/28 93%
P04	4/4 100%	5/6 83%	7/7 100%	1/1 100%	7/7 100%	3/3 100%	27/28 96%
P05	6/6 100%	4/5 80%	2/2 100%	8/8 100%	4/4 100%	3/3 100%	27/28 96%
P06	4/5 80%	3/4 75%	2/2 100%	7/7 100%	6/6 100%	4/4 100%	26/28 93%
P07	2/2 100%	6/7 86%	2/2 100%	8/9 89%	6/6 100%	2/2 100%	26/28 93%
P08	4/5 80%	2/2 100%	3/3 100%	7/8 88%	4/4 100%	6/6 100%	26/28 93%
P09	2/2 100%	3/4 75%	10/10 100%	4/4 100%	5/5 100%	3/3 100%	27/28 96%
P10	8/8 100%	2/2 100%	1/2 50%	6/6 100%	5/6 83%	3/4 75%	25/28 89%
P11	5/5 100%	3/4 75%	1/1 100%	8/8 100%	7/7 100%	2/3 67%	26/28 93%
P12	3/3 100%	3/4 75%	5/5 100%	3/3 100%	3/4 75%	9/9 100%	26/28 93%
P13	5/5 100%	5/6 83%	4/4 100%	5/6 83%	4/4 100%	3/3 100%	26/28 93%
P14	7/8 88%	3/3 100%	4/5 80%	2/2 100%	5/6 83%	4/4 100%	25/28 89%
P15	4/4 100%	5/6 83%	5/5 100%	2/2 100%	6/6 100%	3/5 60%	25/28 89%
P16	5/5 100%	4/4 100%	3/4 75%	2/2 100%	6/7 86%	6/6 100%	26/28 93%
P17	4/6 67%	5/6 83%	2/2 100%	4/4 100%	4/5 80%	5/5 100%	24/28 86%
P18	7/7 100%	3/3 100%	3/3 100%	3/4 75%	8/9 89%	2/2 100%	26/28 93%
P19	3/3 100%	5/5 100%	5/5 100%	3/4 75%	6/7 86%	1/4 25%	23/28 82%
P20	3/3 100%	3/6 50%	3/3 100%	2/2 100%	3/7 43%	6/7 86%	20/28 71%
# >50%	**20/20**	**19/20**	**19/20**	**20/20**	**19/20**	**19/20**	**20/20**

Table 12: **Per-participant aggregation of the user study for the MobileHDR dataset.** Each cell reports the number of trials in which a participant preferred our reconstruction over the indicated baseline, along with the corresponding percentage. The *Overall* column pools results across all six baselines, totaling 28 trials per participant. The bottom row counts how many of the 20 participants exceeded 50% preference in each column.

	LEDiff	HDRCNN	SingleHDR	BracketDiff	X2HDR-SD2.1	X2HDR-Flux	Overall
Win rate (Ours) ↑	94.9%	83.0%	95.9%	94.9%	89.8%	90.9%	**91.4%**
Binomial p ↓	4.5×10^{-22}	5.8×10^{-11}	7.2×10^{-18}	4.5×10^{-22}	2.4×10^{-18}	4.6×10^{-18}	$\mathbf{4.8\times10^{-99}}$

Table 13: **User-study results on MobileHDR test set.** 20 participants compared our reconstructions against each baseline using the GT HDR image as a reference. We report the percentage of trials preferring our method over baselines. Two-sided exact binomial tests against chance (p_0=0.5) show a significant preference rate for our method against the compared baselines.

expansion contributes little and any small detail loss from the latent VAE is more visible than on heavily ill-exposed inputs.

MobileHDR User Study We follow a similar user study protocol as the SI-HDR dataset user study and perform a user study on the MobileHDR test set Liu et al. (2023) as well. We report our results in table 13. Our method has the highest preference rate over all the baselines.

J Failure case analysis

We expand on the limitations noted in section 6 with a more thorough analysis. As noted in the main paper, our method can lag behind baseline methods on already well-exposed inputs, where little dynamic range expansion is required. This limitation is less severe in practice, since such inputs already retain most of the visible scene content and benefit less from HDR reconstruction. To better understand this behavior, we analyze whether the relative performance of our method is related to the amount of high-frequency details visible in the input. Intuitively, when fewer regions are saturated or under-exposed, more edges remain measurable in the input. Based on this insight, we use the *edge density*, defined below, as a simple proxy for visible high-frequency scene detail. We find that this quantity strongly predicts the relative performance of our method against the baselines.

