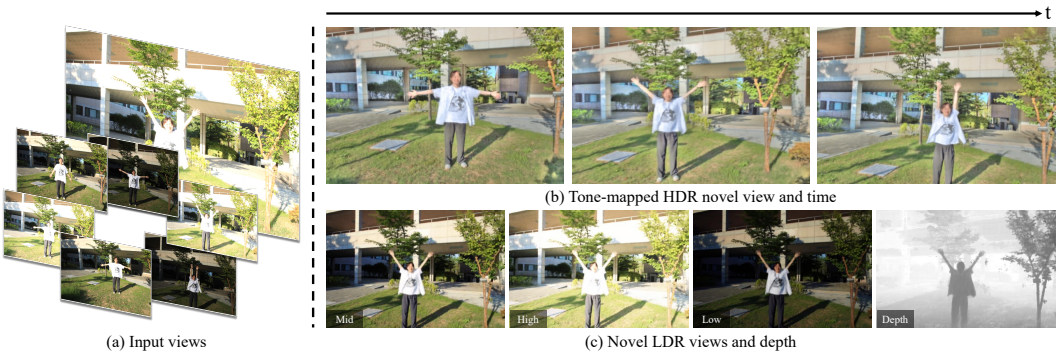
HDR-NSFF: HIGH DYNAMIC RANGE NEURAL SCENE FLOW FIELDS

Anonymous authors
Paper under double-blind review



(a) Input views (b) Tone-mapped HDR novel view and time (c) Novel LDR views and depth

Figure 1: **High Dynamic Range Neural Scene Flow Fields (HDR-NSFF)** reconstruct dynamic HDR radiance field from (a) alternatively exposed videos of dynamic scenes. Our method enables the rendering of (b) HDR novel views across both spatial and temporal domains. Additionally, we can generate (c) novel LDR views along with their corresponding depth maps.

ABSTRACT

Radiance of real-world scenes typically spans a much wider dynamic range than what standard cameras can capture, often leading to saturated highlights or underexposed shadows. While conventional HDR methods merge alternatively exposed frames, most approaches remain constrained to the 2D image plane, failing to model geometry and motion consistently. To address these limitations, we present HDR-NSFF, a novel framework for reconstructing dynamic HDR radiance fields from alternatively exposed monocular videos. Our method explicitly models 3D scene flow, HDR radiance, and tone mapping in a unified end-to-end pipeline. We further enhance robustness by (i) extending semantic-based optical flow with DINO features to achieve exposure-invariant motion estimation, and (ii) incorporating a generative prior as a regularizer to compensate for sparse-view and saturation-induced information loss. To enable systematic evaluation, we construct a real-world GoPro dataset with synchronized multi-exposure captures. Experiments demonstrate that HDR-NSFF achieves state-of-the-art performance in novel view and time synthesis, recovering fine radiance details and coherent dynamics even under challenging exposure variations and large motions.

1 INTRODUCTION

Radiance of real-world scenes typically spans a wider dynamic range than what standard cameras can capture (see Fig. 1). As a result, captures with standard cameras often suffer from overexposed highlights or underexposed shadows, leading to severe information loss in critical regions. A widely adopted strategy to address this limitation is high dynamic range (HDR) imaging, which captures multiple low dynamic range (LDR) frames at different exposures and merges them to form a HDR image. DR has become essential for enhancing realism and preserving radiometric information.

However, most existing HDR methods remain fundamentally constrained to the 2D image plane. Video-based HDR approaches (Chung and Cho, 2023; Xu et al., 2024; Cui et al., 2024) typically

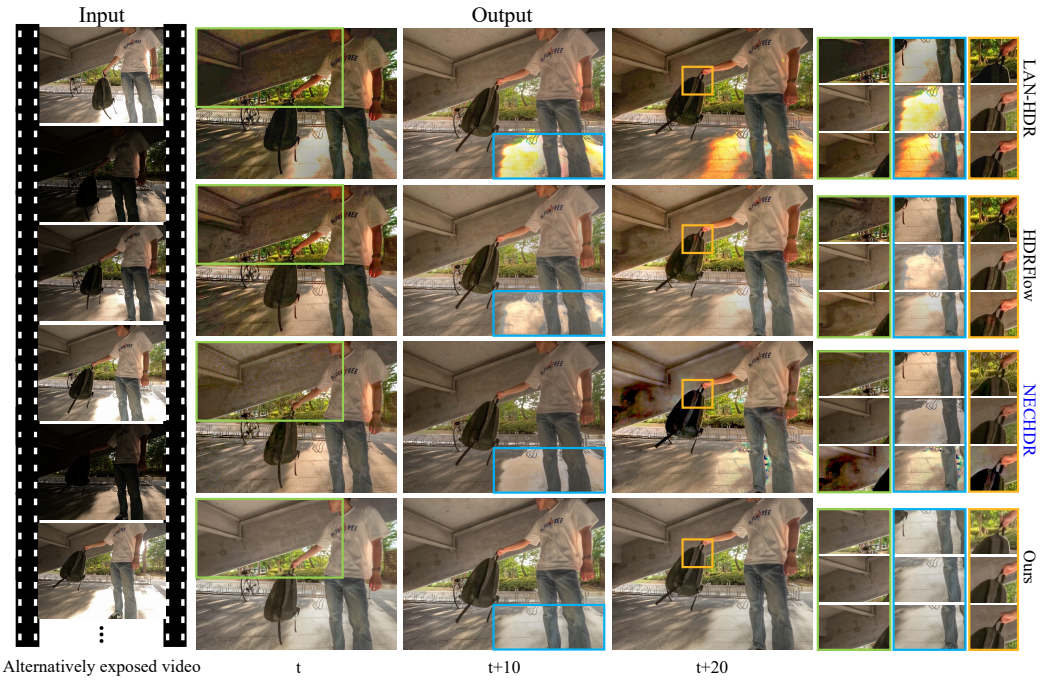


Figure 2: **Comparison of HDR video reconstruction on training views.** Given alternatively exposed video, HDR video reconstruction baselines, *i.e.*, LAN-HDR (Chung and Cho, 2023), HDRFlow (Xu et al., 2024), and NECHDR (Cui et al., 2024) to produce consistent results, while our model ensures temporal coherence and recovers valid information in saturated regions.

align consecutive frames and apply refinement to suppress ghosting and motion artifacts. Operating purely in image space, these methods cannot capture 3D motion, thus failing to handle occlusions, dynamics, varying radiance, and viewpoint changes (see Fig.2). These limitations clearly indicate the need to move beyond conventional 2D fusion toward a 3D representation.

In this work, we propose a framework for reconstructing HDR dynamic radiance fields from alternatively exposed monocular videos. We represent the scene as a continuous function of both space and time, whose outputs include HDR radiance, density, and 3D motion. Due to the additional variation in exposure, the HDR dynamic reconstruction attains a higher order of ill-posedness than reconstruction only across space and time. To address, we propose a dedicated pipeline built upon a dynamic neural radiance field (Li et al., 2021), as it leverages geometric and motion priors that can counteract the ill-posedness of the problem. Then, we jointly optimize the radiance field together with a learnable tone-mapping module, enabling HDR reconstruction and tone mapping in an end-to-end manner.

In designing the pipeline, a critical challenge of alternatively exposed video lies in the severe color inconsistency across frames, which induces combinatorial degradations spanning tone mapping, geometry, and motion priors. Thus, we analyze the robustness of each component to exposure variation and investigate the optimal combination among them. In particular, for motion prior, we observe that the semantic features of DINOv2 (Oquab et al., 2023) demonstrate strong robustness to illumination changes. Inspired by this observation, we extend DINO-Tracker (Tumanyan et al., 2024) to predict dense optical flow that remains reliable under varying exposures, and integrate these predictions into scene flow learning for dynamic HDR reconstruction.

Another major challenge lies in the correlation between the sparse-view nature of monocular videos and the information loss induced by saturation under extreme exposures. In other words, the state of a moving object can only be observed at specific timesteps, and if those observations are saturated, the result is an irrecoverable loss of information. To mitigate this issue, we incorporate generative priors (Wu et al., 2025) to compensate for the loss by augmenting the single training view with multi-view information and distilling it into the radiance field.

In addition to the synthetic dataset (Wu et al., 2024a), we evaluate our method on a newly constructed real-world dataset, which spans a wide range of scenarios including indoor and outdoor environments, diverse objects, and human subjects. Across both domains, our method consistently

outperforms existing baselines, including NeRF-W (Martin-Brualla et al., 2021), 4DGS (Wu et al., 2024b), MotionGS (Zhu et al., 2024) and HDR-Hexplane (Wu et al., 2024a), demonstrating superior reconstruction quality and robustness under challenging exposures.

To summarize, our key contributions are:

- **HDR-NSFF Framework:** We propose the first method that jointly models HDR scene flow fields enabling both novel view rendering and time interpolation.
- **Robust Learning Strategies:** We enhance scene flow learning by extending DINO-Tracker for exposure-robust motion estimation, and introduce generative priors as regularizers to overcome sparse-view limitations.
- **Comprehensive Evaluation:** We provide extensive experiments and a new real-world dataset with alternative exposures, demonstrating state-of-the-art performance in challenging HDR scenarios.

2 Related Work

High Dynamic Range Video Reconstruction. Creating HDR images from multi-exposure inputs is a long-studied problem in computational photography. A long line of work reconstructs HDR video by aligning and fusing alternatively exposed LDR frames (Kang et al., 2003; Kalantari et al., 2017; Chen et al., 2021; Chung and Cho, 2023; Xu et al., 2024; Cui et al., 2024). These approaches typically rely on optical flow or CNN-based alignment in 2D, followed by refinement to suppress ghosting. While effective for moderate motion, they remain vulnerable to occlusions, large displacements, and exposure inconsistencies. In contrast, our work reconstructs HDR video in 3D, enabling consistent rendering even under challenging dynamics.

Dynamic Scene Reconstruction. NeRF-based methods such as NSFF (Li et al., 2021), DynIBaR (Li et al., 2023), HyperNeRF (Park et al., 2021), and factorized grid models like HexPlane (Cao and Johnson, 2023) and K-Planes (Fridovich-Keil et al., 2023) have advanced free-viewpoint rendering of dynamic scenes. These methods represent a scene as a continuous function of space and time, sometimes augmented with deformation fields or canonical templates. They can synthesize novel views or even novel time steps. In parallel, 3D Gaussian Splatting has recently been extended to dynamic settings through 4DGS (Wu et al., 2024b), MotionGS (Zhu et al., 2024), Gaussian Marbles (Stearns et al., 2024), and DeformableGS (Yang et al., 2024b), achieving high efficiency and real-time rendering. Despite their success, all of these methods assume photometrically consistent LDR inputs and do not address the challenges of HDR content. Thus, they struggle to faithfully represent scenes with extreme lighting variations, whereas our approach explicitly targets HDR reconstruction of dynamic radiance fields.

High Dynamic Range Novel View Synthesis. Several recent works integrate HDR modeling into volumetric representations, mainly for static scenes. HDR-NeRF (Huang et al., 2022) and HDR-Plenoxel (Jun-Seong et al., 2022) model radiance together with tone-mapping or exposure functions, enabling HDR novel view synthesis from multi-exposure data. GaussHDR (Liu et al., 2025) extends HDR reconstruction to Gaussian Splatting with local tone mapping, while LTM-NeRF (Huang et al., 2024) embeds spatially varying tone mapping directly into NeRF. These works demonstrate the benefits of HDR-aware radiance fields but assume static content. The most relevant to our work is HDR-HexPlane (Wu et al., 2024a), which extends a factorized grid representation to dynamic HDR scenes by learning per-image exposure mappings. However, it does not explicitly model 3D motion, limiting its ability to represent complex dynamics and to perform temporal synthesis. In contrast, our method incorporates explicit motion modeling, allowing robust HDR reconstruction from real-world alternating-exposure videos and supporting both novel-view and novel-time rendering.

3 Preliminary

Neural Scene Flow Fields. Neural Scene Flow Fields (NSFF) extend NeRF (Mildenhall et al., 2020) by jointly modeling static and dynamic components of a scene. The dynamic branch, F_θ^{dy}, takes spatial location $\mathbf{x}$, view direction $\mathbf{d}$, and time t as inputs, and predicts color c_t^{dy}, density σ_t^{dy},

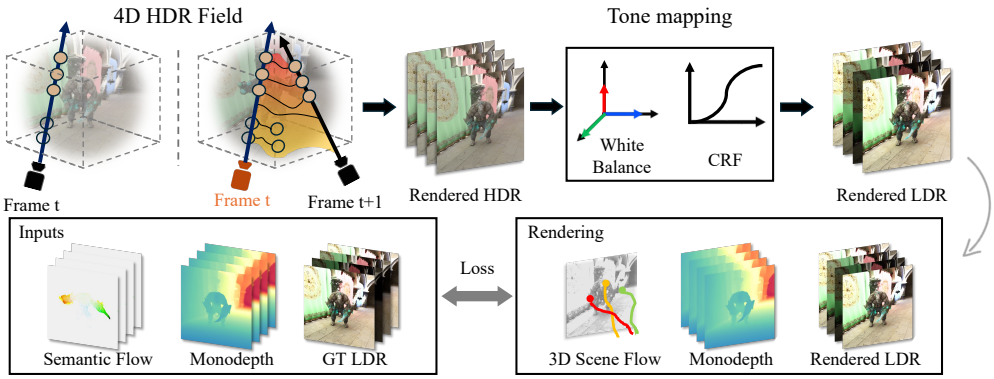


Figure 3: **Overall pipeline of our propsed method.** HDR-NSFF takes an alternatively exposed video as input and estimate 3D scene flow for the sampled points along each ray. Neighboring frames are then warped to render the HDR radiance at the target frame, which is tone-mapped to LDR via a white-balance and camera-response function module. Photometric loss with the ground-truth LDR images, along with optical flow and depth constraints from off-the-shelf models, jointly optimize both the scene flow fields and tone-mapping module in an end-to-end manner.

forward/backward scene flow F_t, and disocclusion weights W_t:

$$(c_t^{\text{dy}}, \sigma_t^{\text{dy}}, F_t, W_t) = F_\theta^{\text{dy}}(\mathbf{x}, \mathbf{d}, t). \tag{1}$$

Scene flow is used to warp 3D points across time for enforcing temporal consistency. The static branch, F_θ^{st}, models time-invariant appearance:

$$(c^{\text{st}}, \sigma^{\text{st}}, v) = F_\theta^{\text{st}}(\mathbf{x}, \mathbf{d}), \tag{2}$$

where v is a blending weight. The final color is obtained by volume rendering with static–dynamic combination:

$$\hat{C}_i(r_i) = \int_{z_n}^{z_f} T_i(z) \left[v(z) c^{\text{st}}(z) \sigma^{\text{st}}(z) + (1 - v(z)) c_i^{\text{dy}}(z) \sigma_i^{\text{dy}}(z) \right] dz. \tag{3}$$

Here, $T_i(z)$ denotes transmittance along the ray. This formulation allows NSFF to capture both persistent geometry and spatio-temporal dependent motion within a unified radiance field. Details are provided in the Appendix B.

4 Approach

Our framework builds upon Neural Scene Flow Fields (NSFF) to reconstruct dynamic HDR radiance fields from alternatively exposed monocular videos, as it offers explicit and stable 3D motion modeling. NSFF exploits physical priors such as depth and optical flow for consistent learning in LDR videos, while recent HDR radiance field methods introduce tone-mapping modules but remain limited to static scenes. However, a direct combination of these ideas is not sufficient for HDR video.

Dynamic HDR videos present fundamental challenges: alternating exposures cause severe color inconsistency, which (i) prevents off-the-shelf models from delivering reliable performance and (ii) limits the effectiveness of tone-mapping regularization. Addressing this requires a systematic approach that disentangles and rethinks each component in light of HDR-specific aspects. Building on this perspective, HDR-NSFF is designed as an integrated framework that introduces tailored modules and empirically grounded analyses, offering a coherent solution for dynamic HDR 4D reconstruction.

HDR-NSFF integrates three core components: (i) NSFF-based radiance field and tone-mapping joint optimization, where we experimentally analyze tone-mapping function (Sec. 4.1), (ii) generative prior regularization to compensate for the sparse-view limitation of monocular input (Sec. 4.2), (iii) exposure-robust semantic flow estimation for reliable motion learning, and robust depth estimation using a carefully selected model verified through empirical analysis (Sec. 5.1). An overview of the pipeline is illustrated in Fig. 3.

4.1 TONE-MAPPING

A key challenge in dynamic HDR reconstruction is the mismatch between multi-exposure LDR observations and the underlying HDR radiance. To bridge this gap, we introduce a tone-mapping module $\mathcal{T}$ with radiometric paramter θ that maps rendered HDR radiance E to the LDR domain:

$$C = \mathcal{T}(E, \theta) = g\big(w(E)\big), \tag{4}$$

where w applies per-channel white balance correction and g denotes the camera response function (CRF). To ensure stable optimization under extreme exposures, we employ a leaky-thresholded CRF that mitigates saturation effects, along with a smoothness regularization that encourages physically plausible CRF shapes through second-order derivative penalties. These components provide both flexibility and regularization, enabling $\mathcal{T}$ to form consistent HDR supervision across varying exposure levels and maintain a coherent radiance field in 3D space. Details are provided in the Appendix C.

