MOTION FROM BLUR: INTRA-EXPOSURE TRAJECTORY RECOVERY FOR FEED-FORWARD 3D DEBLURRING

Anonymous authors
Paper under double-blind review

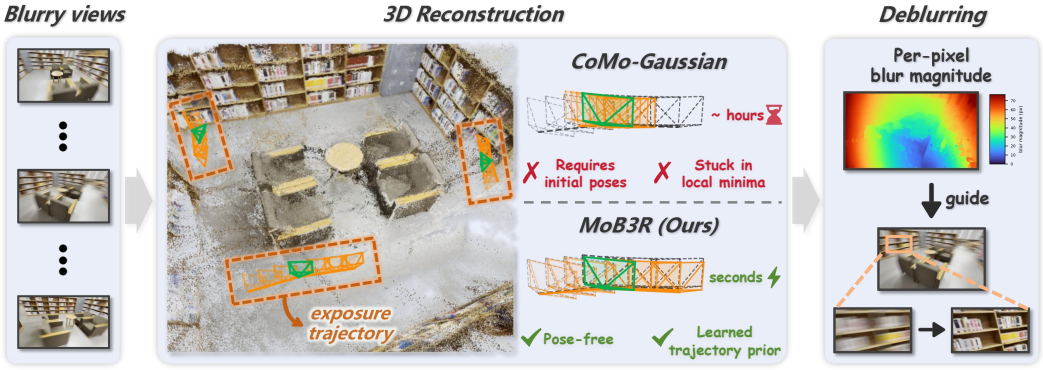

Figure 1: MOB3R reconstructs sharp 3D Gaussians from blurry views and recovers each view's exposure trajectory in a single forward pass. Middle: the estimated exposure trajectory of one input view, shown as an example. Initialized at the given midpoint pose, the per-scene method CoMo-Gaussian requires hours and gets stuck in a local minimum, whereas MOB3R accurately estimates the trajectory in seconds without initial poses, closely matching the ground truth. Right: the recovered trajectory yields a per-pixel blur magnitude map that guides sharp appearance reconstruction.

ABSTRACT

Recent feed-forward models estimate camera poses and scene geometry from multi-view images in a single forward pass, enabling fast 3D reconstruction. However, their performance degrades markedly on motion-blurred inputs. We find that their pose estimates exhibit a systematic bias aligned with the motion-blur direction, which persists even after fine-tuning on blurred data, revealing a mismatch between single-pose prediction and trajectory-based blur formation. We therefore propose MoB3R, a feed-forward framework that recovers the intra-exposure camera trajectory and uses it to reconstruct sharp 3D Gaussians. Rather than treating blur only as an artifact, MoB3R treats its spatial pattern as a signal of camera motion: the geometry stage recovers exposure trajectories with additional motion tokens, and the appearance stage derives a per-pixel motion field to guide sharp reconstruction through cross-attention, supervised by a cycle reblur loss. Experiments on synthetic and real blurred scenes show improved pose estimation and novel-view rendering across blur levels, while reconstructing in seconds rather than the hours required by per-scene optimization.

1 INTRODUCTION

Reconstructing a 3D scene representation from multi-view images for novel view synthesis is a long-standing problem in computer vision, with wide applications in robotics and content creation. Recently, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has achieved high-quality, real-time rendering by representing scenes with explicit 3D Gaussians. However, standard 3DGS assumes sharp input images. In handheld, drone, and driving scenarios, camera motion during exposure often produces motion blur. Each blurred image integrates scene appearance over its own expo-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