Edge density: For an input LDR image I with luminance $L = 0.2126\,R + 0.7152\,G + 0.0722\,B$, we compute the 3×3 Sobel gradient magnitude $G_{ij} = \sqrt{(L * S_x)^2_{ij} + (L * S_y)^2_{ij}}$, where S_x and S_y

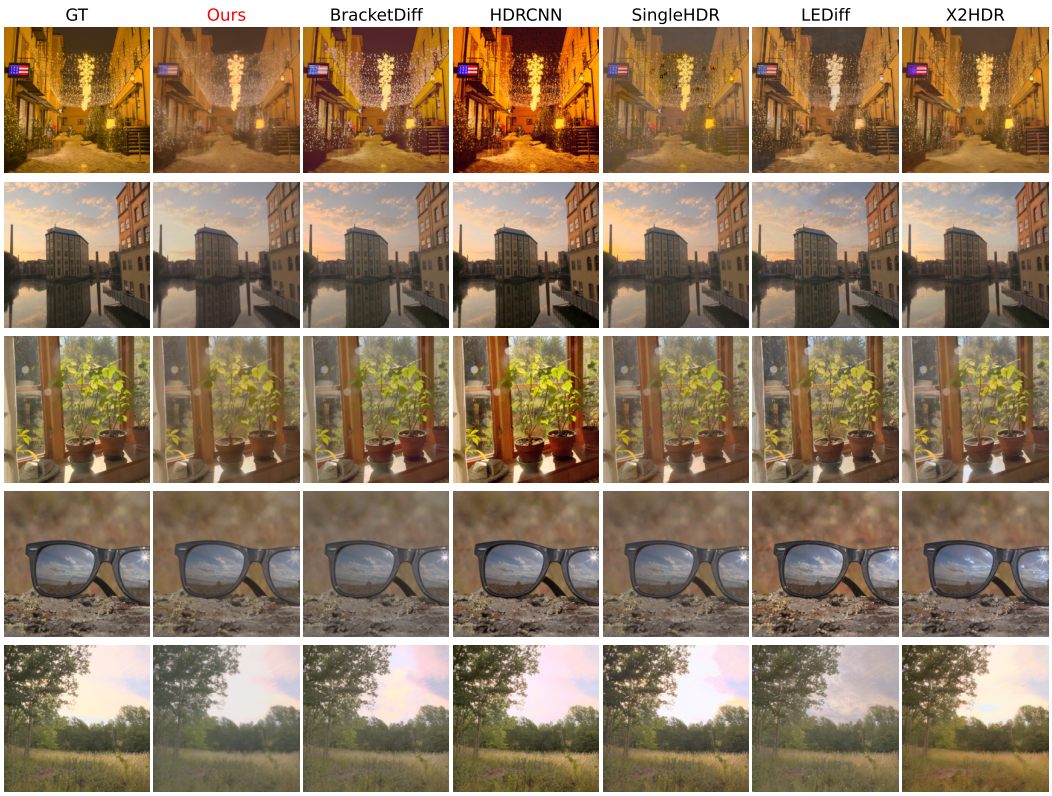

Figure 11: **User study failure cases.** Reconstructions for the five test-set scenes on which participants most often preferred a competing baseline over ours.

are the standard Sobel kernels Sobel et al. (1968). We define the edge density ($\kappa(\cdot)$) as the fraction of pixels with gradient magnitude above a threshold $\tau = 0.13$:

$$\kappa(I) = \frac{1}{HW} \left| \{(i,j) : G_{ij} > \tau\} \right|. \tag{14}$$

We use $\kappa(I)$ as a simple proxy for the amount of visible high-frequency structure in the input LDR image.

In fig. 13, we plot Δ SSIM against the input edge density $\kappa(I)$ across all 96 test scenes. For each scene, Δ SSIM is computed as the difference between the Q^*-SSIM of our method and the average Q^*-SSIM of the baseline methods. We observe from fig. 13 that edge density is a strong predictor of the relative performance gap between our method and the baselines. In particular, our method performs best relative to prior methods on low-edge-density inputs, which commonly correspond to challenging HDR scenes with saturated or under-exposed regions.