4.2 GENERATIVE PRIOR AS A REGULARIZER

Reconstructing dynamic HDR scenes from monocular videos is particularly challenging due to the coupled effects of sparse temporal observations and information loss caused by saturation under extreme exposures. At each timestep, only a single viewpoint is available, and if this observation is saturated, the lost information cannot be recovered directly.

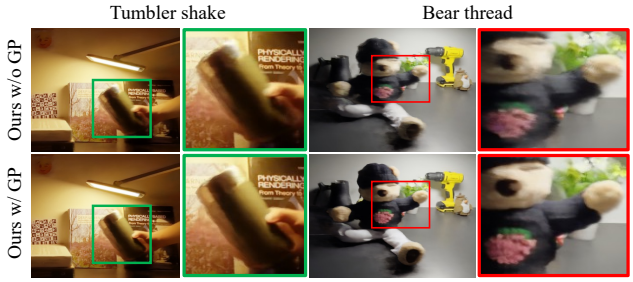


Figure 4: **Ablation on Generative Prior(GP).** GP provides additional plausible views from nearby angles, leading to more consistent and sharper reconstructions.

To mitigate this issue, we adopt a generative prior to compensate for information loss, extending its use from static scene reconstruction (Wu et al., 2025) to dynamic HDR scene reconstruction. The key idea is to periodically render unobserved or intermediate viewpoints, enhance them using a generative prior, and re-introduce these enhanced images as pseudo-observations during optimization. This allows HDR-NSFF to recover semantically plausible structures.

Let $\hat{C}$ be the HDR-NSFF rendering from a candidate novel viewpoint. At scheduled iterations, we obtain a generative enhancement $C^{\text{gen}} = \mathcal{G}(\hat{C})$ using the generative prior $\mathcal{G}$, and incorporate it as an auxiliary supervision signal. For each enhanced view, we apply a patch-wise perceptual loss:

$$\hat{\mathcal{L}}_{\text{gen}} = \sum_{p \in \mathcal{P}} \left\| \phi(\hat{C}_p) - \phi(C^{\text{gen}}_p) \right\|_1, \tag{5}$$

where ϕ denotes the perceptual encoder and p indexes spatial patches. This encourages HDR-NSFF to align with semantically consistent and radiometrically complete reconstructions. Since generative priors may introduce hallucinations when used too aggressively, we carefully control how and when they influence training. Specifically, generative pseudo-observations are activated only after an initial warm-up period:

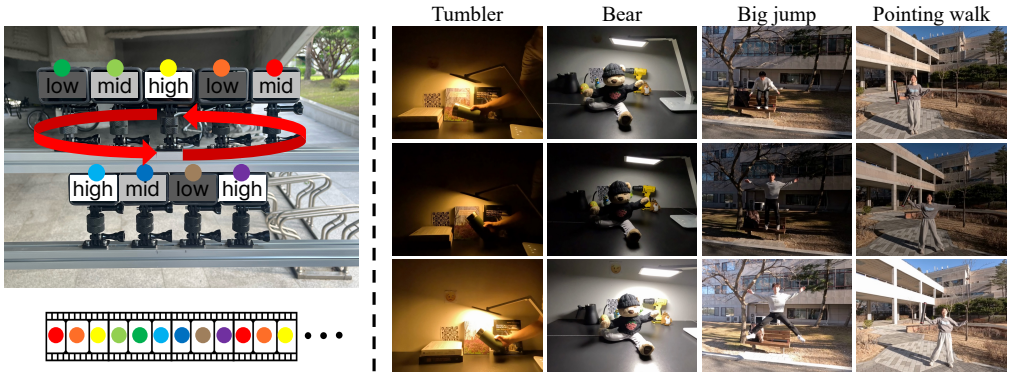
$$\alpha_{\text{gen}}(t) = \begin{cases} 0, & t < T_{\text{warm}} \\ p_{\text{gen}}, & t \geq T_{\text{warm}} \end{cases} \tag{6}$$

where $T_{\text{warm}} = 200{,}000$ iterations and $p_{\text{gen}} = 0.1$ is the sampling probability per iteration. The final effective generative loss is:

$$\mathcal{L}_{\text{gen}} = \alpha_{\text{gen}}(t)\, \beta_{\text{gen}}\, \hat{\mathcal{L}}_{\text{gen}}, \tag{7}$$

with β_{gen} controlling the overall contribution.

Figure 4 shows that the generative prior significantly improves geometric fidelity, radiance completeness, and perceptual consistency across space and time. This regularization is particularly beneficial in saturated regions and sparsely observed views, enabling HDR-NSFF to produce coherent and visually plausible HDR reconstructions under challenging exposure variations.



(a) Capture setup (b) Sampled sequences

Figure 5: **Evaluation setup and sampled sequences from our proposed GoPro dataset**. To evaluate novel view synthesis, we use nine GoPro Hero 13 Black cameras arranged at two height levels with fixed intervals, synchronized to record multi-view video at three exposures (mid, low, high). We construct a monocular alternatively exposed video by selecting one frame per time step across exposures, and use the remaining views for evaluation. Note that the input of our method is a monocular video and the setup described here is designed to evaluate the system.

4.3 Objective Function

We train both the neural scene flow fields and the tone-mapping module by minimizing the Mean Absolute Error (MAE) between rendered LDR views and ground-truth frames. Following NSFF (Li et al., 2021), we replace the rendered color $\hat{C}$ with our tone-mapped output $\mathcal{T}(\hat{E})$, where $\hat{E}$ denotes the rendered HDR radiance. The superscript cb denotes the **combined** rendering that fuses static and dynamic components of the scene. The photometric losses are:

$$\mathcal{L}_{\text{cb}} = \sum_{r_i} \|\mathcal{T}(\hat{E}_i^{\text{cb}}(r_i)) - C_i(r_i)\|_1, \quad \text{and} \tag{8}$$

$$\mathcal{L}_{\text{photo}} = \sum_{r_i} \sum_{j \in \mathcal{N}(i)} \|\mathcal{T}(\hat{E}_{j\to i}(r_i)) - C_i(r_i)\|_1, \tag{9}$$

where r denotes a camera ray. Here, $\hat{E}_{j\to i}(r_i)$ denotes the HDR radiance warped from a frame j to i. We also adopt the optical flow and single-view depth prior, denoted $\mathcal{L}_{\text{Flow}}$ and $\mathcal{L}_{\text{depth}}$ to regularize monocular reconstruction followed by NSFF (Li et al., 2021). For the CRF and generative prior objective functions, we apply $\mathcal{L}_{\text{smooth}}$ and $\mathcal{L}_{\text{gen}}$, respectively. The total objective function of our HDR-NSFF is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cb}} + \mathcal{L}_{\text{photo}} + \beta_{\text{data}}\mathcal{L}_{\text{data}} + \beta_{\text{reg}}\mathcal{L}_{\text{reg}} + \beta_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{gen}}, \tag{10}$$

where β are coefficients weight each term. The details can be found in the the appendix.

4.4 Datasets

Proposed GoPro Dataset. While standard alternatively exposed videos are sufficient for training HDR-NSFF, a single-camera setup cannot support evaluating novel view/time synthesis under varying exposures. To address this, we construct a real-world dataset captured with nine GoPro Hero 13 cameras configured at three exposures (low, mid, high). All cameras were fixed during capture and temporally synchronized using the built-in software. The cameras were manually aligned to face the same direction, forming an approximately parallel multi-view configuration. This setup enables consistent viewpoint sampling while maintaining controlled exposure variations.

This dataset provides the first benchmark for dynamic HDR reconstruction in real-world settings with explicit multi-exposure variation across viewpoints and comprises 12 diverse scenes covering indoor and outdoor environments under fast and complex motions (see Fig. 5). Inspired by prior work (Yoon et al., 2020), we adopt a similar strategy but adapt it for exposure alternation: at each timestamp, we select one frame per viewpoint from a single camera for training while reserving the remaining views for evaluating novel view and novel exposure performance.

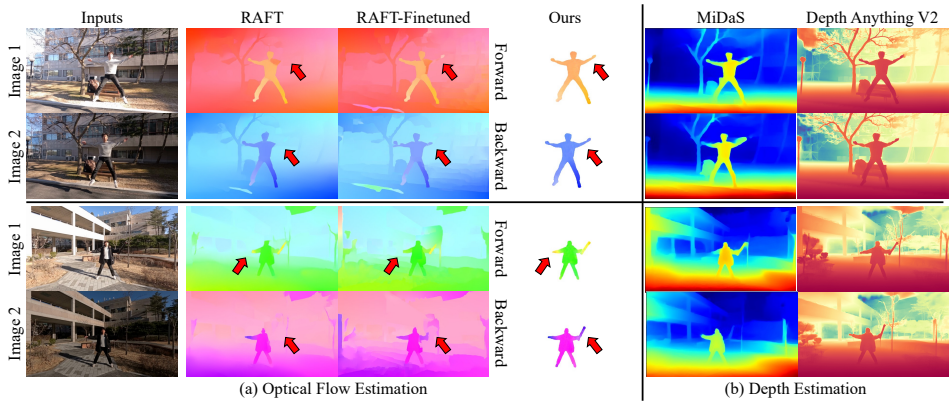


Figure 6: **Visualization of flow and depth estimation between varying exposed images.** (a) RAFT often fails under varying exposure conditions, yielding noticeable errors. Fine-tuning on synthetic varying exposed data (RAFT-Finetuned) improves performance moderately, but our semantic-based approach achieves higher accuracy. As highlighted by the red arrows, RAFT and RAFT-Finetuned miss correct motion. (b) For depth estimation, Depth Anything V2 (Yang et al., 2024a) recovers finer structural details and sharper object boundaries compared to MiDaS (Ranftl et al., 2020).

5 Experiments

HDR-NSFF takes as input an alternatively exposed monocular video and jointly reconstructs HDR radiance, 3D motion, and tone-mapping. Before evaluating the full reconstruction pipeline, we first analyze how each module should be designed to handle exposure-varying inputs. We study this under *Exposure-Robust Learning Strategies*, independently assessing (i) optical flow, (ii) depth estimation, and (iii) tone-mapping.

After establishing robust learning strategies, we evaluate HDR-NSFF on full HDR 4D reconstruction tasks, including novel view synthesis, novel time synthesis, and combined view-time synthesis on both real and synthetic datasets. We also include a two-stage 2D-to-4D HDR baseline that first reconstructs HDR video using LAN-HDR (Chung and Cho, 2023), HDRFlow (Xu et al., 2024), and NECHDR (Cui et al., 2024), then applies MoSca (Lei et al., 2025) for 4D reconstruction, enabling a direct comparison between end-to-end HDR 4D modeling and HDR-preprocessing pipelines.

All methods are evaluated using PSNR, SSIM (Wang et al., 2003), and LPIPS (Zhang et al., 2018). For HDR visualization, we use same Photomatix Pro tone-mapping operator.

5.1 Exposure robust learning strategies

Semantic based optical flow. A key challenge in reconstructing HDR dynamic scenes from alternatively exposed video is that frame-to-frame color inconsistencies significantly degrade the reliability of conventional optical flow methods. Standard alignment techniques such as RAFT (Teed and Deng, 2020) often fail under severe exposure variations (see Fig. 6 (a)).

In this context, we focus on the abundant embedding space of the self-supervised vision foundation model DINOv2 (Oquab et al., 2023), which has demonstrated strong robustness to photometric corruptions and perturbations, as shown by experiments on ImageNet-C (Hendrycks and Dietterich, 2019). We further investigate and observe the feature consistency, *i.e.*, robustness, across multi-exposure settings. These analyses are provided in the appendix. Built upon these observations, we adopt a DINOv2-based point tracking method, DINO-Tracker (Tumanyan et al., 2024), as motion estimation method, with a simple yet effective modification to ensure compatibility with our pipeline.

Since tracking errors accumulate with increasing frames under exposure variance, we redefine tracking points at each timestep and estimate only the flow between adjacent frame pairs in both forward and backward directions, as required by our pipeline. we also introduce motion masks from SAM2 (Ravi et al., 2024) to restrict DINO-Tracker to operate only within motion regions for preventing from noisy tracking performance in background. As a result, our semantic-based optical flow achieves robust motion estimation even in the presence of severe exposure variation (Fig. 6), providing consistent motion cues that are critical for HDR-NSFF.

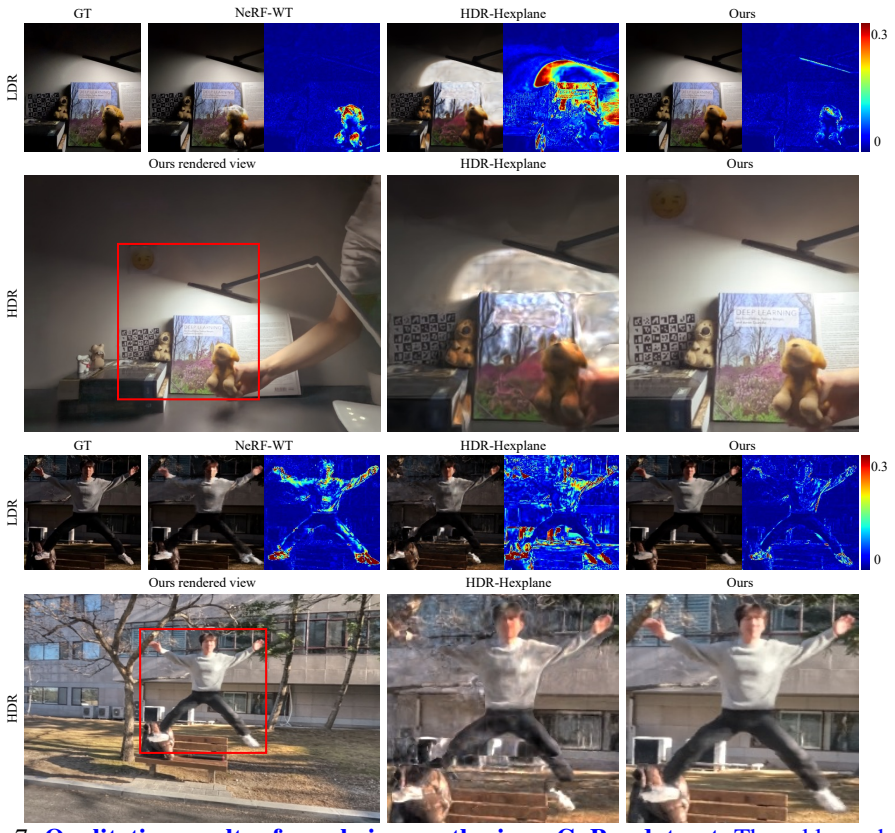


Figure 7: **Qualitative results of novel view synthesis on GoPro dataset.** The odd-numbered rows show the LDR-rendered novel views along with their corresponding L1 error maps against the ground-truth novel view (leftmost). Our method consistently yields the smallest error across all scenes. The even-numbered rows present tone-mapped HDR novel views. Compared with HDR-HexPlane (Wu et al., 2024a), our approach produces more accurate radiance, geometry, and motion representations.

Depth analysis under varying exposure. We assess the robustness of off-the-shelf depth estimators under exposure variation by synthetically generating ±2 EV versions of input images. RGB images are first converted into pseudo-RAW using a learned ISP inversion model (Xing et al., 2021), after which ±2 EV renderings are produced via standard sRGB mapping. Following the protocol of Ke et al. (Ke et al., 2024), we evaluate AbsRel on NYUv2 and ScanNet (Silberman et al., 2012; Dai et al., 2017). As shown in Table 1, while all methods degrade under ±2 EV shifts, Depth-Anything-V2 remains significantly more robust than alternatives. Based on this observation, we adopt Depth-Anything-V2 as the geometric prior in our pipeline.

Methods	NYUv2			ScanNet		
	Original	+2EV	-2EV	Original	+2EV	-2EV
MiDaS	9.08	13.68	9.35	8.66	13.78	10.22
DPT	9.21	12.96	8.95	8.27	13.62	9.57
Marigold	5.81	11.26	6.66	7.24	14.26	8.33
Depth-Anything-V2	4.87	7.63	5.10	4.82	10.57	6.36

Table 1: **Depth estimation results under exposure variance.** We employ AbsRel as the evaluation metric.