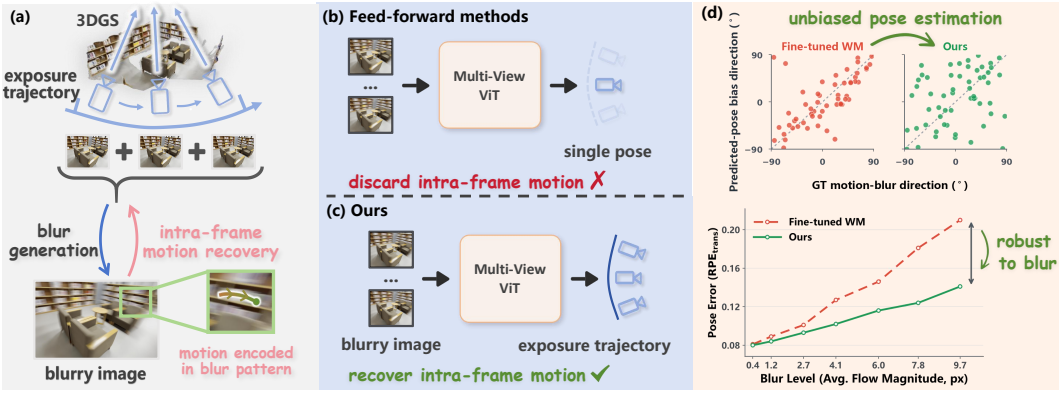

Figure 2: Motivation and overview. (a) A blurry image integrates sharp views along the camera trajectory during exposure, so its blur pattern encodes the intra-frame motion. (b) Existing feed-forward methods map each image to a single pose and thus discard intra-frame motion. (c) Our method recovers the exposure trajectory from blurry images. (d) Top: GT motion-blur direction versus the bias direction of globally aligned pose predictions; Fine-tuned WM exhibits a strong correlation between the two directions, which is largely removed by ours. Bottom: translational pose error across blur levels; our method maintains lower errors as blur increases. The analysis protocol is detailed in Sec. 3.2.

sure trajectory, weakening cross-view consistency. As a result, 3DGS absorbs blur into the scene representation, yielding distorted geometry and degraded novel-view rendering.

To address this issue, existing methods explicitly model the blur formation process and optimize a latent sharp scene for each scene. One line of work models blur in 2D pixel space with spatially varying kernels (Ma et al., 2022; Lee et al., 2023a; Peng & Chellappa, 2023; Peng et al., 2024), while another models it in 3D space through a physically based imaging model (Wang et al., 2023; Lee et al., 2023b; Oh et al., 2024; Zhao et al., 2024; Lee et al., 2025). This joint recovery is highly ill-posed, since a blurred observation can be explained by either a sharper scene with larger motion or a blurrier scene with smaller motion. These methods therefore rely on accurate initial poses from COLMAP (Schönberger & Frahm, 2016) and dense multi-view constraints. Under sparse views, however, the initial poses are inaccurate and the multi-view constraints are weak, while blur further flattens the photometric loss around the poses. As a result, the optimization often converges to a local minimum near the erroneous initialization, where blur is absorbed into the scene rather than attributed to camera motion (Fig. 1). Each scene also takes tens of minutes to hours to optimize.

Recently, feed-forward models (Wang et al., 2024; 2025a;b; Liu et al., 2025) have emerged as a promising alternative for 3D reconstruction. They regress camera poses and geometry in a single forward pass, complete reconstruction within seconds, and remain robust under sparse views. Extending this paradigm to blurred inputs is therefore attractive, yet their accuracy degrades sharply as blur increases, and direct fine-tuning on blurred data brings only limited improvement. We further conduct a quantitative analysis of how motion blur affects 3D reconstruction. As shown in Fig. 2(d), the predicted pose-bias direction of the fine-tuned model is strongly correlated with the motion-blur direction, while its pose error grows rapidly with blur magnitude. This persistent bias suggests that fine-tuning alone is insufficient. The underlying issue lies in the camera representation: feed-forward models predict a single instantaneous pose per image, whereas a blurred image integrates an exposure trajectory. Without an explicit representation of intra-frame motion, the model attributes part of the exposure motion to the single pose, producing a bias aligned with the motion direction.

Building on the above analysis, we identify two critical issues that must be addressed for accurate and robust feed-forward reconstruction from blurred images. (i) **Modeling intra-frame motion is essential.** This requires extending the camera output of feed-forward models from a single pose to an exposure trajectory while retaining single-pass inference. (ii) **Exploiting the recovered motion for appearance reconstruction is nontrivial.** Per-scene methods couple motion and appearance within a joint optimization, where each constrains the other. In contrast, feed-forward models follow a two-stage paradigm that first recovers geometry and then infers appearance, leaving the appearance stage