The strong correlation in fig. 13 motivates a closer look at what low edge density represents. On the SI-HDR test set, low input edge density can occur either because the scene is intrinsically smooth, with little high-frequency spatial content, or because the input LDR image is severely under- or over-exposed, causing its limited dynamic range to suppress gradients that would otherwise be visible. Both cases fall in the low-edge-density regime where our method performs strongest relative to the baselines.

The latter case is especially important for single-shot HDR reconstruction. In fig. 15, we show additional over-exposed input LDRs for scenes where our method outperforms the baselines. These examples indicate that over-exposure can reduce apparent edge density in the input image, even when the underlying HDR scene contains substantial structure. Compared to the failure cases in fig. 14, these scenes contain fewer high-frequency textures such as foliage, which helps explain why our method performs better in this regime.

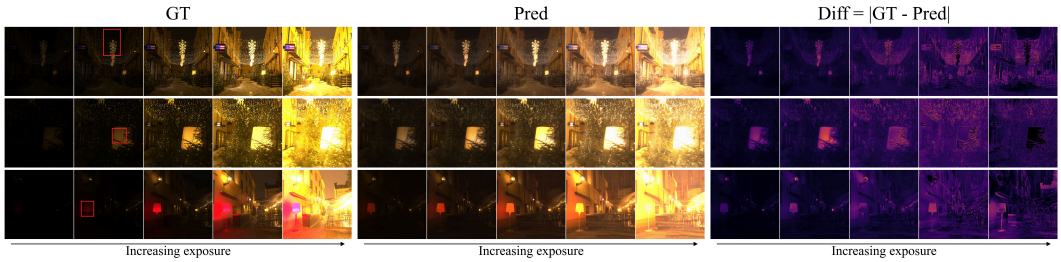

Figure 12: **Per-exposure failure diagnostics.** Since our method synthesizes the full exposure bracket, errors can be localized to specific exposures by comparing our generated bracket (Pred) against the ground-truth bracket (GT) frame-by-frame, using the per-frame $|\text{GT} - \text{Pred}|$ difference (Diff). A red box on the GT column marks the relevant region in each row. In rows 2–3, our bracket (Pred) shows a street lamp already lit at this exposure, where GT is still dark. In row 1, the hanging string lights instead show higher error at the high-exposure end of the bracket.

Scene set	n	$\rho_{\text{VAE}} \uparrow$	$\rho_{\text{full}} \uparrow$
Best-10 (smooth / dark scenes)	10	0.89	0.72
Worst-10 (textured outdoor scenes)	10	0.69	0.45

Table 14: **VAE-induced high-frequency energy loss.** ρ_{VAE} is the ratio between the high-frequency energy (Laplacian std of luminance) of a GT image that has been PU-21-encoded, encoded–decoded through our SVD VAE, and compared back in PU-21 space, and that of the original PU-21-encoded GT. ρ_{full} is the same ratio computed on our predicted HDR (also PU-21 encoded). Encoding inputs in PU-21 places the VAE in its trained distribution range and isolates its compression bottleneck from input-distribution artifacts. We report the mean over the 10 scenes where our method most outperforms baselines (Best-10) and the 10 where it most underperforms (Worst-10). The VAE alone discards 31% of input HF energy on textured scenes versus only 11% on smooth scenes, accounting for $\approx 75\%$ of the cross-scene gap in our pipeline's HF preservation.

The edge-density analysis also highlights the failure regime, i.e., appropriately exposed scenes with high-edge-density. Since these scenes require preserving fine spatial detail rather than recovering missing dynamic range, it is natural to investigate the effect of the SVD VAE, which compresses images into a lower-dimensional latent representation. This motivates our VAE round-trip analysis in section J.1, where we show that the latent bottleneck can attenuate high-frequency texture.