Tone-mapping module analysis.. Prior HDR radiance studies explored three CRF designs: a non-learnable fixed CRF (Wu et al., 2024a), a fully learnable MLP-based CRF (Huang et al., 2022), and a piecewise parametric CRF with per-channel white-balance factors (Jun-Seong et al., 2022). The fixed CRF offers strong regularization but insufficient flexibility, whereas the MLP CRF is overly flexible and often unstable. We adopt the piecewise CRF, which provides a balanced formulation. On our GoPro dataset, it achieves the best novel-view synthesis performance (Table 2), indicating that moderate flexibility with structured regularization is most effective under varying exposures.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o Tone-mapping	17.79	0.7048	0.0705	15.59	0.5577	0.1339
Fix CRF	25.55	0.8391	0.0487	20.43	0.6904	0.0911
MLP CRF	28.76	0.8861	0.0394	21.48	0.7256	0.0776
Piecewise CRF	31.01	0.9301	0.0233	22.55	0.7714	0.0697

Table 2: **Comparison of tone-mapping designs.**

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF	18.02	0.6792	0.2061	17.59	0.5473	0.2329
4DGS	20.94	0.7905	0.1541	17.83	0.5524	0.2230
MotionGS	14.61	0.3976	0.3617	12.33	0.2303	0.4696
NeRF-WT	29.70	0.9333	0.0598	19.25	0.6335	0.1770
HDR-HexPlane	20.70	0.6694	0.1917	20.55	0.6629	0.1716
Ours (w/o GP & DT)	31.04	0.9364	0.0621	24.93	0.8068	0.1048
Ours (w/o GP)	32.66	0.9447	0.0557	25.65	0.8205	0.1012
Ours	32.63	0.9444	0.0554	25.50	0.8208	0.0972

Table 3: **Averaged quantitative results of novel view synthesis on GoPro dataset.** Ours achieves the best overall performance, with DINO-Tracker (DT) offering the strongest improvement in motion-consistent reconstruction and the generative prior (GP) further enhancing perceptual quality.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF	19.01	0.7258	0.1976	18.84	0.5873	0.2531
HDR-HexPlane	20.46	0.6583	0.1933	19.59	0.6107	0.1855
Ours (w/o GP & DT)	31.31	0.9392	0.0648	24.85	0.7979	0.1372
Ours (w/o GP)	32.79	0.9451	0.0596	25.40	0.8075	0.1378
Ours	32.75	0.9448	0.0594	25.26	0.8070	0.1339

Table 4: **Averaged quantitative results of novel view and time synthesis on GoPro dataset.** Our method outperform baseline models.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF	15.98	0.6457	0.1388	16.04	0.5697	0.1527
NeRF-WT	31.10	0.9366	0.0342	21.50	0.7490	0.0895
HDR-HexPlane	29.95	0.9055	0.0527	23.87	0.7999	0.1071
Ours	35.07	0.9465	0.0483	27.19	0.8836	0.0576

Table 5: **Averaged quantitative results of novel view and time synthesis on synthetic data.** Our method outperform baseline models.

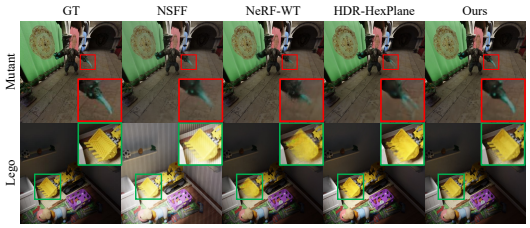


Figure 8: **Qualitative results of novel view and time synthesis on synthetic data.** Since, our approach explicitly models scene flow, it excels at time interpolation.

5.2 Results

Novel view synthesis. We evaluate novel view synthesis on our proposed GoPro dataset. For each time instance, we render the scene from all camera poses not used during training and apply the corresponding learned tone-mapping functions to convert the HDR renders to LDR. We then compare these tone-mapped views against the GT LDR images. It directly assesses two key aspects: (1) the quality of dynamic scene modeling, and (2) the accuracy of tone-mapping functions. Table 3 shows that our approach achieves significant improvements in rendering fidelity compared to baselines, both in highly dynamic regions and across the entire scene. Figure 7 its effectiveness in reconstructing HDR scenes with fine detail across varying exposures. Methods without appearance embedding (NSFF (Li et al., 2021), 4DGS (Wu et al., 2024b), MotionGS (Zhu et al., 2024)) fail to reconstruct consistent HDR views under alternating exposures. NeRF-WT (Quei-An, 2020) and HDR-Hexplane (Wu et al., 2024a) provide limited robustness but still struggle in real-world dynamic settings.

Novel view and time synthesis. We also evaluate novel view and time synthesis to demonstrate our method's ability to handle dynamic scenes with sparse temporal sampling (see Fig. 8). Following NSFF (Li et al., 2021), we remove every other frame from the original video sequences during training, and use the intermediate frames at held-out camera viewpoints for testing. Table 5 shows that our results outperform competing models across all evaluation metrics.

For real-world evaluation on our GoPro dataset, we extend this setting to simultaneously test novel view and time synthesis. While all camera views are retained to ensure realistic multi-view coverage, we subsample frames from each video and evaluate the model at unseen time instances and camera viewpoints. This joint evaluation directly measures the fidelity of both HDR radiance reconstruction and learned 3D motion under exposure-varying, dynamic scenes. Importantly, in this experiment as well, our model consistently surpasses all baseline methods (see Table 4).

While HDR-NSFF explicitly models 3D scene motion, enabling reliable synthesis across both space and time. In contrast, HDR-HexPlane does not incorporate explicit motion modeling, which limits its ability to handle space and time interpolation in dynamic HDR scenes.

Qualitative comparison of HDR reconstruction. To validate our HDR reconstruction, we qualitatively compare our results with ground-truth HDR images (see Fig. 9). Tone-mapped HDR views from our model closely match ground truth, preserving fine details in both under- and overexposed regions. Histograms of pixel intensities further show that our reconstructions cover the full radiance range, recovering values from very low to high intensities. In addition, novel LDR views rendered at multiple exposures confirm that our method accurately controls exposure.

Figure 9: **Qualitative results on novel LDR/HDR view synthesis.** We visualize LDR rendering results at varying exposure levels (low, mid, and high), tone-mapped HDR rendering by ours and corresponding ground-truth HDR references. We also visualize histograms of our HDR images and ground truth. For better visualization, we plot HDR histogram using smoothed kde method.

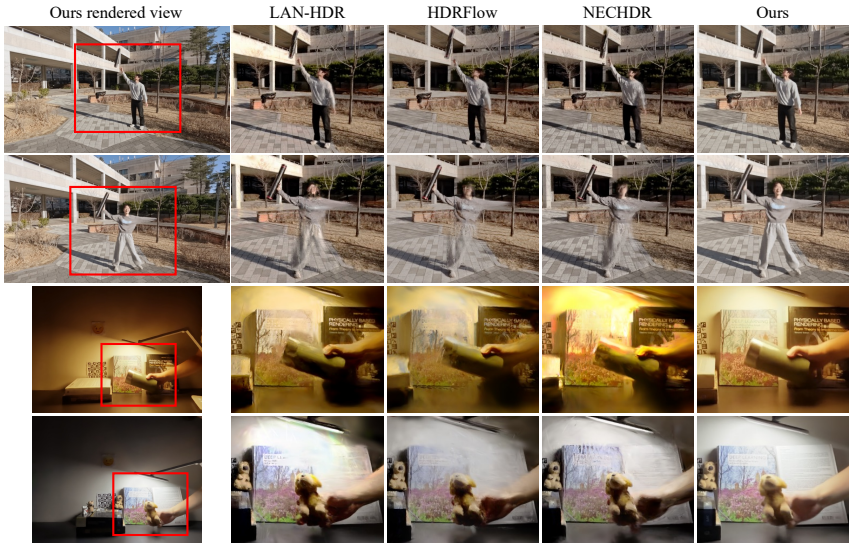


Figure 10: **Qualitative comparison with 2D-to-4D HDR reconstruction.** Two-stage baselines first reconstruct HDR video using LAN-HDR, HDRFlow, and NECHDR, then apply MoSca for 4D reconstruction. Our method produce more coherent HDR radiance and stable geometry.

Two-stage 2D-to-4D HDR reconstruction. We compare HDR-NSFF with a two-stage baseline that first reconstructs HDR video using 2D HDR approaches Chung and Cho (2023); Xu et al. (2024); Cui et al. (2024)and then applies MoSca (Lei et al., 2025) for 4D reconstruction. Although recent 2D HDR video methods incorporate motion cues within the image domain, they do not build a 3D representation of geometry and motion. Consequently, exposure-inconsistent frames reconstructed in 2D may still contain radiometric or geometric deviations, and these inaccuracies propagate to the subsequent 4D reconstruction stage, leading to less stable radiance fields and geometry (see Fig. 10). In contrast, HDR-NSFF jointly models radiance, geometry, and motion within a end-to-end manner, resulting in more consistent reconstructions under challenging exposure alternation. You can find more details in Appendix D.3.

6 Conclusion

In this work, we introduced HDR-NSFF, the first framework that jointly reconstructs HDR radiance, 3D motion, and tone-mapping from alternatively exposed monocular videos. By explicitly modeling scene flow and integrating learnable tone-mapping, our approach addresses the fundamental limitations of prior HDR methods that operate purely in 2D image space. We further enhanced robustness through semantic-based optical flow, depth priors, and generative prior, enabling reliable reconstructions under severe exposure variations and sparse temporal observations. Extensive experiments on both real and synthetic datasets demonstrated that HDR-NSFF consistently outperforms baselines across novel view synthesis, novel time synthesis, and combined view-time synthesis. In particular, our method achieves sharper geometry, more faithful HDR radiance, and temporally coherent results compared to state-of-the-art dynamic scene and HDR reconstruction models.

REFERENCES

Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *CVPR*, 2023.

Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2502–2511, 2021.

Haesoo Chung and Nam Ik Cho. Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12760–12769, 2023.

Jiahao Cui, Wei Jiang, Zhan Peng, Zhiyu Pan, and Zhiguo Cao. Exposure completing for temporally consistent neural high dynamic range video rendering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10027–10035, 2024.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 465–474. 2023.

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022.

Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, and Qing Wang. Ltm-nerf: Embedding 3d local tone mapping in hdr neural radiance field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10944–10959, 2024.

Kim Jun-Seong, Kim Yu-Ji, Moon Ye-Bin, and Tae-Hyun Oh. Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In *European Conference on Computer Vision*, pages 384–401. Springer, 2022.

Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. In *ACM TOG*, 2017.

Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions On Graphics (TOG)*, 22(3):319–325, 2003.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Proc. ECCV*, 2024.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

Jiahui Lei, Yijia Weng, Adam W Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6165–6177, 2025.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6498–6508, 2021.

Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *CVPR*, 2023.

Jinfeng Liu, Lingtong Kong, Bo Li, and Dan Xu. Gausshdr: High dynamic range gaussian splatting via learning unified 3d and 2d local tone mapping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5991–6000, 2025.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European conference on computer vision*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. In *ACM TOG*, 2021.

Chen Quei-An. Nerf pl: a pytorch-lightning implementation of nerf. 2020.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *ECCV*, 2020.

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision*, pages 367–385. Springer, 2024.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2003.

Guanjun Wu, Taoran Yi, Jiemin Fang, Wenyu Liu, and Xinggang Wang. Fast high dynamic range radiance fields for dynamic scenes. In *2024 International Conference on 3D Vision (3DV)*, 2024a.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024b.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *CVPR*, pages 26024–26035, 2025.

Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *CVPR*, 2021.

Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, and Xin Yang. Hdrflow: Real-time hdr video reconstruction with large motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24851–24860, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024a.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *CVPR*, 2024b.

Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *CVPR*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Ruijie Zhu, Yanzhe Liang, Hanzhi Chang, Jiacheng Deng, Jiahao Lu, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:101790–101817, 2024.

APPENDIX

This appendix provides additional details complementary to the main manuscript. In Section A, we describe the implementation details, including counterparts, dataset configuration, and experimental setup. Section B elaborates on the Neural Scene Flow Fields framework, covering the regularization terms. Section C presents HDR-NSFF–specific components such as tone-mapping, semantic tracking, generative prior formulation, and the full objective function. Finally, Section D includes additional experimental results, ablation studies, and the extension to 3DGS-based dynamic reconstruction, along with demonstrations of 2D-to-4D HDR reconstruction. A supplementary video further showcases novel-view rendering results.

CONTENTS

A IMPLEMENTATION DETAILS

A.1 COUNTERPARTS

In this chapter, we briefly explain the method we compared as a counterparts in our experiments.

NeRF-WT. NeRF-W (Martin-Brualla et al., 2021) introduces per-image appearance and transient embedding, modelling to handle dynamic changes such as lighting variations and moving objects. In our experiments, we adapted NeRF-W to a dynamic HDR video (named NeRF-WT) using appearance embedding for ISP modelling and transient part for scene dynamics. We follow the hyperparameters given in the codebase. For implementation we used the codebase in `https://github.com/kwea123/nerf_pl`

HDR-Hexplane. HDR-Hexplane (Wu et al., 2024a) adopted Hexplane (Cao and Johnson, 2023) for the dynamic 3D representation and MLP with exposure embeddings accompanied with fixed gamma function to optimize ISP module. We follow the hyperparameters following manuscript. For implementation we used the codebase in `https://github.com/hustvl/HDR-HexPlane`

A.2 DATASET

Synthetic. We select four synthetic scenes for evaluation: *Lego*, *Mutant*, *Jumping Jack*, and *Stand Up*. Each image has a resolution of 800×800, with exposure values spanning from -2EV to 5EV. To maximize the influence of exposure change, we carefully adjust the camera viewpoints and lighting directions.

The sampling rate is determined based on the motion speed of each scene. Specifically, the *Lego* scene is subsampled by selecting every 10th frame, whereas the remaining scenes are sampled by skipping every two frames.

Real. For the real dataset, we preset exposure time for each cameras before acquisition. We set exposure time differently for each sequence. Sequence lengths and corresponding exposure information are detailed in the Table S1 All sequences are synchronized using the GoPro software.

A.3 EXPERIMENTAL SETUP

To facilitate understanding of the experimental setup employed for the real dataset experiments, we provide an illustrative diagram in Fig. S1 In the novel view synthesis experiment, performance is evaluated by measuring the differences between synthesized results and the images captured from cameras that were excluded from the training set, for all camera views i. In the novel view and time

Name	Exp. Time [s]	# of frames
Big jump	$\frac{1}{960}, \frac{1}{2880}, \frac{1}{7680}$	324
Side walk	$\frac{1}{960}, \frac{1}{2880}, \frac{1}{7680}$	324
Jumping jack	$\frac{1}{720}, \frac{1}{1920}, \frac{1}{7680}$	324
Pointing walk	$\frac{1}{720}, \frac{1}{1920}, \frac{1}{7680}$	324
Tube toss	$\frac{1}{720}, \frac{1}{1920}, \frac{1}{7680}$	324
Bear	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324
Dog	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324
Tumbler	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324
Fire extinguisher	$\frac{1}{480}, \frac{1}{960}, \frac{1}{1920}$	324
Laptop	$\frac{1}{480}, \frac{1}{960}, \frac{1}{1920}$	324
Bag	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324
Ball	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324

Table S1: **Parameter setting for real dataset**

synthesis experiment, we evaluate performance by holding out certain segments of the time sequence and measuring how accurately these withheld segments are inferred.

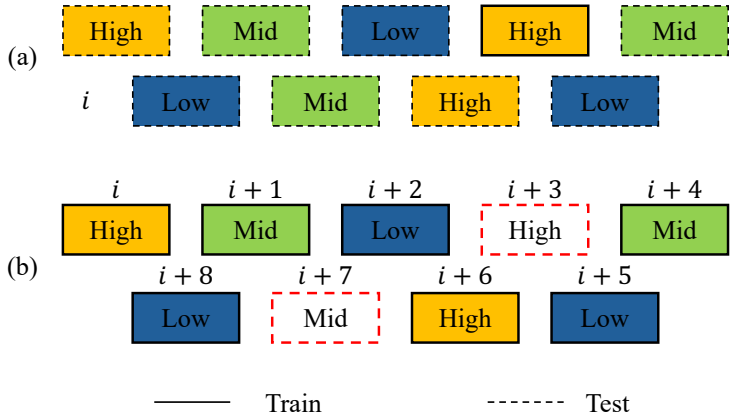


Figure S1: **Illustration of two experimental setting.** We illustrate two experimental settings described in Sec. 4.2 in manuscript: (a) Novel view synthesis (b) Novel view and time synthesis.