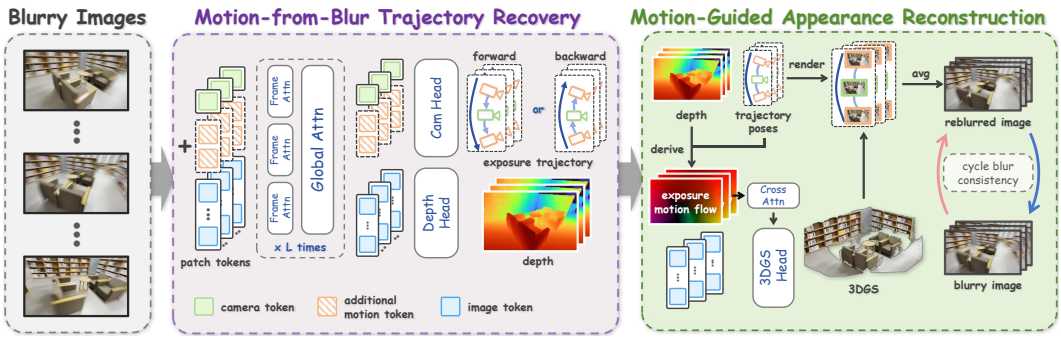

Figure 3: Overview of **MoB3R**. Given blurry multi-view images, the geometry stage recovers depth and the intra-exposure trajectory, while a reversal-invariant pose loss accounts for the two trajectory directions. The appearance stage derives an exposure motion field from the recovered trajectory and depth, and uses it to guide sharp Gaussian reconstruction through cross-attention. A reblur loss averages sharp renderings along the predicted trajectory to enforce deblur–reblur cycle consistency.

agnostic to the motion that produces the blur. The recovered motion must therefore be explicitly propagated to the appearance stage.

To this end, we propose **MoB3R**, a feed-forward framework that reconstructs sharp 3D Gaussians from blurred multi-view images. Our key insight is that instead of treating motion blur merely as an artifact to remove, we view its spatial pattern as a signal encoding camera motion during exposure (Fig. 2(a)). MoB3R follows a two-stage pipeline: the geometry stage recovers scene geometry and intra-exposure trajectories, while the appearance stage uses the recovered motion to reconstruct sharp appearance. For geometry reconstruction, additional motion tokens initialized from the pre-trained camera token extend single-pose prediction to trajectory recovery (Fig. 2(c)). We train geometry progressively, from single-frame trajectory learning to multi-view geometry training. To handle the forward–backward ambiguity, we define a reversal-invariant loss. For appearance reconstruction, we derive a per-pixel image-space motion field from the recovered trajectory and depth, whose magnitude reflects the local blur level. We exploit this field to guide sharp appearance recovery through cross-attention in the Gaussian branch. Finally, a reblur loss further enforces deblur–reblur cycle consistency.

Our contributions are summarized as follows:

- We identify a motion-aligned pose bias arising from the mismatch between single-pose prediction and trajectory-based blur formation, and show that recovering the intra-exposure trajectory substantially weakens this bias.
- We propose **MoB3R**, which learns a generalizable trajectory prior from data to infer intra-exposure motion in a single forward pass. Unlike per-scene methods that fit motion to each scene, MoB3R requires neither initial poses nor dense views, avoiding the local minima of per-scene optimization under sparse views. The recovered motion is further propagated to the appearance stage as a per-pixel motion field, bridging the gap left by the two-stage feed-forward paradigm.
- Experiments on synthetic and real blurred scenes demonstrate improved pose estimation and novel-view rendering, and show that the inferred trajectories are substantially more accurate than those optimized per scene under sparse views, while reconstruction takes seconds rather than hours.

2 Related Work

Feed-forward 3D reconstruction. Recent feed-forward methods infer scene geometry and camera poses directly from unposed images, extending earlier posed-view depth models (Yao et al., 2018). DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) regress dense 3D structure from image pairs, while multi-view models such as VGGT (Wang et al., 2025a), MUSt3R (Cabon

et al., 2025), Fast3R (Yang et al., 2025), π^3 (Wang et al., 2025b), Depth Anything 3 (Lin et al., 2025), WorldMirror (Liu et al., 2025), Scal3R (Xie et al., 2026), VGGT-Ω (Wang et al., 2026), and VGG-T^3 (Elflein et al., 2026) improve scalability and generalization. Feed-forward Gaussian methods further predict renderable scene representations from sparse, unposed views (Charatan et al., 2024; Chen et al., 2024; Wewer et al., 2024; Xu et al., 2025a; Smart et al., 2024; Ye et al., 2025; Hong et al., 2025; Zhang et al., 2025). Although efficient, these methods assume sharp observations and assign a single camera pose to each image. MoB3R instead extends the camera representation to an intra-exposure trajectory for reconstruction from blurred inputs.