Qualitative failure cases: Fig. 14 shows the five scenes on which our method trails existing methods by the largest PSNR margin. Most of them are dense outdoor or textured-foliage scenes. While our reconstructions visibly lose a small amount of fine high-frequency detail relative to the strongest baselines, they remain visually comparable overall. This is consistent with our user study (table 3): even across scenes of this type, our method is preferred over the baselines 72% of the time on average. The scenes where our method is least preferred in the user study (fig. 11) show qualitative similarities to the well-exposed cases where it underperforms the baselines in PSNR (fig. 14). Both sets of failure cases are well-exposed scenes with dense high-frequency detail, such as foliage and tiled patterns.

Per-exposure failure diagnostics: Because our method synthesizes the full exposure bracket rather than directly regressing the HDR image, we can localize errors to specific exposures by comparing our generated bracket against the ground-truth bracket frame-by-frame, rather than only inspecting the final fused HDR output. fig. 12 shows this diagnostic on three representative failure scenes. The street lamps in rows 2 and 3 have incorrect relative exposure with respect to the rest of the scene at lower exposures: our bracket shows each lamp already lit at exposures where the GT bracket is still dark. The hanging string lights in row 1 instead show higher error at the high-exposure end of the bracket.

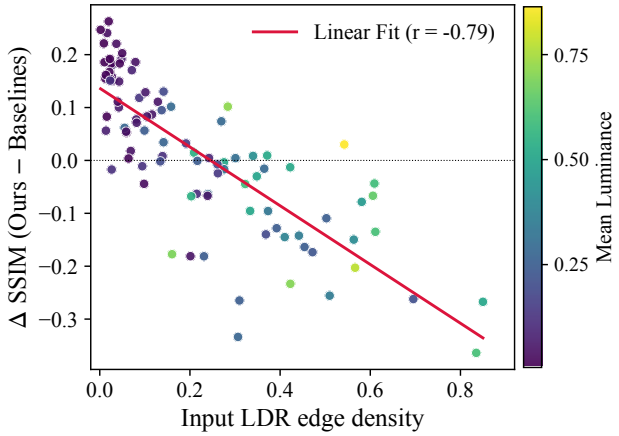

Figure 13: **Per-scene Δ SSIM vs. input edge density ($n = 96$).** Each point is one test scene. $\Delta\,\mathrm{SSIM} = \mathrm{SSIM}_{\mathrm{Ours}} - \overline{\mathrm{SSIM}}_{\mathrm{baselines}}$ (positive $\Rightarrow$ Ours better). The Pearson correlation coefficient between edge density and Δ SSIM is $r = -0.79$: We outperform the baselines on inputs with low edge-density. After controlling for the mean LDR luminance (normalized to $[0, 1]$), edge density remains strongly correlated with Δ SSIM (partial $r = -0.67$), while luminance contributes no residual signal once edge density is held fixed (partial $r = +0.04$). This suggests that the observed trend is driven primarily by the amount of visible structure in the input, rather than by its mean brightness.