B Details of Neural Scene Flow Fields

To model dynamic scenes, NSFF (Li et al., 2021) extend the concept of NeRF (Mildenhall et al., 2020) by representing 3D motion as scene flow fields. NSFF learns a combination of static and dynamic NeRF representations. The dynamic model, denoted as F_θ^{dy}, explicitly models view and time dependent variations by incorporating time t as an additional input. Beyond color and density, it also predicts forward and backward 3D scene flow $F_t{=}(\mathbf{f}_{t\to t+1}, \mathbf{f}_{t\to t-1})$ and occlusion weights $W_t{=}(w_{t\to t+1}, w_{t\to t-1})$ to handle 3D motion disocclusion:

$$(c_t, \sigma_t, F_t, W_t) = F_\theta^{\text{dy}}(\mathbf{x}, \mathbf{d}, t). \tag{11}$$

To supervise scene flow estimation, NSFF uses temporal photometric consistency. Specifically, for each time i, scene flow is predicted for the 3D points sampled along rays, and this predicted flow is used to warp corresponding points from neighboring times $j \in \mathcal{N}(i)$ to time i. The color and opacity information of the warped points is then used to render the image at time i:

$$\hat{C}_{j\to i}(r_i) = \int_{z_n}^{z_f} T_j(z)\, \sigma_j\big(r_{i\to j}(z)\big)\, c_j\big(r_{i\to j}(z), d_i\big)\, dz, \tag{12}$$

$$\text{where} \quad r_{i\to j}(z) = r_i(z) + \mathbf{f}_{i\to j}\big(r_i(z)\big). \tag{13}$$

Temporal photometric consistency is enforced by minimizing the mean squared error (MSE) between the warped rendered view and the ground-truth image:

$$\mathcal{L}_{\text{photo}} = \sum_{r_i} \sum_{j \in \mathcal{N}(i)} \|\hat{C}_{j \to i}(r_i) - C_i(r_i)\|_2^2. \tag{14}$$

The static NeRF, F_θ^{st}, represents a time-invariant scene using a multilayer perceptron (MLP). Given an input position $\mathbf{x}$ and view direction $\mathbf{d}$, it outputs the RGB color c, volume density σ, and an unsupervised 3D mixing weight v that determines the blending between static and dynamic components:

$$(c, \sigma, v) = F_\theta^{\text{st}}(\mathbf{x}, \mathbf{d}). \tag{15}$$

Here, c_t and σ_t denote the color and volume density at position $\mathbf{x}$ at time t. The final color is computed by blending the static and dynamic components using the following rendering equation:

$$\hat{C}_i^{cb}(r_i) = \int_{z_n}^{z_f} T_i^{cb}(z)\, \sigma_i^{cb}(z)\, e_i^{cb}(z)\, dz, \tag{16}$$

where $\sigma_i^{cb}(z) c_i^{cb}(z)$ is a linear combination of static and dynamic scene components, weighted by $v(z)$:

$$\sigma_i^{cb}(z) c_i^{cb}(z) = v(z) c(z) \sigma(z) + (1 - v(z)) c_i(z) \sigma_i(z). \tag{17}$$

T_i represents the transmittance at time i, while z_n and z_f denote the near and far depths along the ray. The final rendered output $\hat{C}_i^{cb}(r_i)$ is optimized against the ground-truth pixel color $C_i(r_i)$ using a photometric loss:

$$\mathcal{L}_{cb} = \sum_{r_i} \|\hat{C}_i^{cb}(r_i) - C_i(r_i)\|_2^2. \tag{18}$$

Reconstructing dynamic scenes from monocular input is inherently ill-posed, and relying solely on photometric consistency often leads to convergence at poor local minima. Therefore, NSFF incorporates three additional guided losses: a term enforcing monocular depth and optical flow consistency, a motion trajectory term promoting cycle-consistency and spatiotemporal smoothness, and a compactness prior encouraging binary scene decomposition and reducing floaters via entropy and distortion losses.

Following section, we elaborate on data-driven prior loss (Flow loss and Single-view depth loss) and additional regularization terms introduced by NSFF (Li et al., 2021): Scene Flow Cycle Consistency and Low-Level regularization term. We employ additional regularization terms consistently in both our model and NSFF.

Flow Loss. Flow loss operates by minimizing the discrepancy between observed 2D pixel correspondences, computed from pretrained optical flow networks and predicted 2D pixel correspondences, obtained by projecting predicted 3D scene flows. This aligns 3D scene flow with pretrained 2D motion estimation.

Given two adjacent frames at times i and $j = i \pm 1$, L_{flow} is calculated as follows. Let p_i represent a pixel location at frame i. The corresponding pixel location at frame j, denoted by $p_{i \to j}$, can be computed using pretrained 2D motion estimation $u_{i \to j}$ as $p_{i \to j} = p_i + u_{i \to j}$.

The model predicts an expected scene flow $\hat{F}_{i \to j}(r_i)$ corresponding to 3D location $\hat{X}_i(r_i)$ along the ray r_i passing through the pixel p_i via volumetric rendering. Thus, the predicted 3D displacement can be expressed as $\hat{X}_i(r_i) + \hat{F}_{i \to j}(r_i)$. Then, by applying the perspective projection operator Π_j, corresponding to the camera viewpoint at frame j , the expected 2D pixel position $\hat{p}_{i \to j}(r_i)$ at frame j is calculated as:

$$\hat{p}_{i \to j}(r_i) = \Pi_j \left(\hat{X}_i(r_i) + \hat{F}_{i \to j}(r_i) \right). \tag{19}$$

Finally, the geometric consistency loss is computed by measuring the discrepancy between these two pixel positions (observed vs. predicted) using the L1-norm:

$$\mathcal{L}_{\text{flow}} = \sum_{r_i} \sum_{j \in \{i \pm 1\}} ||\hat{p}_{i \to j}(r_i) - p_{i \to j}(r_i)||_1. \tag{20}$$

Single-view Depth Prior. Encourages rendered depths to match predictions from a pretrained depth model:

$$\mathcal{L}_{\text{depth}} = \sum_{r_i} ||\hat{Z}_i^*(r_i) - Z_i^*(r_i)||_1, \tag{21}$$

where the superscript $(*)$ denotes scale-shift invariant normalization. These priors are combined into:

$$\mathcal{L}_{\text{data}} = \mathcal{L}_{\text{flow}} + \beta_{\text{depth}}\mathcal{L}_{\text{depth}}. \tag{22}$$

Scene Flow Cycle Consistency. To ensure plausible scene motion, the loss ensures coherence between forward and backward predicted scene flows for adjacent frames, mathematically defined as:

$$\mathcal{L}_{\text{cyc}} = \sum_{x_i} \sum_{j\in\{i\pm1\}} w_{i\to j}||f_{i\to j}(x_i) + f_{j\to i}(x_{i\to j})||_1, \tag{23}$$

where $f_{i\to j}(x_i)$ indicates the predicted displacement (scene flow) of point x_i from time i to j.

Low-Level Regularization. Spatial-temporal smoothness is enforced through $l1$ regularization on scene flow estimated between neighboring sampled 3D points along rays. This encourages 3D point trajectories to be piecewise linear. Another sparsity regularization term, calculating an $l1$ loss in flow estimation is also applied. This encourage minimal scene flow magnitudes across most spatial regions. It is composed of three equally weighted components: spatial smoothness, temporal smoothness, and minimal flow magnitude:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{sp}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{min}}.$$

Spatial Smoothness. Following NSFF, the spatial smoothness term encourages nearby 3D samples along the same camera ray to predict similar scene flows. For each sampled 3D location $\mathbf{x}_i$ on ray $\mathbf{r}_i$, we consider its neighboring samples $\mathcal{N}(\mathbf{x}_i)$ and penalize the weighted ℓ_1 discrepancy:

$$\mathcal{L}_{\text{sp}} = \sum_{\mathbf{x}_i} \sum_{\mathbf{y}_i\in\mathcal{N}(\mathbf{x}_i)} \sum_{j\in\{i\pm1\}} w^{\text{dist}}(\mathbf{x}_i,\mathbf{y}_i)\left\|f_{i\to j}(\mathbf{x}_i) - f_{i\to j}(\mathbf{y}_i)\right\|_1, \tag{1}$$

where the weight is based on Euclidean distance: $w^{\text{dist}}(\mathbf{x},\mathbf{y}) = \exp(-2\|\mathbf{x}-\mathbf{y}\|_2)$.

Temporal Smoothness. Inspired by the piecewise-linear motion prior of Vo et al., the temporal term encourages each 3D trajectory to maintain low kinetic energy. This is implemented by minimizing the squared norm of the sum of forward and backward scene flows:

$$\mathcal{L}_{\text{temp}} = \frac{1}{2}\sum_{\mathbf{x}_i}\left\|f_{i\to i+1}(\mathbf{x}_i) + f_{i\to i-1}(\mathbf{x}_i)\right\|_2^2. \tag{2}$$

Minimal Flow Prior. Finally, following the observation that most points in the scene exhibit small motion, we impose an ℓ_1 penalty on all predicted scene flows to encourage near-zero flows where appropriate:

$$\mathcal{L}_{\text{min}} = \sum_{\mathbf{x}_i} \sum_{j\in\{i\pm1\}} \left\|f_{i\to j}(\mathbf{x}_i)\right\|_1. \tag{3}$$

C Details of High Dynamic Range Neural Scene Flow Fields

Our method is built upon the NSFF framework and therefore inherits its core loss formulation and optimization structure. However, reconstructing HDR dynamic radiance fields from alternatively exposed monococular videos introduces unique challenges not addressed in the original NSFF design. To handle severe exposure fluctuations, saturation artifacts, and inconsistent appearance across viewpoints, we adapt NSFF's formulation by modifying the photometric loss (Eq. 9) and combined loss (Eq. 8), and by introducing additional regularizers tailored for HDR reconstruction. Specifically, our HDR-NSFF incorporates (i) a physically informed tone-mapping module, and (ii) a generative prior that recovers saturated or missing information. We describe each of these components in detail below, along with the updated objective function used for end-to-end optimization.

C.1 TONE-MAPPING

Our goal is to reconstruct HDR dynamic radiance fields, encompassing both 3D space and motion, from 2D multi-exposure LDR RGB images. A crucial component in this process is the tone-mapping module, which bridges the gap between varying 2D observations and a coherent 3D HDR representation. Specifically, tone-mapping module, $\mathcal{T}$ can be expressed as:

$$C = \mathcal{T}(E, \theta) = g\big(w(E)\big), \tag{24}$$

where E denotes the rendered radiance, w the white balance correction, g the camera response function (CRF), and θ the radiometric parameters.

The white balance function w applies per-channel scaling using the white balance parameter $\theta_w = [w_r, w_g, w_b]^\top \in \mathbb{R}^3$, producing a white balance-corrected image E_w. The CRF g is then applied to E_w, mapping it to the final LDR image C. The CRF is parameterized as a piecewise linear function, defined using 256 points uniformly sampled in the $[0, 1]$ range. Values exceeding the dynamic range are thresholded accordingly. During training, we adopt leaky-thresholding, to reduce saturation loss in rendered images:

$$g_{\text{leaky}}(x) = \begin{cases} \alpha x, & x < 0 \\ g(x), & 0 \le x \le 1 \\ -\frac{\alpha}{\sqrt{x}} + \alpha + 1, & x > 1, \end{cases} \tag{25}$$

where α is the thresholding coefficient. This approach ensures effective color correction and dynamic range handling during HDR-NSFF training.

We incorporate a smoothness loss to enforce that CRF varies smoothly in a physically plausible manner Debevec et al. (2023). We penalize the second-order derivative of the CRFs: It is defined as follows:

$$\mathcal{L}_{\text{smooth}} = \sum\nolimits_{i=1}^{N} \sum\nolimits_{e \in [0,1]} g_i''(e), \tag{26}$$

where $g''(e)$ denotes the second order derivative of CRFs *w.r.t.* its input domain.

In the absence of a known camera response function (CRF), the choice of tone-mapping module $\mathcal{T}(\cdot, \theta)$ determines the flexibility with which HDR radiance can be effectively recovered from LDR inputs. Moreover, to build consistent HDR representations in 3D space, the tone-mapping module must also act as a regularizer, preventing fluctuations in HDR results under multi-exposure conditions. This combination of flexibility and regularization largely influences the overall quality and stability of HDR field reconstruction.

C.2 DINO-TRACKER

DINO-Tracker is a self-supervised framework designed to accurately track points over long sequences of video frames. Given an initial query point in an early frame of video, it estimates the trajectory of these points throughout subsequent frames. The method leverages pretrained deep features from the DINOv2-ViT (Oquab et al., 2023) model, which are refined by learning residual features via a small, trainable CNN module. DINO feature and residual feature are aggregated to find correspondence heatmap computed by cost volume. Lastly, additional CNN-refiner follows to further enhance matching.

Optimization is performed using several losses

- **Flow Loss** (L_{flow}): Ensures predicted trajectories align closely with short-term optical flow correspondences.
- **DINO Best-Buddies Loss** ($L_{\text{dino-bb}}$): Contrastively aligns refined features based on semantic matches from original DINO embeddings.
- **Refined Best-Buddies Loss** ($L_{\text{rfn-bb}}$): Similar to DINO best-buddies loss but applied to newly detected reliable matches among refined features.
- **Cycle-Consistency Loss** ($L_{\text{rfn-cc}}$): Encourages consistency in predicted trajectories, penalizing trajectories that fail a cycle-consistency criterion.
- **Prior Preservation Loss** (L_{prior}): Regularizes the refined features to remain close in norm and direction to original DINO features, ensuring semantic coherence is preserved.

Figure S2: **DINOv2 feature visualization under varying exposures.** Despite large changes in brightness, DINOv2 embeddings remain consistent, showing robust clustering across different exposure levels.

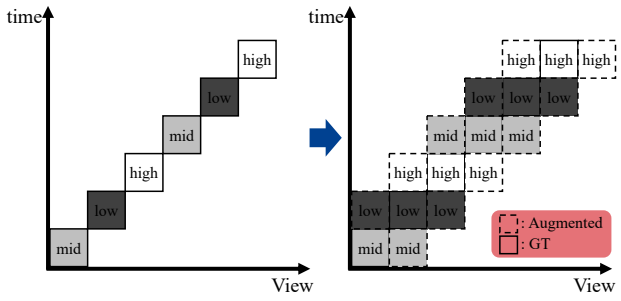


Figure S3: **Multi-view augmentation using generative prior.**

In contrast to the original DINO-Tracker, our proposed approach introduces a novel utilization of this framework explicitly aimed at enhancing the robustness and accuracy of 2D dense correspondence estimation. Specifically, we propose deriving dense matching from consecutive frames using the trained DINO-Tracker model itself. Leveraging the semantic matching capability inherent to DINO features, our method provides robust optical flow estimates even in challenging conditions such as alternatively exposed video settings, where conventional texture-based methods typically degrade due to information loss. Figure S2 shows that DINOv2 features is robust to exposure variance.

C.3 Generative prior for recovering saturated information

In HDR-NSFF, we additionally employ generative prior (Wu et al., 2025) as a regularizer to stabilize training under severe exposure inconsistencies. Generative prior provides a diffusion-based enhancement prior that guides the radiance field toward semantically consistent reconstructions when input frames suffer from brightness fluctuations or missing details. Concretely, we periodically generate pseudo-observations by enhancing intermediate renderings with the Difix prior and incorporate them into the optimization loop. This regularization not only improves geometric and radiometric stability but also enforces stronger multi-view consistency in dynamic scenes, where exposure variations and motion often break correspondences across views. As a result, HDR-NSFF achieves more coherent reconstructions that generalize better to unseen exposures and viewpoints.

Generative prior regularization. To mitigate the sparse-view limitation of monocular input, we adopt enhanced views generated via a generative prior (Wu et al., 2025). For these views, we apply a patch-wise perceptual loss to encourage realistic and view-consistent appearance:

$$\mathcal{L}_{\text{gen}} = \sum_{p \in \mathcal{P}} \|\phi(\hat{C}_p) - \phi(C_p^{\text{gen}})\|_1, \tag{27}$$

where ϕ denotes a perceptual feature extractor, and p indexes sampled patches. Since generative priors may introduce hallucinations, we carefully balance their contribution by (i) delaying their use until a stable stage of training (200K iterations), and (ii) training with enhanced views at a low probability (10%) per iteration.