3D reconstruction from blurred images. Single-image deblurring networks restore each frame independently and do not model multi-view geometry or camera motion (Nah et al., 2017; Kupyn et al., 2018; Tsai et al., 2022; Zamir et al., 2022; Chen et al., 2022). Methods for blurred 3D reconstruction instead incorporate blur formation into NeRF (Mildenhall et al., 2021) or 3DGS (Kerbl et al., 2023; Yu et al., 2024; Huang et al., 2024) rendering. Kernel- and prior-based approaches combine latent sharp rays using spatially varying blur kernels or sharpness priors (Ma et al., 2022; Lee et al., 2023a; Peng & Chellappa, 2023; Peng et al., 2024; Lee et al., 2024), whereas trajectory-based approaches average sharp renderings along the camera trajectory during exposure (Wang et al., 2023; Lee et al., 2023b; Oh et al., 2024; Zhao et al., 2024; Lee et al., 2025). BARD-GS (Lu et al., 2025) further extends blur-aware reconstruction to dynamic scenes. These methods optimize the sharp scene and blur formation per scene, incurring high computational cost and relying on sufficient multi-view constraints. Recent methods explore learned diffusion priors for 3D deblurring. CoherentGS (Xu et al., 2025b) uses image diffusion to correct artifacts in 3DGS renderings, but still relies on slow per-scene optimization. DeblurNVS (Shi et al., 2026) combines feed-forward reconstruction with multi-view geometric latent diffusion, but does not explicitly model motion during exposure. In contrast, MoB3R recovers the intra-exposure trajectory in a single forward pass and uses the recovered motion to guide sharp Gaussian reconstruction.

3 METHOD

3.1 PROBLEM FORMULATION

Given N motion-blurred images $\mathcal{B} = \{B_i\}_{i=1}^{N}$ of a static scene, our goal is to recover a sharp 3D Gaussian representation $\mathcal{G}$ (Kerbl et al., 2023) without known camera poses. We approximate each blurred image by averaging sharp renderings at S camera poses $\{T_i^j\}_{j=1}^{S}$ sampled during its exposure, $B_i \approx \frac{1}{S}\sum_{j=1}^{S}\mathcal{R}(\mathcal{G}, T_i^j)$, where $\mathcal{R}$ denotes Gaussian splatting. Standard feed-forward models assign one pose to each image and cannot represent this trajectory. MoB3R instead predicts a sharp representation $\widehat{\mathcal{G}}$ and all exposure trajectories directly from $\mathcal{B}$ in a single forward pass. Camera intrinsics are predicted by the camera head, remain fixed during exposure, and are omitted from our notation.

3.2 PRELIMINARY ANALYSIS: MOTION-ALIGNED POSE BIAS

Before designing our model, we first ask how exposure motion affects the pose predicted from a blurred image. We fit a vanilla 3DGS to the sharp views of each scene in the DL3DV (Ling et al., 2024) test set and synthesize blurred frames by averaging sharp renderings along manually specified exposure trajectories, so that only the exposure motion changes. We apply fine-tuned WorldMirror (Liu et al., 2025) to these frames, globally align its predictions to the sharp reference cameras, and convert both the ground-truth exposure motion and the aligned pose residual into image-space flow to compare their dominant directions. The full protocol is provided in Appendix D.1.

As shown in Fig. 2(d), the pose error grows rapidly as blur increases, while the pose-bias direction remains strongly correlated with the blur direction. Thus, blur does not merely increase pose error at random; it introduces a bias aligned with the exposure motion. This result indicates that a single-pose model assigns part of the unmodeled motion to its pose prediction. It motivates MoB3R to recover the exposure trajectory explicitly and use the recovered motion to reconstruct sharp appearance.