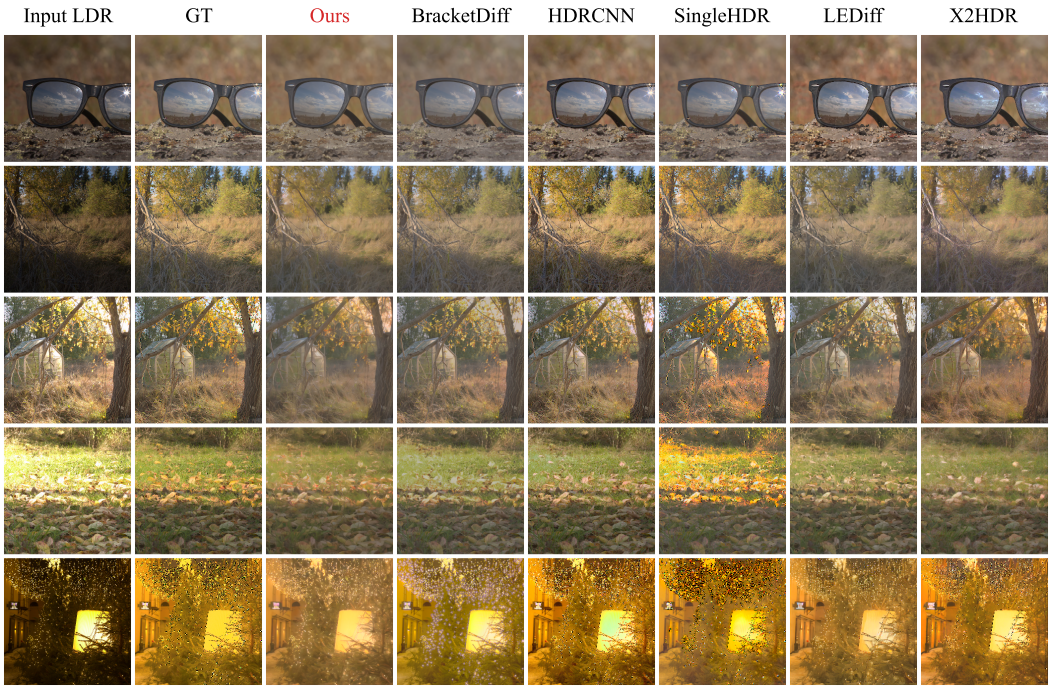

Figure 14: **Failure-case qualitative comparison.** We show five well-exposed scenes containing high-frequency textures. HDR images are visualized with Reinhard tone-mapping. In these cases, our method loses some fine texture detail compared to existing methods, consistent with the VAE bottleneck analysis in section J.1. Nevertheless, the reconstructions remain visually comparable overall.



Figure 15: **Bright-input scenes where our method outperforms most baselines.** Input LDR thumbnails (top row) and tone-mapped ground-truth HDR images (bottom row) for the 10 brightly exposed test scenes (mean LDR luminance ≥ 0.40). These examples show that over-exposure can reduce apparent edge density in the input LDR, even when the underlying HDR scene contains substantial structure. In such bright-input, moderate-to-low-edge-density cases, our method outperforms most baselines.

J.1 Quantifying VAE-induced high-frequency attenuation

We analyze the contribution of the SVD VAE to the loss of fine detail observed on well-exposed, high-edge-density inputs. For each ground-truth HDR image in the SI-HDR test set, we clip intensities to the 99th percentile, rescale the peak to $4000\,\text{cd/m}^2$, and apply PU-21 encoding to obtain a perceptually uniform image in the range $[0, 1]$. We then encode and decode this image with the SVD VAE encoder and decoder, respectively. Using PU-21 avoids confounding VAE compression artifacts with the distribution mismatch between the training domain of the VAE and HDR images Wu et al. (2026).

To measure detail preservation, we compute the luminance of each image and apply a 3×3 Laplacian filter; the standard deviation of this Laplacian filtered image serves as our high-frequency energy measure. We report ρ_{VAE}, the ratio of this measure after VAE reconstruction to that of the original image. $\rho_{\text{VAE}} < 1$ indicates high-frequency detail lost solely by the VAE. For comparison, we compute the same ratio for our full HDR pipeline, ρ_{full}, after PU-21 encoding the predicted HDR output.

We average these ratios over two subsets: the 10 scenes where our method outperforms baselines on PSNR, and the 10 scenes where it underperforms the baselines. Results are shown in table 14. On textured scenes, the VAE preserves only 69% of the high-frequency measure, which drops to 45% after the full pipeline. On smoother scenes (due to over/under-exposure), the VAE preserves 89%, dropping to 72% after the full pipeline. The cross-scene gap in ρ_{VAE} accounts for roughly 75% of the gap in ρ_{full}, indicating that the SVD VAE is a major contributor to the high-frequency deficit on textured scenes.

This helps explain why non-generative methods such as SingleHDR and HDRCNN, or methods such as BracketDiff that do not rely on a latent VAE bottleneck, can potentially outperform our method on high-texture, well-exposed inputs. Importantly, this failure mode is tied to the VAE compression stage rather than to the exposure-bracket generation formulation itself. Moreover, these failure cases are largely well-exposed scenes that already preserve most visible content, and therefore benefit less from dynamic range expansion.