Method	Big jump						Side walk					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	16.23	0.6268	0.194	13.75	0.4476	0.2755	16.62	0.6346	0.1874	14.1	0.5524	0.2139
4DGS	20.02	0.7283	0.1751	14.21	0.3724	0.319	18.46	0.702	0.1724	13.56	0.3927	0.254
MotionGS	11.4	0.1354	0.4549	9.42	0.0862	0.5628	13.58	0.2387	0.3692	8.79	0.1025	0.513
NeRF-WT	27.09	0.9051	0.0738	16.31	0.5023	0.2301	25.29	0.9061	0.0641	13.23	0.4243	0.2143
HDR-Hexplane	21.51	0.6235	0.2117	18.04	0.5653	0.2074	19.12	0.4931	0.2392	17.01	0.6214	0.1664
Ours	30.03	0.9239	0.0596	21.72	0.7494	0.1058	29.91	0.9263	0.0515	20.39	0.7132	0.1041

Method	Jumping jack						Pointing walk					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	17.44	0.7543	0.1251	18.19	0.5493	0.1274	17.16	0.7534	0.1206	14.07	0.6198	0.1745
4DGS	19.58	0.7486	0.1239	19.15	0.5733	0.1516	19.17	0.7373	0.1319	13.52	0.3664	0.2084
MotionGS	13.82	0.2474	0.3217	11.51	0.1244	0.4059	13.72	0.2472	0.3278	10.46	0.089	0.4774
NeRF-WT	30.93	0.9364	0.0384	19.94	0.6028	0.1384	27.36	0.9243	0.0488	15.26	0.5363	0.2018
HDR-Hexplane	17.24	0.4671	0.2112	19.89	0.5946	0.1466	16.97	0.4398	0.2111	16.67	0.6537	0.1674
Ours	31.83	0.9457	0.0353	23.66	0.7871	0.0721	29.72	0.9338	0.0433	20.05	0.7036	0.1046

Method	Tube toss						Bear thread					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	17.75	0.7547	0.1237	16.94	0.7071	0.0913	15.45	0.566	0.4656	14.89	0.3535	0.5652
4DGS	19.53	0.7429	0.127	17.72	0.6961	0.087	18.20	0.7811	0.2736	12.21	0.2426	0.5023
MotionGS	13.73	0.2469	0.3227	9.85	0.1433	0.3943	12.90	0.4995	0.494	9.83	0.1518	0.6414
NeRF-WT	31.63	0.9474	0.0315	19.44	0.8194	0.0704	22.13	0.8607	0.1618	13.55	0.2925	0.4098
HDR-Hexplane	17.13	0.4732	0.2211	16.43	0.6485	0.1096	22.08	0.7903	0.2597	18.09	0.5367	0.3357
Ours	32.08	0.9482	0.0348	24.78	0.9105	0.0366	30.19	0.9224	0.1337	23.93	0.7568	0.2448

Method	Tumbler						Dog					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	17.8	0.5645	0.2897	14.89	0.4618	0.3354	17.87	0.5403	0.2884	14.96	0.5193	0.3316
4DGS	22.05	0.8171	0.1753	17.54	0.6221	0.2124	27.15	0.9164	0.1069	19.42	0.7279	0.1661
MotionGS	15.41	0.5143	0.4434	12.16	0.3455	0.5881	15.26	0.4963	0.4309	11.73	0.3149	0.5667
NeRF-WT	31.97	0.9456	0.0625	20.48	0.7477	0.1733	31.7	0.9466	0.0719	20.15	0.7525	0.143
HDR-Hexplane	24.98	0.825	0.1828	22.24	0.7556	0.1909	25.02	0.8445	0.1905	20.46	0.7165	0.2318
Ours	34.92	0.9428	0.0727	27.85	0.8863	0.0927	33.14	0.944	0.0841	24.85	0.8422	0.1441

Method	Fire extinguisher						Laptop					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	21.04	0.8215	0.1300	26.5	0.6232	0.1742	23.17	0.8437	0.102	27.41	0.7219	0.1107
4DGS	22.73	0.8568	0.1324	23.94	0.6189	0.2659	23.86	0.8818	0.1023	28.34	0.8203	0.0957
MotionGS	18.34	0.6445	0.2728	20.56	0.4259	0.4292	18.64	0.6568	0.2378	20.8	0.4533	0.2908
NeRF-WT	32.78	0.9506	0.0496	24.39	0.6706	0.2079	37.4	0.9801	0.0197	29.26	0.9034	0.048
HDR-Hexplane	23.57	0.8876	0.0797	28.78	0.7612	0.080	23.51	0.8993	0.0721	29.9	0.8375	0.0611
Ours	36.82	0.9686	0.0298	32.04	0.8494	0.0777	37.28	0.9759	0.0231	33.32	0.9141	0.0407

Method	Bag						Ball					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	18.33	0.6867	0.2029	19.32	0.5594	0.1516	17.38	0.6033	0.2433	16.08	0.4525	0.2439
4DGS	19.54	0.7723	0.1744	18.13	0.6577	0.1706	21.01	0.8016	0.1534	16.17	0.5384	0.2431
MotionGS	14.59	0.4542	0.3284	12.43	0.3155	0.3460	13.95	0.3903	0.3366	10.47	0.2114	0.4195
NeRF-WT	29.9	0.9518	0.0488	22.34	0.7912	0.09	28.26	0.9453	0.0461	16.7	0.5584	0.1974
HDR-Hexplane	18.84	0.6643	0.2116	19.82	0.6863	0.1631	18.38	0.6246	0.2101	19.3	0.5772	0.1992
Ours	32.53	0.9532	0.0495	27.13	0.8937	0.0603	33.11	0.9482	0.0473	26.22	0.8431	0.0827

Table S2: **Quantitative results of novel view synthesis on GoPro dataset.** The green and yellow colors stand for the best and the second best, respectively.

C.4 Objective function

Finally, our HDR-NSFF is end-to-end optmized using the following loss:

$$\mathcal{L} = \mathcal{L}_{\text{cb}} + \mathcal{L}_{\text{photo}} + \beta_{\text{data}}\mathcal{L}_{\text{data}} + \beta_{\text{reg}}\mathcal{L}_{\text{reg}} + \beta_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \beta_{\text{gen}}\mathcal{L}_{\text{gen}}, \tag{28}$$

where the β coefficients weight each term. Additional regularization terms, $\mathcal{L}_{\text{reg}}$ leveraging scene flow priors.

D Additional experiment results

We provide additional experimental results that could not be included in the main manuscript, due to page limit. Specifically, Tables S2, S3, & S4 present per-scene quantitative results for each experiment. Figures S8–S19illustrates qualitative outcomes for additional real datasets not shown in

Method	Big jump						Side walk					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	17.09	0.6856	0.1858	15.05	0.4765	0.3175	17.48	0.7095	0.1621	15.49	0.609	0.2292
HDR-Hexplane	19.01	0.5076	0.2301	15.13	0.4166	0.2604	19.08	0.4914	0.2389	16.33	0.5778	0.1699
Ours	30.13	0.924	0.0662	22.09	0.756	0.1515	30.18	0.927	0.0564	21.03	0.7304	0.1413
Method	Jumping jack						Pointing walk					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	18.26	0.7873	0.1212	19.82	0.62	0.1448	18.21	0.7752	0.1286	15.56	0.6301	0.248
HDR-Hexplane	17.22	0.4626	0.2134	19.56	0.5833	0.1471	16.98	0.4383	0.2107	16.4	0.6171	0.1702
Ours	32.03	0.9457	0.0395	24.2	0.7954	0.0995	30.05	0.9346	0.0473	20.74	0.7194	0.1287
Method	Tube toss						Bear thread					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	18.5	0.7968	0.1095	18.4	0.7628	0.0967	16.91	0.5648	0.4884	15.96	0.3872	0.5795
HDR-Hexplane	17.15	0.4737	0.2204	16.35	0.6457	0.1096	21.81	0.785	0.2604	17.35	0.493	0.3452
Ours	32.27	0.9482	0.0378	25.19	0.9125	0.0498	30.4	0.9235	0.1437	24.29	0.7626	0.2869
Method	Dog						Tumbler					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	18.42	0.5972	0.3588	15.28	0.4708	0.3672	18.84	0.6352	0.2457	16.25	0.4823	0.3693
HDR-Hexplane	25.02	0.8438	0.1881	17.38	0.4989	0.2992	24.93	0.8247	0.1837	20.42	0.6685	0.2142
Ours	33.45	0.9452	0.0875	21.16	0.6793	0.2781	35.14	0.944	0.0752	26.12	0.8476	0.1415
Method	Fire extinguisher						Laptop					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	22.94	0.8639	0.1094	27.67	0.6782	0.1842	24.19	0.8908	0.0760	28.87	0.7801	0.0971
HDR-Hexplane	23.56	0.8854	0.0804	27.84	0.7416	0.0847	23.54	0.8992	0.0719	29.67	0.8359	0.0610
Ours	37.00	0.9694	0.0331	32.46	0.8568	0.0984	37.41	0.9763	0.0237	33.72	0.9187	0.0454
Method	Bag						Ball					
	Full			Dynamic only			Full			Dynamic only		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NSFF	19.33	0.7352	0.1815	20.51	0.6336	0.1439	17.95	0.6685	0.2042	17.25	0.5172	0.2595
HDR-Hexplane	18.85	0.6641	0.2116	19.76	0.6818	0.1627	18.36	0.6243	0.2104	18.91	0.5684	0.2014
Ours	32.26	0.9525	0.0525	26.91	0.8873	0.0748	32.64	0.9468	0.0499	25.25	0.8185	0.1105

Table S3: **Quantitative results of novel time synthesis on real data.** The green and yellow colors stand for the best and the second best, respectively.

Methods	synthetic dataset-Full				Methods	synthetic dataset-Dynamic			
	Lego	Mutant	Standup	Jumping Jack		Lego	Mutant	Standup	Jumping Jack
	PSNR					PSNR			
NSFF	15.45	16.97	13.47	15.53	NSFF	15.94	18.43	10.25	13.74
NeRF-WT	29.55	33.06	32.55	29.25	NeRF-WT	22.32	27.58	19.77	16.33
HDR-Hexplane	28.58	30.88	30.83	29.50	HDR-Hexplane	24.61	29.71	21.59	19.57
Ours	34.64	36.13	35.80	33.72	Ours	28.77	31.80	24.98	23.21
	SSIM					SSIM			
NSFF	0.6472	0.6348	0.4958	0.6551	NSFF	0.6145	0.5152	0.1601	0.5795
NeRF-WT	0.9595	0.9114	0.9556	0.9200	NeRF-WT	0.8517	0.8289	0.7741	0.5412
HDR-Hexplane	0.9443	0.8526	0.9112	0.9137	HDR-Hexplane	0.8626	0.8443	0.7665	0.7262
Ours	0.9670	0.9278	0.9564	0.9348	Ours	0.9062	0.9115	0.8816	0.8349
	LPIPS					LPIPS			
NSFF	0.1556	0.1243	0.2368	0.1364	NSFF	0.1528	0.1708	0.3097	0.1345
NeRF-WT	0.0171	0.0316	0.0224	0.0655	NeRF-WT	0.0592	0.0845	0.0988	0.1154
HDR-Hexplane	0.0257	0.0708	0.0603	0.0539	HDR-Hexplane	0.1217	0.0724	0.1547	0.0794
Ours	0.0147	0.0305	0.0249	0.1229	Ours	0.0426	0.0590	0.0749	0.0538

Table S4: **Quantitative results of novel view and time synthesis on synthetic dataset.** The green and yellow colors stand for the best and the second best, respectively.

the main paper. Moreover, supplementary videos include more HDR, LDR, and novel view rendering results. Please refer supplementary video for further visualization results.

D.1 Ablation Study

We analyze the impact of our proposed semantic-based optical flow on the novel view synthesis task using 8 real dataset samples. We compare two variants of our method: (1) Ours (w/ RAFT), in which the RAFT optical flow is used without modification, and (2) Ours (w/ RAFT Finetuned), where RAFT is fine-tuned on synthetic multi-exposure data. Note that, as shown in Figure 6, the original RAFT model was not trained on multi-exposed images, resulting in high errors when applied directly in our

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours w/ RAFT (Teed and Deng, 2020)	30.42	0.9269	0.0246	21.38	0.7369	0.0675
Ours w/ Finetuned	30.68	0.9234	0.0253	21.51	0.7377	0.0689
Ours w/ Dino-Tracker (Tumanyan et al., 2024)	31.01	0.9301	0.0233	22.55	0.7714	0.0697

Table S5: **Ablation study of flow model.** To compare the effect of flow regularization, we compare NVS performance of our approach against the baseline optical flow model (RAFT Teed and Deng (2020)) and a stronger baseline fine-tuned RAFT on a multi-exposure adaptation of the FlyingThings3D dataset.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MoSca w/ TM	26.67	0.920	0.060	15.96	0.463	0.064
MoSca w/ TM + DT	**29.35**	**0.937**	**0.040**	**21.22**	**0.751**	**0.064**

Table S6: **Quantitative results of MoSca with our proposed modules on GoPro outdoor scenes for HDR novel view synthesis.** MoSca with Tone-mapping (TM) aligns exposure-varied inputs into a consistent HDR radiance but still suffers from unreliable motion due to CoTracker's exposure sensitivity. Incorporating DINO-Tracker (DT) provides exposure-robust semantic flow, significantly improving both HDR reconstruction and dynamic motion stability.

setting. By fine-tuning it on synthetic data, the performance is improved. As shown in Table S5, our proposed method achieves the best results.

D.2 Extension to 3DGS-based dynamic reconstruction

While our method is developed on top of NSFF, its core components—tone-mapping and exposure-robust semantic flow via DINO-Tracker—are not tied to the NSFF framework. In this section, we demonstrate that our approach is broadly applicable and can be seamlessly integrated into 3D Gaussian Splatting (3DGS)–based dynamic reconstruction pipelines.

Applicability beyond NSFF. The primary objective of our design is to make dynamic 4D reconstruction robust to the exposure variance inherent in alternatively exposed monocular videos. To verify that the proposed components are method-agnostic, we extend our pipeline to a representative 3DGS-based method, MoSca (Lei et al., 2025). For this purpose, we integrate our modules into MoSca without modifying its core architecture. First, our learnable tone-mapping (TM) module transforms LDR frames acquired under varying exposures into a unified HDR radiance space, providing exposure-invariant appearance supervision throughout the optimization. In addition, we replace MoSca's original motion estimation, CoTracker (Karaev et al., 2024) with semantic flow obtained from DINO-Tracker(DT), which offers robust correspondence cues under severe illumination and exposure variations. With these modifications, we assess whether the proposed components can still enhance HDR dynamic reconstruction in a framework that depends on explicit Gaussian tracking rather than continuous scene-flow modeling.

Experimental setup. We run all experiments on the proposed GoPro outdoor dataset. For this ablation, we compare two variants of MoSca augmented with our components: MoSca + TM, which incorporates only our tone-mapping (TM) module, and MoSca + TM + DT, our full configuration equipped with both exposure-aware appearance normalization and robust semantic flow (DINO-Tracker, DT). We note that the original MoSca configuration with CoTracker-based motion cues consistently failed to converge under alternatively exposed inputs. Thus, we exclude it from comparison. We evaluate novel view synthesis (NVS) results and report quantitative metrics in addition to qualitative visualizations.

Quantitative results. Table S6 shows that applying only the tone-mapping module improves the HDR appearance reconstruction, as it successfully aligns exposure-varied images into a common radiometric domain. However, this configuration struggles to recover consistent motion: the CoTracker-based correspondences often fail under extreme exposure variations, leading to incorrect dynamic geometry.

Figure S4: **Qualitative results of MoSca with our proposed modules on GoPro outdoor scenes for HDR novel view synthesis.** MoSca with Tone-mapping module (TM) produces HDR-like appearances but suffers from temporal inconsistency and distorted dynamic geometry due to exposure-sensitive motion estimation. In contrast, MoSca with TM and DINO-Tracker (DT) leverages exposure-robust semantic flow, yielding stable geometry and photometrically consistent HDR novel views under challenging exposure alternation.

When both TM and DINO-Tracker are applied, the model achieves the best performance across all metrics. The semantic flow provides exposure-robust motion cues, enabling the 3DGS optimization to recover temporally stable Gaussian trajectories and coherent HDR appearance over time.

Qualitative results. Figure S4 shows qualitative results of MoSca based our approach on our GoPro Outdoor scenes. While MoSca+TM produces HDR-like frames, the reconstructed geometry becomes inconsistent across time due to unreliable motion supervision. In contrast, MoSca+TM+DT produces stable and photometrically consistent HDR novel views, successfully handling complex exposure alternation and large dynamic motions.

These results demonstrate that the improvements brought by our approach—our tone-mapping module and exposure-robust semantic flow—are not specific to NSFF. The same components substantially enhance a 3DGS-based method, improving both radiance reconstruction and motion stability. This indicates that the proposed modules provide general utility for 4D HDR reconstruction from alternatively exposed monocular videos, regardless of the underlying 3D representation.

D.3 2D-TO-4D HDR RECONSTURCTION

We compare our method against a two-stage baseline that first reconstructs an HDR video from alternatively exposed input frames using existing 2D HDR video methods—LAN-HDR (Chung and Cho, 2023), HDRFlow (Xu et al., 2024), and NECHDR (Cui et al., 2024)—and subsequently applies a dynamic 4D reconstruction framework. For each method, the predicted HDR frames are tone-mapped using a fixed μ-law operator, which compresses HDR radiance values E into the LDR domain through a logarithmic mapping:

$$M(E) = \frac{\log(1 + \mu E)}{\log(1 + \mu)}, \tag{29}$$

where $\mu = 500$ controls the compression strength. The tone-mapped frames are then provided as input to MoSca (Lei et al., 2025) for 4D reconstruction. For a fair comparison with the two-stage

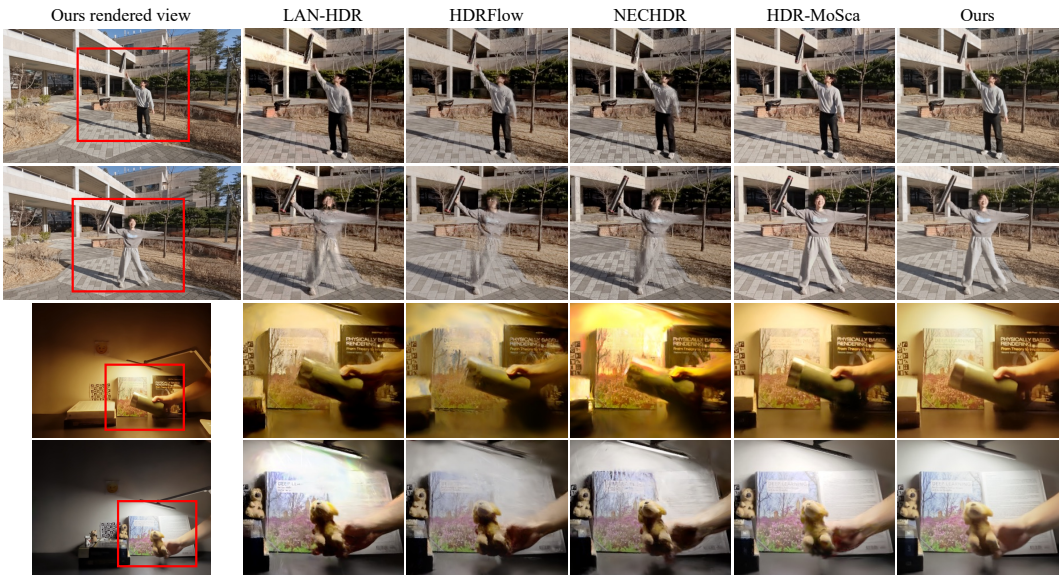


Figure S5: **Qualitative comparison with 2D-to-4D HDR reconstruction.**A two-stage baseline reconstructs HDR video using LAN-HDR (Chung and Cho, 2023), HDRFlow (Xu et al., 2024), or NECHDR (Cui et al., 2024), applies a fixed μ-law tone mapping, and performs 4D reconstruction with MoSca (Lei et al., 2025). For reference, we also include a MoSca-based variant of our method (HDR-MoSca). While 2D-to-4D baselines fail to recover, our approach yields temporally coherent HDR radiance and stable geometry.