3.3 METHOD OVERVIEW

MoB3R reconstructs sharp 3D Gaussians from blurry multi-view images through two stages, as shown in Fig. 3. We adopt WorldMirror as the base model, which follows VGGT (Wang et al., 2025a). For each image B_i, the image encoder (Dosovitskiy et al., 2021) produces patch tokens t_i^{img} and one camera token t_i^{cam}, and an alternating-attention transformer applies frame-wise and cross-view self-attention to them. From the output tokens, a camera head predicts the pose $\widehat{T}_i$, while three DPT heads (Ranftl et al., 2021) predict the depth $\widehat{D}_i$, the point map $\widehat{P}_i$, and the Gaussian depth and features $(\widehat{D}_i^g, \widehat{\mathbf{F}}_i^g)$, denoted by DPT_d, DPT_p, and DPT_g. MoB3R extends the single-pose output to an exposure trajectory (Sec. 3.3.1), then derives image-space motion from the trajectory and depth to condition sharp Gaussian reconstruction (Sec. 3.3.2). It adds only motion tokens, a motion-field encoder, and cross-attention to the base model.

3.3.1 MOTION-FROM-BLUR TRAJECTORY RECOVERY

Our goal is to recover the camera trajectory during each frame's exposure while preserving the pretrained single-pose prediction. We therefore retain the original camera output as the exposure midpoint and extend it with two endpoint predictions.

Motion tokens. Following VGGT, the first frame uses a dedicated learnable camera token $\bar{t}^{\mathrm{cam}}$, while all remaining frames use a shared learnable camera token t^{cam}. After attention, these tokens become view-specific outputs. We retain each view's camera token as the midpoint token, denoted by $t_i^m \equiv t_i^{\mathrm{cam}}$, and add two endpoint motion tokens t_i^s and t_i^e. Both endpoint tokens are initialized as the corresponding camera token plus small independent perturbations, preserving its pose-aware representation while breaking the endpoint symmetry. After the alternating-attention transformer, their output tokens are denoted by $\widehat{t}_i^s$, $\widehat{t}_i^m$, and $\widehat{t}_i^e$ and share the same pretrained camera head, $(\widehat{T}_i^s, \widehat{T}_i^m, \widehat{T}_i^e) = \mathrm{CameraHead}(\widehat{t}_i^s, \widehat{t}_i^m, \widehat{t}_i^e)$, where all outputs are absolute camera poses.

Exposure trajectory. We assume constant camera motion within each half of the exposure and uniformly sample the trajectory from the predicted endpoints through the midpoint. We use $S = 7$ samples, with the fourth sample fixed to the midpoint pose $\widehat{T}_i^4 = \widehat{T}_i^m$, giving the trajectory $\widehat{\mathcal{T}}_i = \{\widehat{T}_i^j\}_{j=1}^S = \mathrm{Interpolate}_{\mathrm{SLERP}}(\widehat{T}_i^s, \widehat{T}_i^m, \widehat{T}_i^e)$. Here, $\mathrm{Interpolate}_{\mathrm{SLERP}}$ uses quaternion SLERP for rotation and linear interpolation for translation, as in BAD-Gaussians (Zhao et al., 2024). This representation retains the pretrained midpoint camera estimate while extending it to a continuous exposure trajectory.

Reversal-invariant pose loss. For a single blurred image, exposure integrates renderings along the camera trajectory but does not preserve their temporal order. Reversing the trajectory only exchanges the start and end poses; it leaves the sampled poses, and hence the averaged blurred rendering, unchanged for a static scene. The image therefore cannot distinguish the forward and reverse directions of the same intra-exposure motion. To avoid penalizing either valid ordering, we evaluate the endpoint pose loss under both assignments and retain the lower value, $\mathcal{L}_{\mathrm{pose}} = \mathcal{L}_{\mathrm{mid}} + \min(\mathcal{L}_{\mathrm{fwd}}, \mathcal{L}_{\mathrm{rev}})$, where $\mathcal{L}_{\mathrm{fwd}}$ matches start to start and end to end, while $\mathcal{L}_{\mathrm{rev}}$ swaps the two endpoints. Here, $\mathcal{L}_{\mathrm{mid}}$ is the standard camera loss on the midpoint pose. The midpoint and endpoint terms follow the camera loss of VGGT and WorldMirror; endpoint supervision is applied only to rotation and translation. We retain the original depth and point-map losses for geometry learning, with the weights used by WorldMirror.