K Comparing with Flux-Based X2HDR

The original X2HDR-Flux model fine-tunes FLUX.1-dev, which uses a 12B-parameter rectified-flow transformer as its denoising network. In contrast, our method fine-tunes the 1.5B-parameter SVD diffusion UNet. We report a comparison of our method with X2HDR-Flux in table 15. The gap between X2HDR-SD2.1 and X2HDR-Flux indicates that backbone capacity is an important factor for HDR reconstruction. Nevertheless, our method outperforms both X2HDR variants on Q^*-PSNR

Method	Q^*-PSNR ↑	Q^*-SSIM ↑	Q^*-MAE ↓	Q^*-LPIPS ↓	Q-JOD ↑
Ours	$\mathbf{22.36 \pm 3.48}$	0.664 ± 0.156	$\mathbf{0.060 \pm 0.027}$	0.289 ± 0.111	7.51 ± 0.99
X2HDR-SD2.1	21.24 ± 3.32	0.585 ± 0.150	0.068 ± 0.030	0.335 ± 0.120	7.17 ± 0.97
X2HDR-Flux	21.37 ± 3.10	$\mathbf{0.726 \pm 0.108}$	0.067 ± 0.025	$\mathbf{0.232 \pm 0.106}$	$\mathbf{7.59 \pm 1.24}$

Table 15: **Comparison with different X2HDR backbones.** We compare against both our controlled X2HDR-SD2.1 implementation and the original Flux-based X2HDR model. Our method achieves the best Q^*-PSNR and Q^*-MAE, while X2HDR-Flux performs best on Q^*-SSIM, Q-JOD and Q^*-LPIPS.

and Q^*-MAE, while remaining comparable on Q-JOD. This suggests that explicitly generating an exposure bracket followed by learned fusion provides a strong reconstruction formulation, even when using a substantially smaller diffusion backbone. Scaling our method to video diffusion models with larger backbones would be an interesting next step for further performance improvements.

L Q-JOD and user study results

Q-JOD quantifies reference-based reconstruction quality under a specified display model, while the user study measures which reconstructions are perceived as closer to the ground truth by humans. We will clarify in the paper that Q-JOD aligns most reliably with human preference for large perceptual degradations corresponding to substantial JOD differences Hanji et al. (2022), rather than for small score gaps.

Reliability of small Q-JOD differences Quoting prior work Hanji et al. (2022), "the metrics can detect only very large image differences in this task and conducting a controlled experiment is the recommended option." These metrics include Q-JOD. Using a subset of the SI-HDR dataset, Hanji et. al [7] further estimate a minimum measurable difference of 0.786 Q-JOD at a 5% error rate in human studies. Thus, smaller Q-JOD gaps between our method and SingleHDR/HDRCNN (less than 0.15 Q-JOD on average) should not be interpreted as a reliable perceptual ranking. As this threshold was estimated on the same SI-HDR dataset used in our evaluation, it provides directly relevant evidence for interpreting these small differences with caution.

Perception-Distortion Tradeoff Distortion-based metrics like Q-JOD measure fidelity to a reference, while perceptual quality depends on how closely outputs follow the natural-image distribution Blau and Michaeli (2018). These objectives can conflict: single-image HDR reconstruction is an ill-posed task, yet deterministic methods such as HDRCNN and SingleHDR must produce a single output amongst multiple plausible solutions, which can favor averaged estimates. Hence, an output may achieve low reference distortion without appearing natural or being perceptually preferred Blau and Michaeli (2018); Delbracio and Milanfar (2023).

This is also illustrated in fig. 4: HDRCNN obtains a higher Q-JOD (7.85) despite unnatural blocky artifacts in the sky, whereas our generative method produces a more perceptually plausible sky (Q-JOD 7.46). We also outperform both methods on LPIPS, which is a more perceptually aligned metric.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We demonstrate evidence of the claims in the abstract in the Experiments and Results section.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In the limitations and future work section, we discuss the limitations of our method. We also visualize some failure cases in Figure 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We have mentioned the major hyperparameters in the Implementation details section of the paper.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We will release the code after acceptance.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We specify implementation details in the paper; additional details will be included in the supplement.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report standard deviations for the performance metrics over the test set in the main results table.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We mention the GPU used, memory usage, and optimization time for our algorithm.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We conform to the ethics guidelines.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: Our work has no particularly negative societal impacts.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our work poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer:[Yes]

    Justification: We have cited the datasets and models we used for comparison. These datasets have been widely used in published papers on this topic.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We will release our fine-tuned model with the code release of the paper.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: We include the task interface and protocol information in the Appendix.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: Yes, we obtained an IRB exemption for our study from our institute.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: This paper is not related to LLM research or applications.

Guidelines:

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