2D-to-4D baseline, which uses MoSca for 4D reconstruction, we additionally report a MoSca-based variant of our pipeline (HDR-MoSca). This isolates the effect of the 2D HDR reconstruction stage from the choice of 4D representation.

While these 2D HDR video methods produce visually plausible results under mild motion, they fundamentally operate within the 2D image plane and therefore inherit the limitations outlined in Figure 2. In scenarios with noticeable camera motion they struggle to reliably handle occlusions, complex dynamics, and the severe exposure inconsistency intrinsic to alternatively exposed videos.

When such inconsistent HDR frames are used for 4D reconstruction, the downstream model receives observations that are radiometrically unstable and geometrically incoherent, preventing reliable estimation of density, radiance, and dynamic motion. Since the second stage has no mechanism to correct errors originating from the 2D reconstruction stage, these radiometric inconsistencies propagate forward and degrade the overall quality of the 4D reconstruction. Figure S5 shows that the two-stage 2D-to-4D baseline often produces temporally inconsistent radiance fields and fails to reconstruct stable geometry under exposure variation.

In contrast, our method performs end-to-end dynamic HDR radiance reconstruction directly from alternatively exposed inputs, jointly reasoning about radiance, geometry, and motion within a unified 4D representation. This formulation leverages geometric and motion priors unavailable to 2D methods, enabling consistent tone reproduction, recovery of valid information in saturated regions, and significantly more stable reconstruction under challenging exposure variations.

Figure S6: **Training view and ours reconstrcuted views on GoPro indoor dataset.** The odd-numbered rows show sample LDR frames from the input sequence, captured with alternating exposures (low, high, and mid). The even-numbered rows present our tone-mapped HDR reconstruction results for the corresponding input views. Under- and over-exposed regions are reliably recovered across all scenes. Even in areas where severe saturation is expected, such as the smile-emoji picture on the wall or the region directly beneath the light source, our method accurately reconstructs fine details. Even the mid-exposure frames contain locally saturated regions, yet these areas are consistently restored with high fidelity.

Figure S7: **Training view and ours reconstrcuted views on GoPro outdoor dataset.** The odd-numbered rows show sample LDR frames from the input sequence, captured with alternating exposures (low, high, and mid). The even-numbered rows present our tone-mapped HDR reconstruction results for the corresponding input views. Across the outdoor scenes, our method robustly reconstructs HDR radiance even under strong and diverse motion patterns. It performs reliably in scenarios involving large vertical motion in Big Jump, lateral motion in Pointing Walk, and fast object motion in Tube Toss. Despite these challenging dynamics, the model successfully restores both saturated regions caused by strong sunlight reflections on buildings and under-exposed regions cast in shadow, demonstrating consistent reconstruction quality across a wide range of exposure conditions.

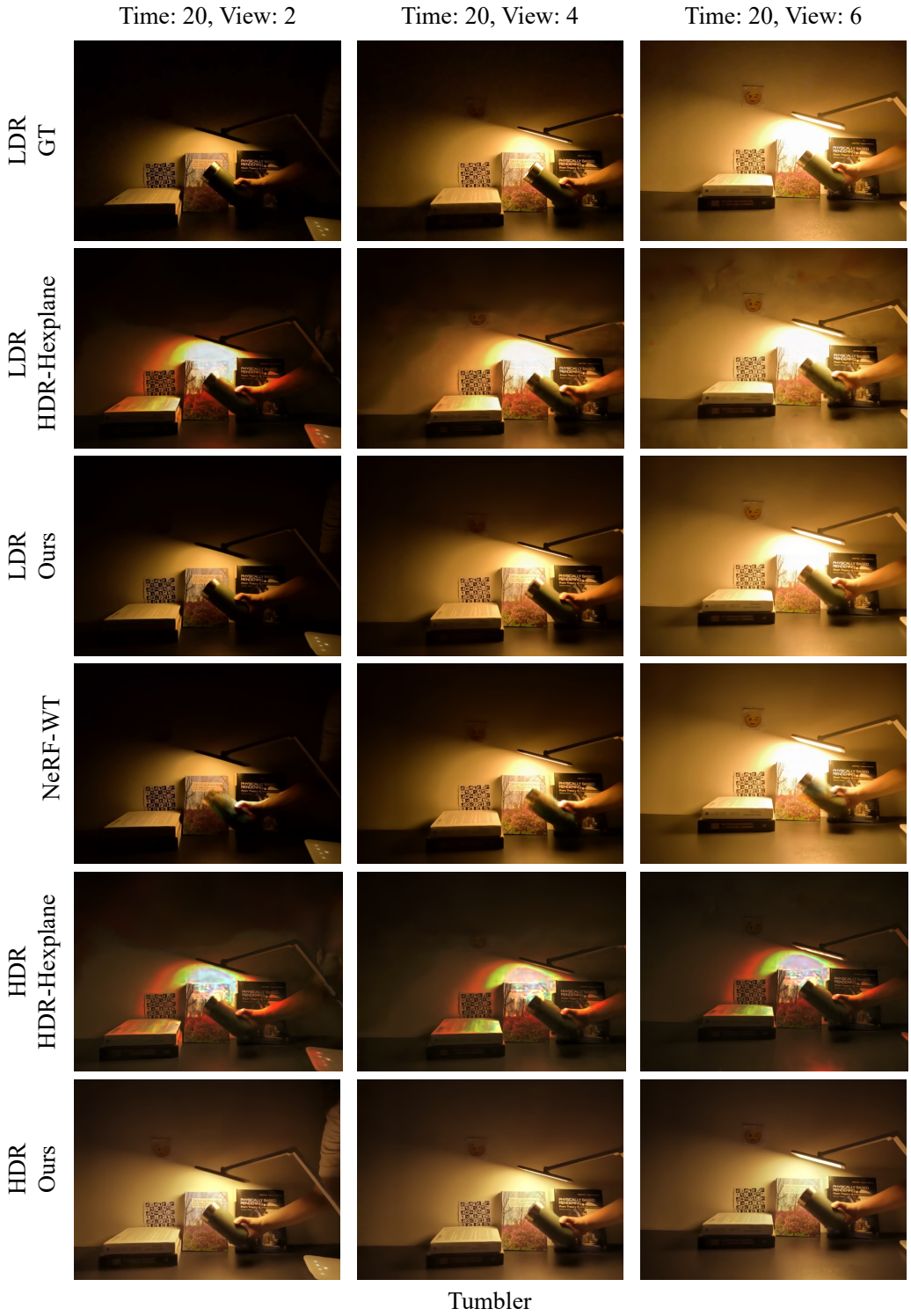


Figure S8: **Qualitative results of novel view synthesis on *Tumbler* data.**

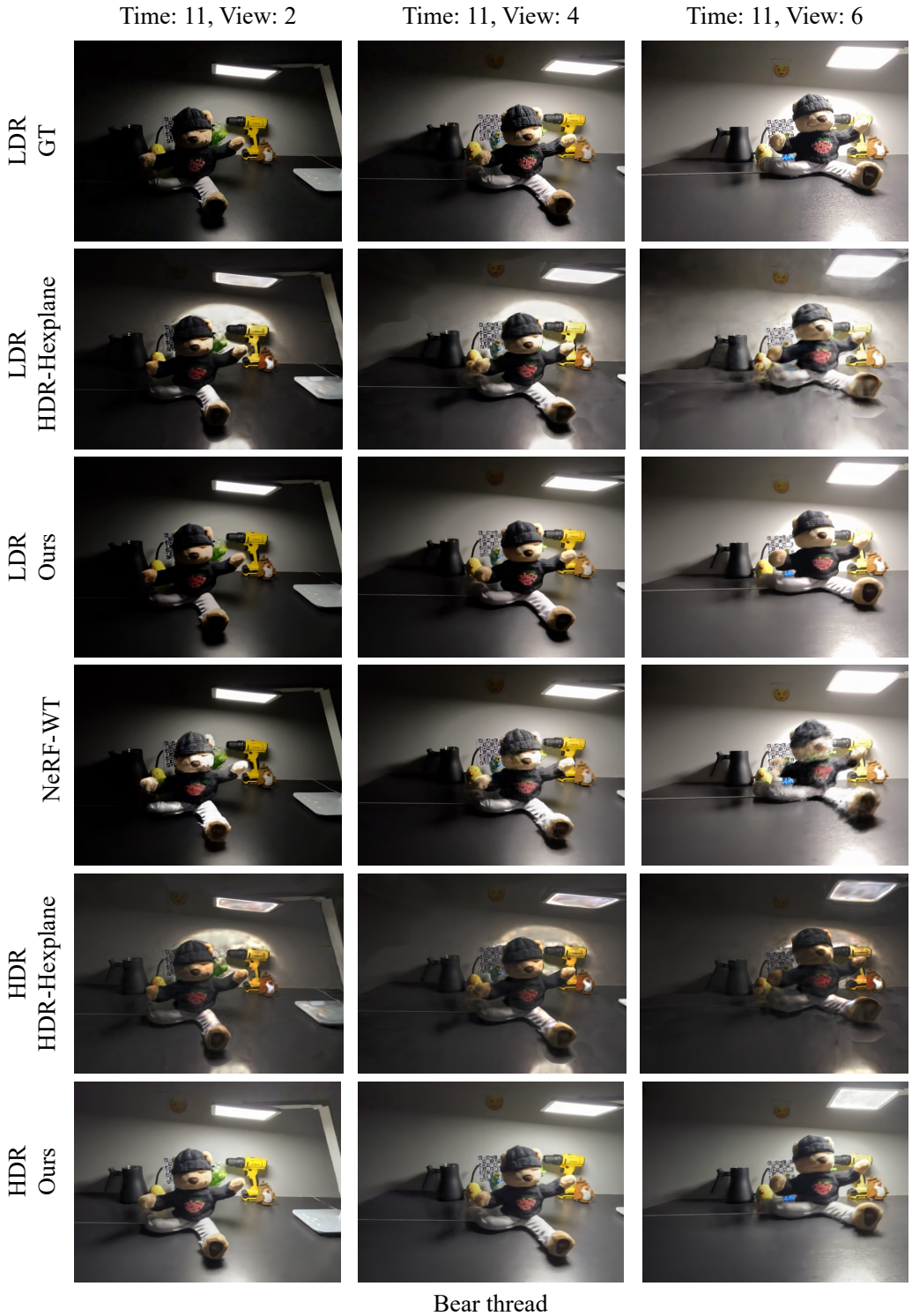


Figure S9: **Qualitative results of novel view synthesis on *Bear thread* data.**