3.3.2 MOTION-GUIDED APPEARANCE RECONSTRUCTION

Exposure motion field. The recovered trajectory describes camera motion in 3D, whereas blur varies across image locations. We project the trajectory into each image using the depth prediction $\widehat{D}_i$ from DPT_d. Let π_i^k denote projection under camera $\widehat{T}_i^k$. For pixel $\mathbf{p}$, the flow from the midpoint to endpoint $k \in \{s, e\}$ is

$$\mathbf{f}_i^k(\mathbf{p}) = \pi_i^k\Big((\pi_i^m)^{-1}\big(\mathbf{p}, \widehat{D}_i(\mathbf{p})\big)\Big) - \mathbf{p}, \qquad k \in \{s, e\}, \tag{1}$$

where $(\pi_i^m)^{-1}$ back-projects pixel $\mathbf{p}$ with its predicted midpoint depth, and π_i^k projects the resulting 3D point into endpoint camera k. We concatenate the two flows into a four-channel motion field $M_i = [\mathbf{f}_i^s, \mathbf{f}_i^e]$, which records the direction and extent of local exposure motion on both sides of the midpoint.

Motion-guided features. The WorldMirror Gaussian branch combines the feature from DPT_g, $F_i^{\mathrm{DPT}} \equiv \widehat{\mathbf{F}}_i^g$, with an image feature from a shallow appearance branch. We denote the fused feature by $F_i = F_i^{\mathrm{DPT}} + F_i^{\mathrm{img}}$. We patchify F_i into Gaussian tokens $Z_i = P(F_i)$ and encode the motion field into aligned motion tokens $U_i = E_m(M_i)$. Cross-attention (Vaswani et al., 2017) is performed within each frame, using Gaussian tokens as queries and motion tokens as keys and values: $\widetilde{Z}_i = Z_i + \mathrm{CA}(Z_i, U_i)$, where P is the patchification operator, E_m is the motion-field encoder, and CA denotes cross-attention. The residual connection preserves the pretrained feature while adding local motion information. The unpatchified features are decoded by the Gaussian attribute decoder to produce sharp 3D Gaussians.

Method	4 Views						12 Views						32 Views					
	Deblur			NVS			Deblur			NVS			Deblur			NVS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Per-scene optimization																		
BAD-Gaussians (Zhao et al., 2024)	19.52	0.3284	0.2243	18.37	0.2891	0.2478	21.83	0.4652	0.1581	21.14	0.4387	0.1816	23.71	0.5819	0.1289	23.08	0.5614	0.1304
BAGS (Peng et al., 2024)	20.89	0.3817	0.1846	18.94	0.3246	0.2317	21.62	0.4573	0.1748	20.73	0.4215	0.1809	23.15	0.5637	0.1271	22.58	0.5392	0.1486
CoMo-Gaussian (Lee et al., 2025)	21.58	0.4217	0.1789	19.93	0.3681	0.1982	26.87	0.6814	0.0874	25.63	0.6458	0.1121	28.31	0.7296	0.0847	27.18	0.7063	0.0862
CoherentGS (Xu et al., 2025b)	21.18	0.4053	0.1086	19.76	0.3638	0.1379	26.61	0.6748	0.0739	25.19	0.6317	0.0958	27.93	0.7162	0.0634	27.02	0.7011	0.0786
MoB3R (Ours) + Opt	26.58	0.6037	0.1229	25.07	0.5614	0.1361	27.71	0.7048	0.0787	25.89	0.6431	0.1014	28.39	0.7446	0.0756	27.54	0.7097	0.0836
Feed-forward																		
WorldMirror (Liu et al., 2025)	23.41	0.4682	0.1546	21.18	0.3947	0.1905	22.87	0.4413	0.1482	22.41	0.4296	0.1639	23.09	0.4487	0.1528	22.74	0.4381	0.1531
NAFNet+WorldMirror (Chen et al., 2022; Liu et al., 2025)	23.82	0.4831	0.1388	21.63	0.4108	0.1827	23.34	0.4572	0.1475	22.89	0.4461	0.1486	23.57	0.4643	0.1372	23.18	0.4537	0.1519
Fine-tuned WorldMirror	24.48	0.5116	0.1321	22.05	0.4297	0.1684	23.77	0.4759	0.1469	23.56	0.4828	0.1417	23.91	0.4794	0.1398	23.38	0.4702	0.1496
MoB3R (Ours)	25.41	0.5569	0.1193	23.28	0.4756	0.1512	26.08	0.6154	0.1027	24.96	0.5882	0.1216	27.43	0.6778	0.0811	25.79	0.6214	0.1046