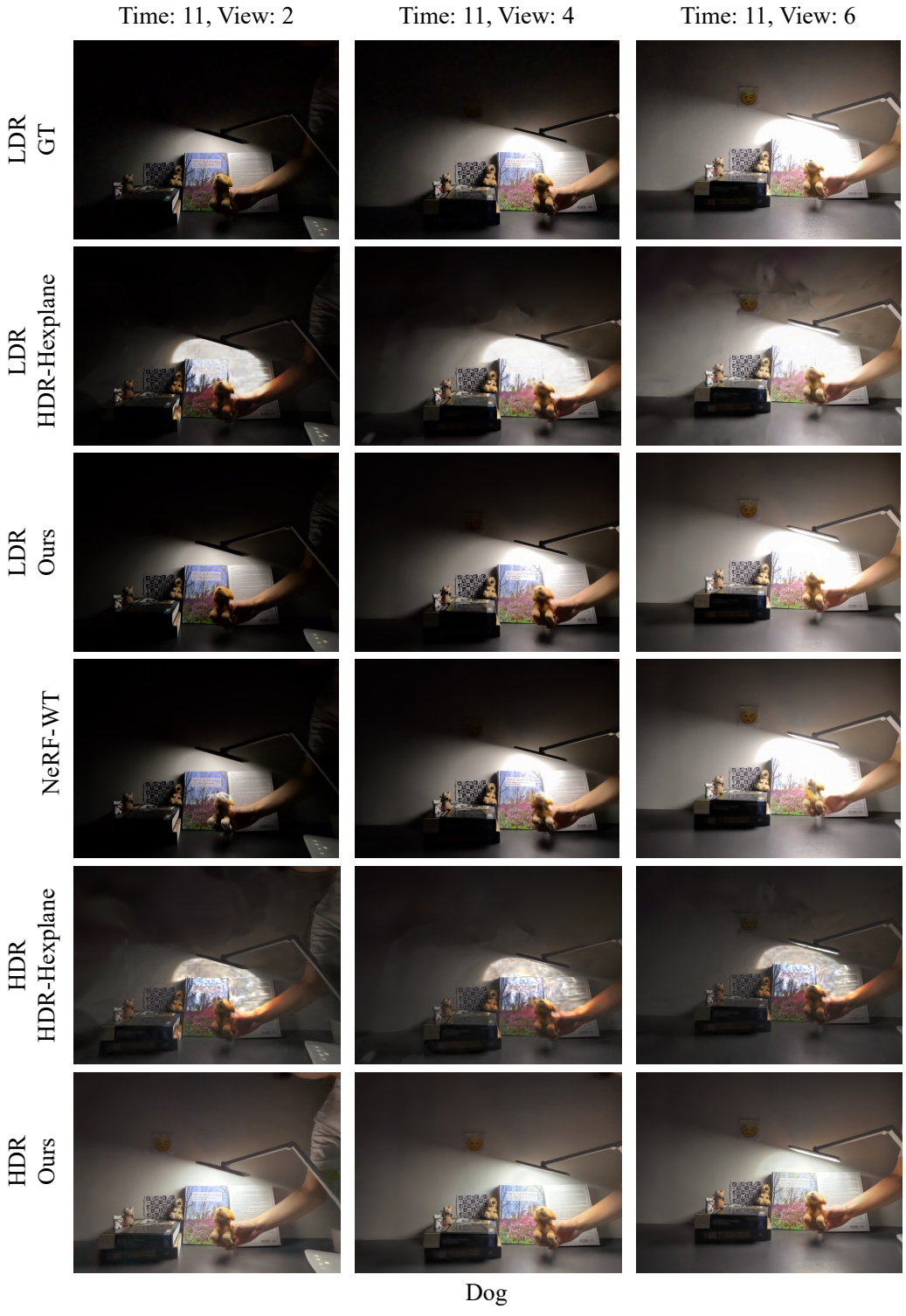


Figure S10: **Qualitative results of novel view synthesis on *Dog* data.**

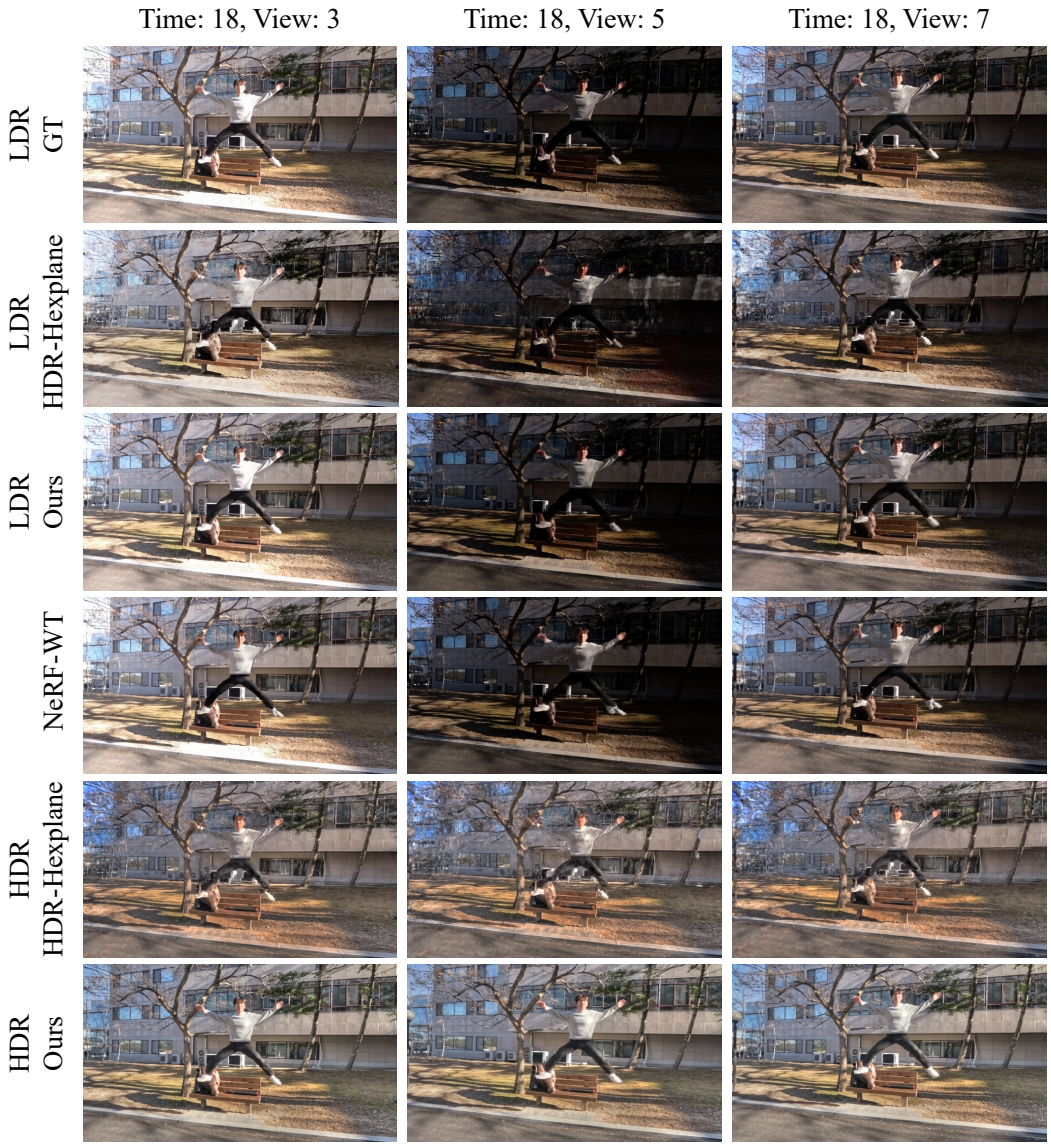


Figure S11: **Qualitative results of novel view synthesis on *Big jump* data.**

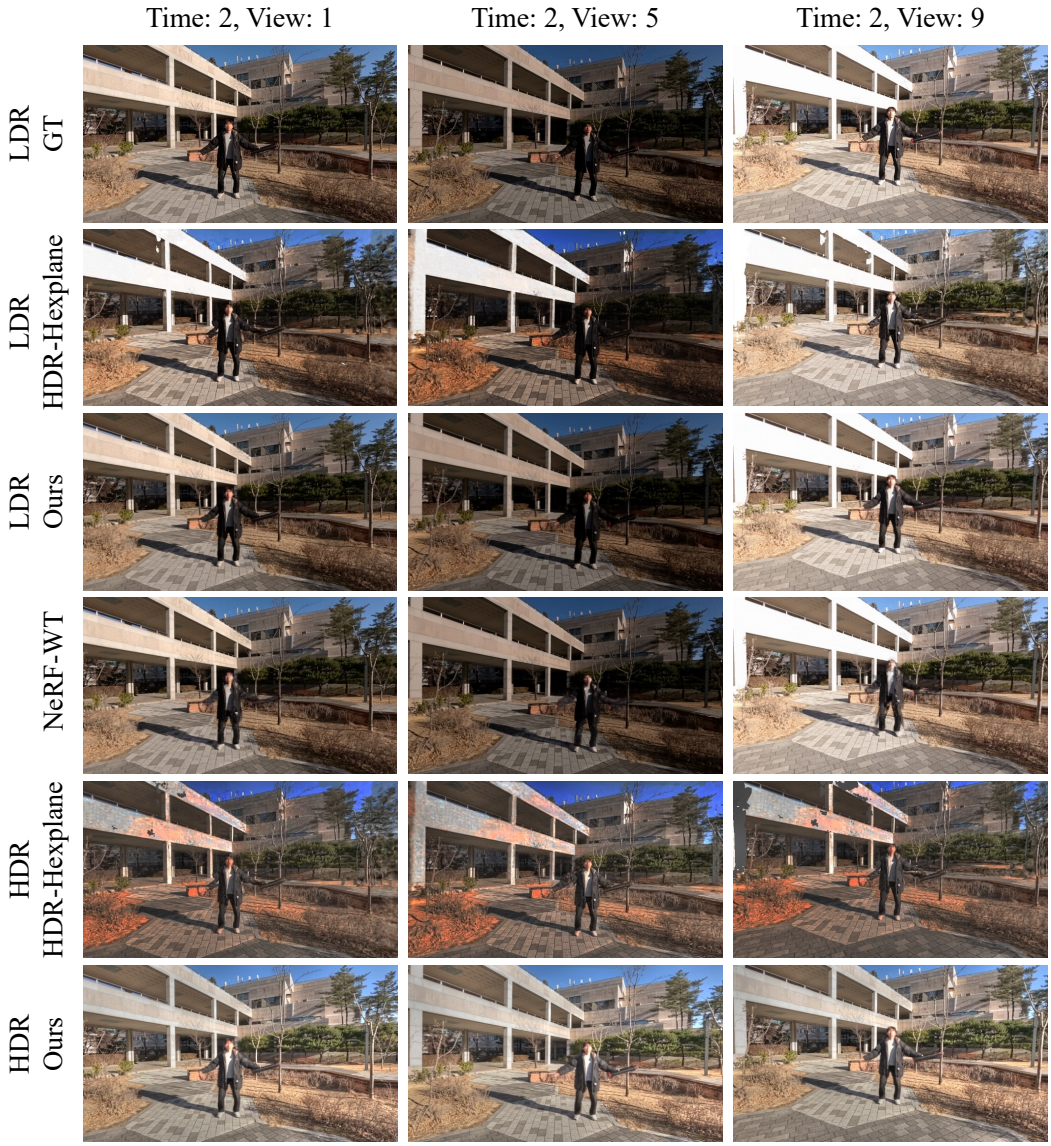


Figure S12: **Qualitative results of novel view synthesis on *Jumping jack* data.**

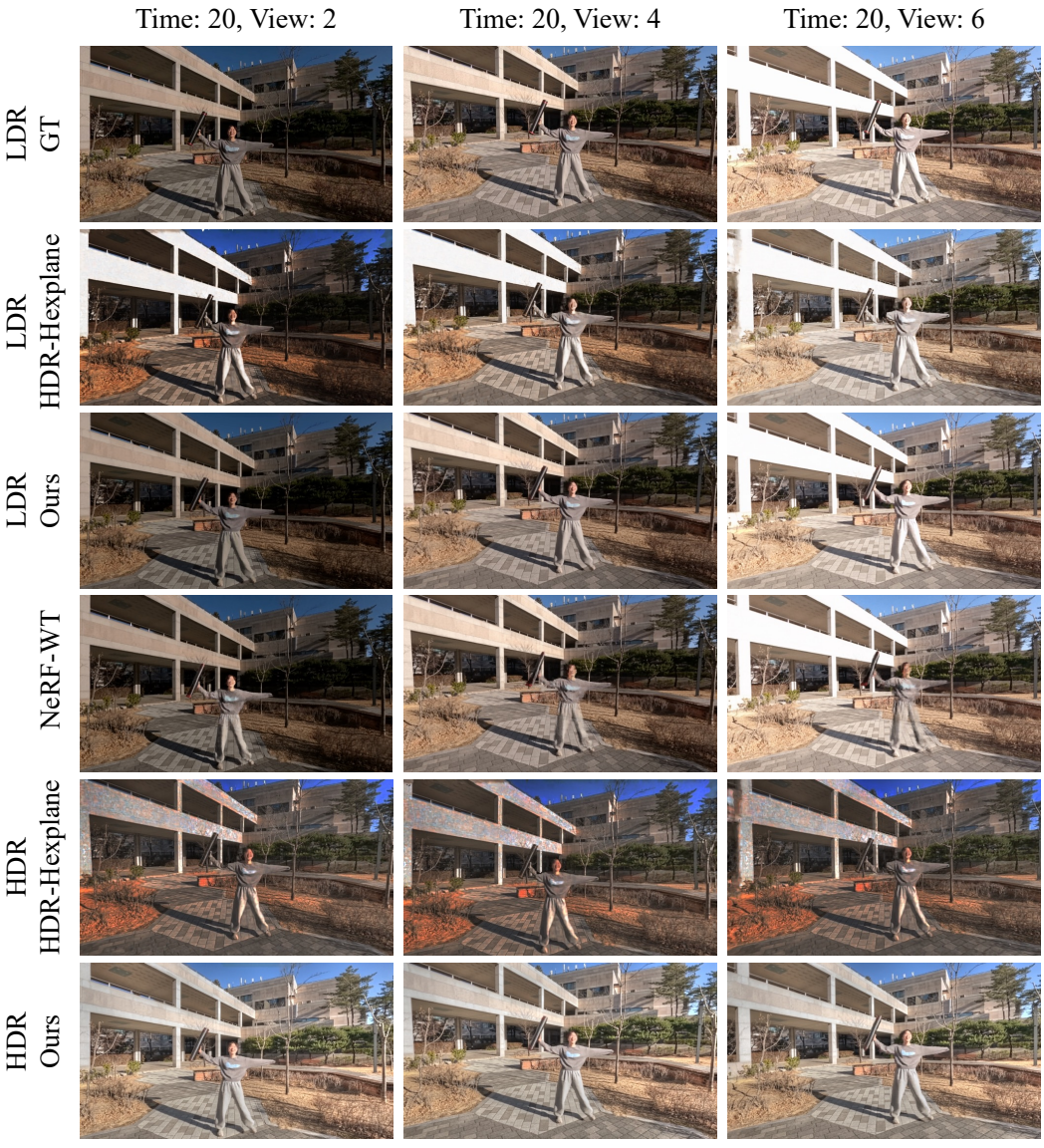


Figure S13: **Qualitative results of novel view synthesis on *Pointing walk* data.**

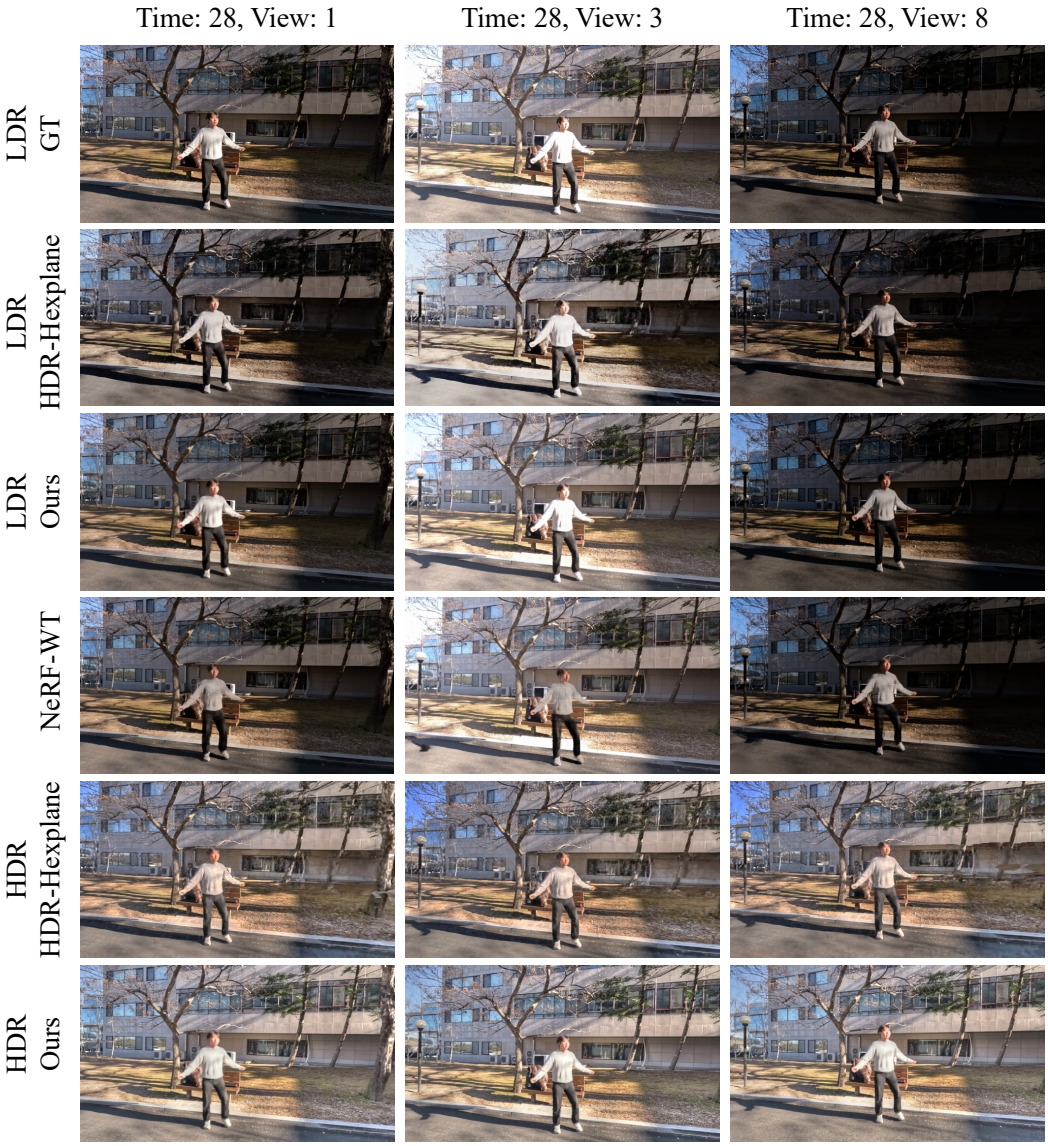


Figure S14: **Qualitative results of novel view synthesis on *Side walk* data.**

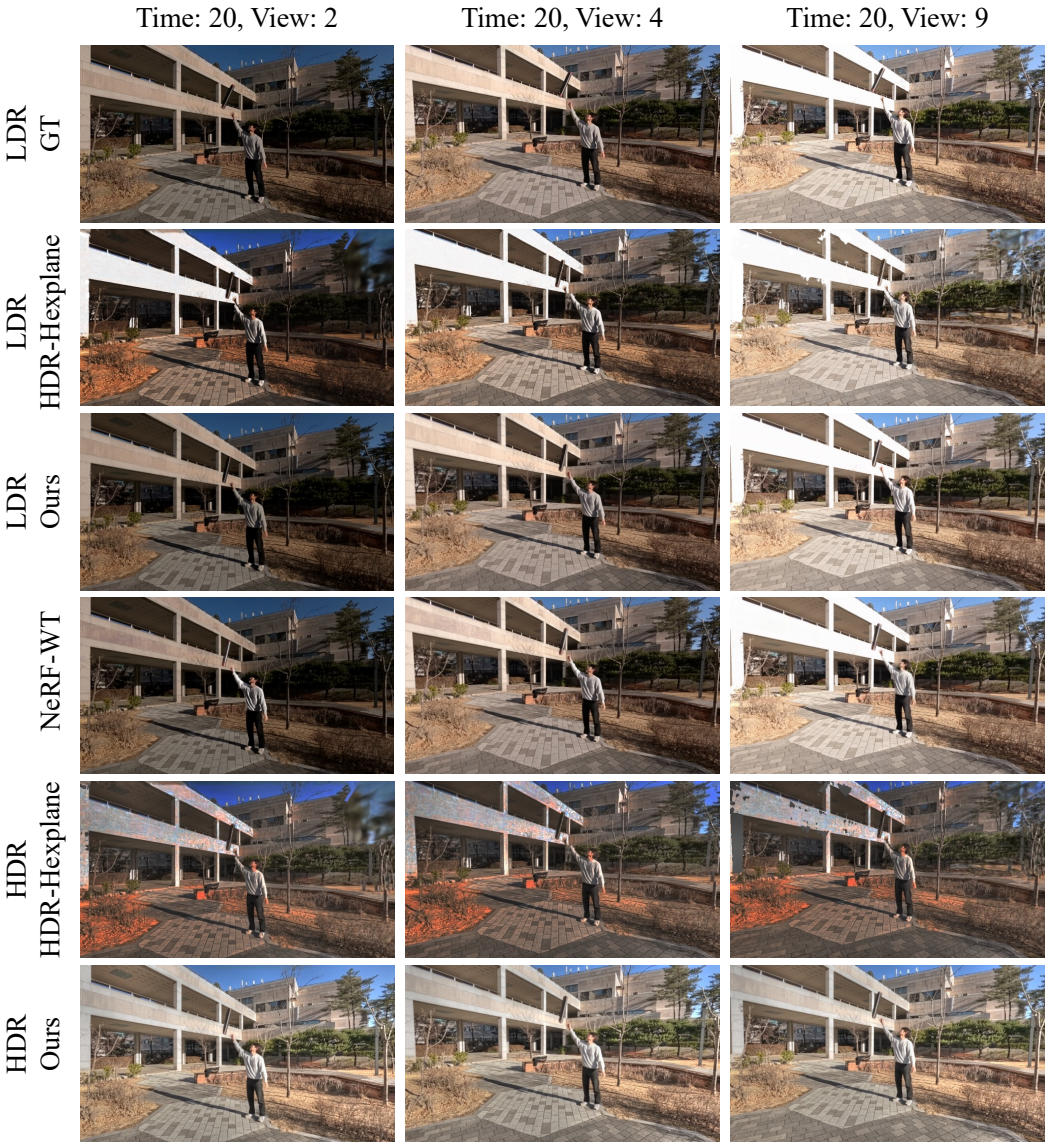


Figure S15: **Qualitative results of novel view synthesis on *Tube toss* data.**

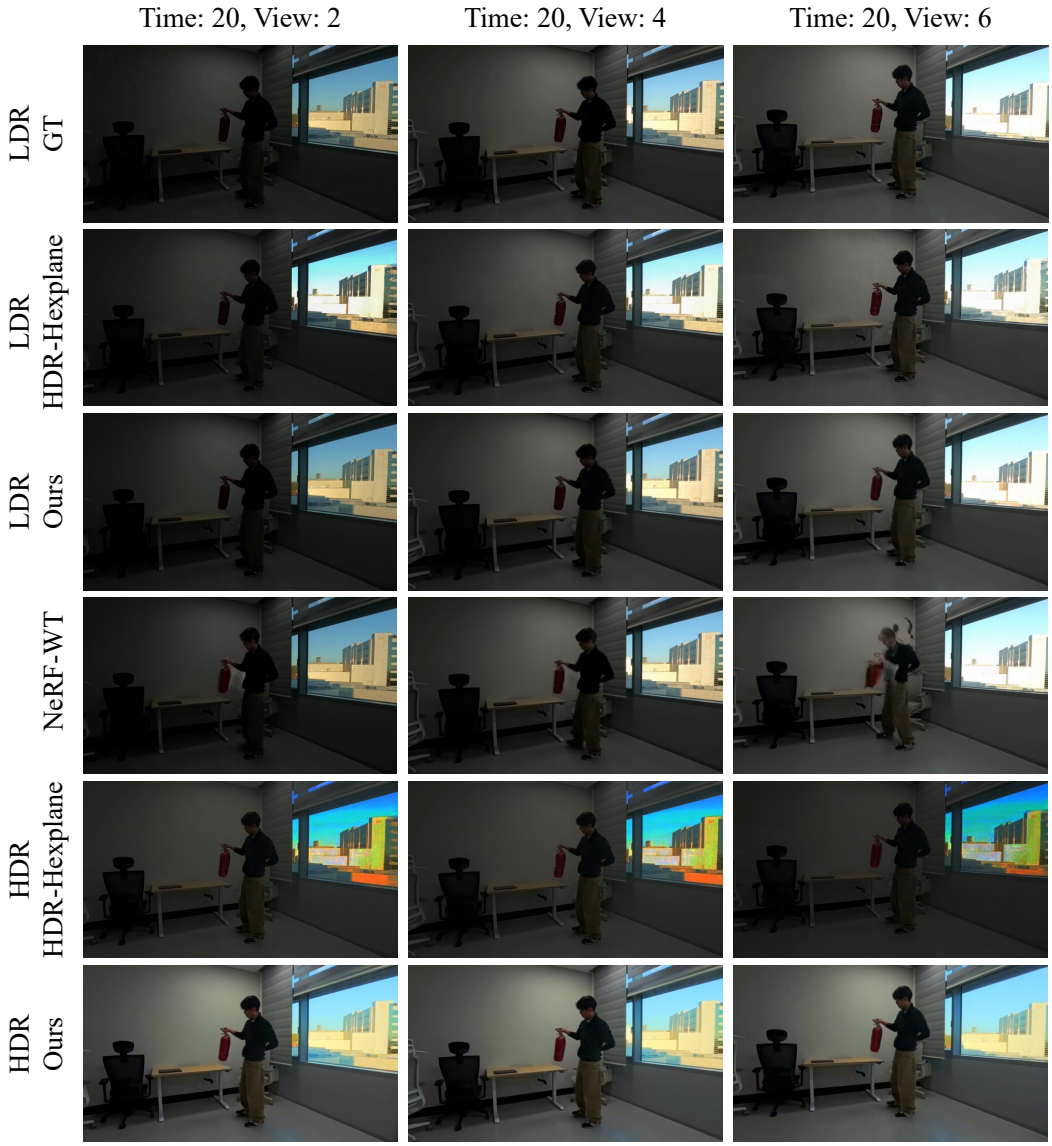


Figure S16: **Qualitative results of novel view synthesis on *Fire extinguisher* data.**

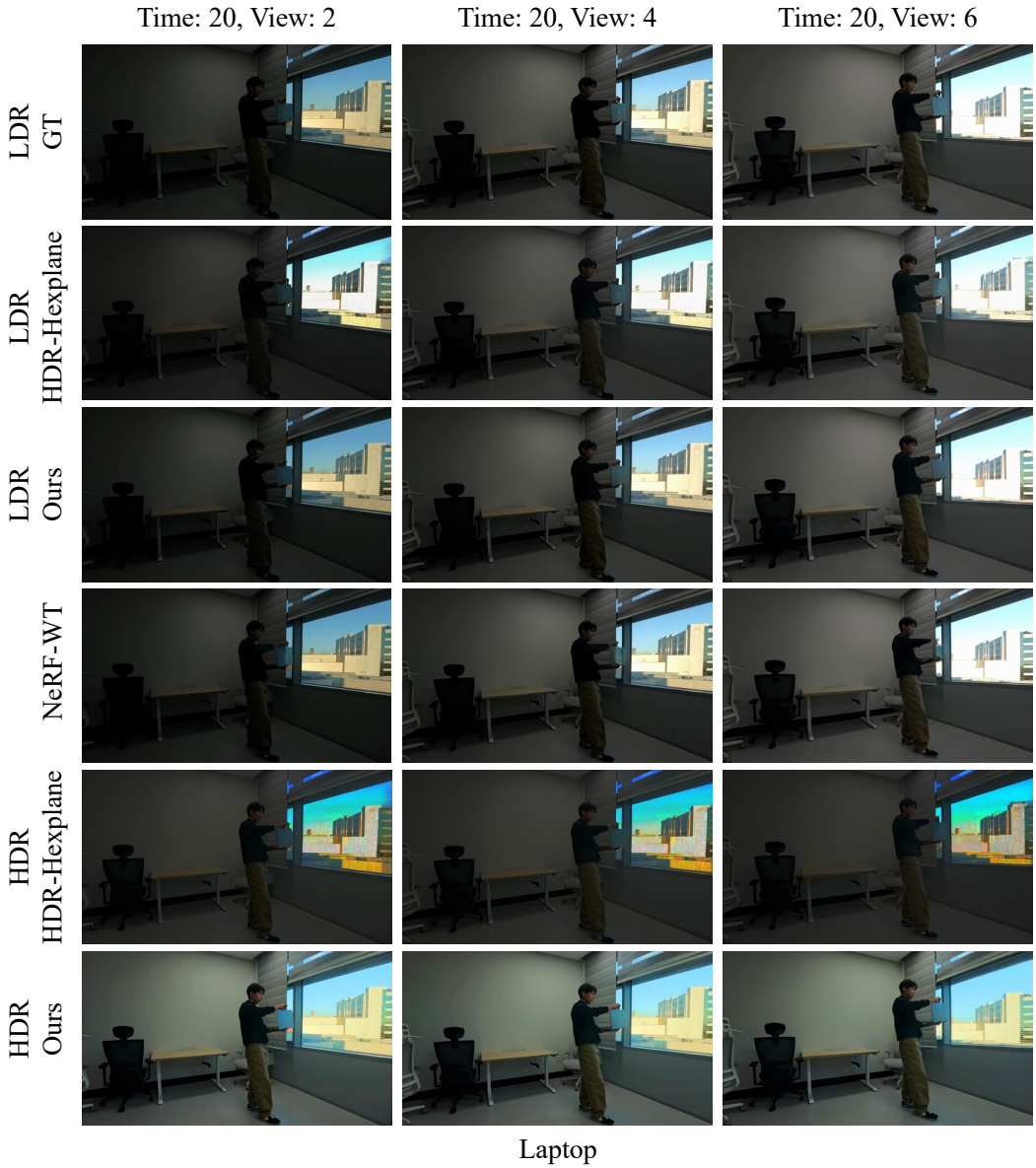


Figure S17: **Qualitative results of novel view synthesis on *Laptop* data.**

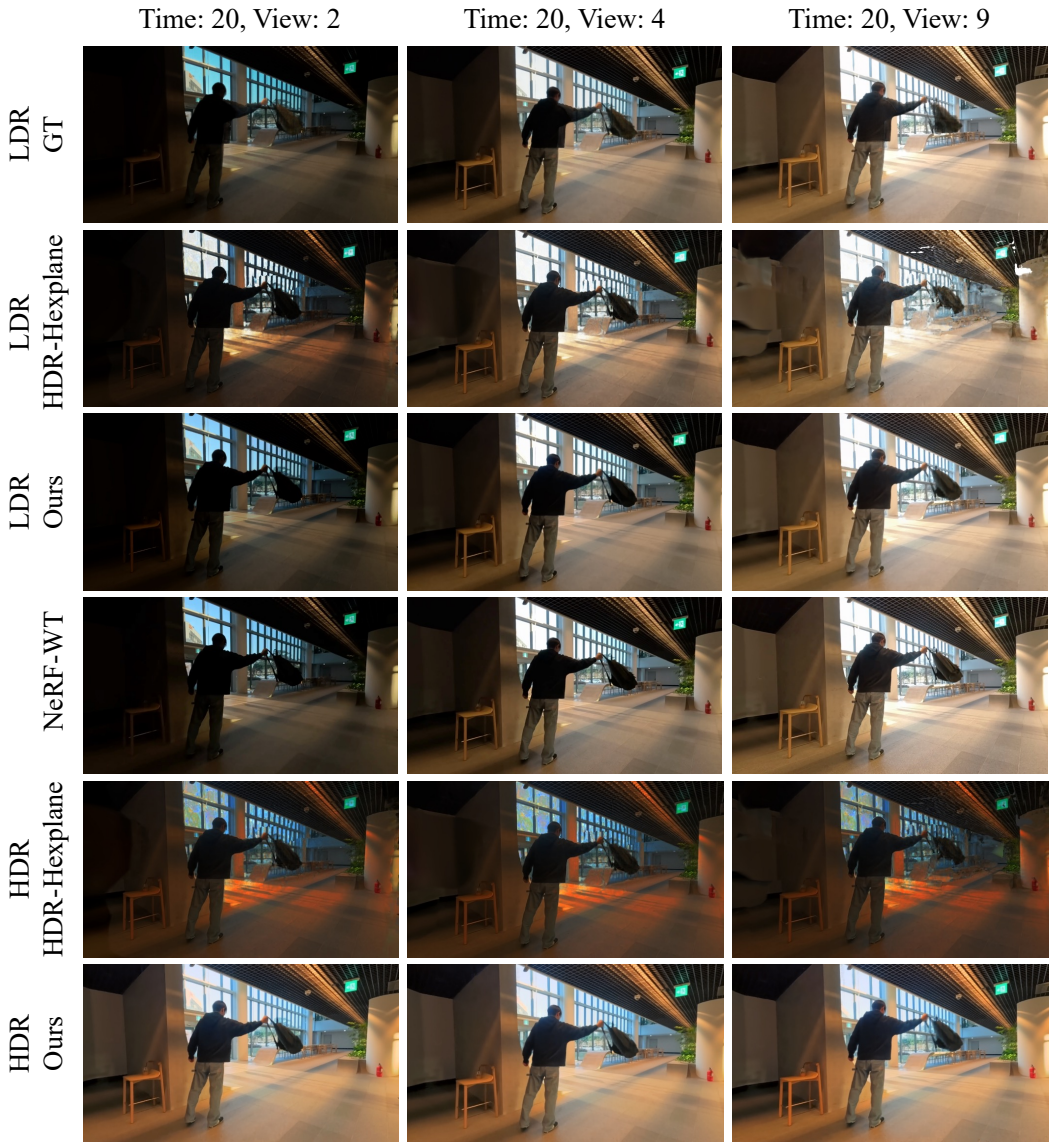


Figure S18: **Qualitative results of novel view synthesis on *Bag* data.**

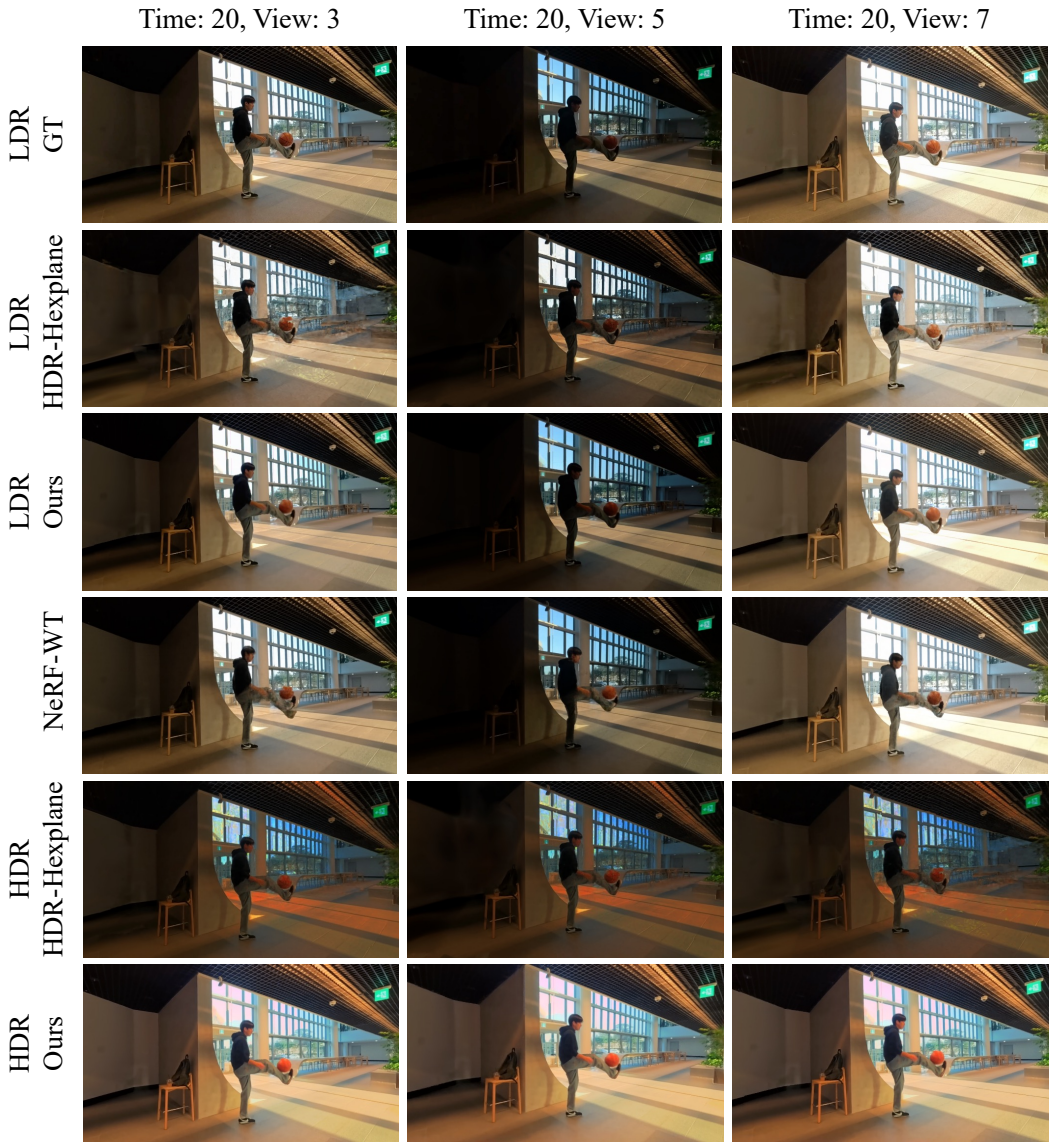


Figure S19: **Qualitative results of novel view synthesis on *Ball* data.**

USE OF LARGE LANGUAGE MODELS

A large language mode was used only for minor assistance in writing and improving the clarity of presentation.