Table 1: Quantitative comparison on DL3DV (Ling et al., 2024). We evaluate deblurring (Deblur) and novel view synthesis (NVS) quality under 4, 12, and 32 input views. We mark the Best and Second best results within each method category.

Sharp and reblur supervision. We use the same image loss for sharp rendering and reblurring, $\ell_{\mathrm{img}}(X, Y) = \|X - Y\|_2^2 + \lambda_{\mathrm{LPIPS}} \mathrm{LPIPS}(X, Y)$, where X and Y denote the predicted and reference images, respectively. We set $\lambda_{\mathrm{LPIPS}} = 0.05$, following WorldMirror. The sharp loss compares Gaussian renderings with the available sharp target views. To enforce blur-formation consistency, we render the predicted 3DGS at the $S = 7$ predicted trajectory poses and average the results:

$$\widehat{B}_i = \frac{1}{S} \sum_{j=1}^{S} \mathcal{R}\left(\widehat{\mathcal{G}}, \widehat{T}_i^j\right), \qquad \mathcal{L}_{\mathrm{reblur}} = \ell_{\mathrm{img}}(\widehat{B}_i, B_i), \tag{2}$$

where $\widehat{B}_i$ is the reblurred prediction for input B_i. The appearance objective is $\mathcal{L}_{\mathrm{app}} = \mathcal{L}_{\mathrm{sharp}} + \mathcal{L}_{\mathrm{reblur}} + \mathcal{L}_{\mathrm{GS}}$, where $\mathcal{L}_{\mathrm{GS}}$ groups the Gaussian-depth and rendered-depth consistency terms of WorldMirror. We keep their original weights and set the weight of $\mathcal{L}_{\mathrm{reblur}}$ to one.

3.3.3 Multi-Stage Training

We train MoB3R in two stages that follow its geometry-to-appearance pipeline. In stage 1, we train the geometry network with a progressive curriculum. We first use single frames to learn intra-exposure motion without relying on cross-view cues, and then train on multiple views while retaining the per-frame trajectory loss and adding the camera, depth, and point-map supervision of WorldMirror. We fully fine-tune the geometry network throughout this stage; empirically, this is more effective than LoRA (Hu et al., 2022) because trajectory recovery changes the original single-pose output space. In stage 2, we freeze the backbone and geometry heads, and fine-tune the complete Gaussian branch, together with the motion-field encoder and cross-attention module using $\mathcal{L}_{\mathrm{app}}$.

4 Experiments

4.1 Experimental Setup

Training datasets. No existing dataset provides large-scale paired blurred–sharp multi-view images. We therefore construct training data from DL3DV (Ling et al., 2024), TartanAir (Wang et al.,

Method	4 Views						12 Views						32 Views					
	Deblur			NVS			Deblur			NVS			Deblur			NVS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Per-scene optimization																		
BAD-Gaussians (Zhao et al., 2024)	21.36	0.4019	0.1934	20.13	0.3615	0.2011	24.04	0.5737	0.1262	23.30	0.5509	0.1498	25.96	0.6862	0.1041	25.38	0.6747	0.1052
BAGS (Peng et al., 2024)	22.67	0.4747	0.1537	20.51	0.4290	0.1875	23.57	0.5726	0.1428	22.78	0.5495	0.1461	24.97	0.6627	0.1063	24.49	0.6471	0.1249
CoMo-Gaussian (Lee et al., 2025)	26.19	0.6406	0.1109	24.61	0.5939	0.1207	30.38	0.8000	0.0602	29.10	0.7663	0.0788	31.52	0.8395	0.0526	30.44	0.8158	0.0539
CoherentGS (Xu et al., 2025b)	25.88	0.6257	0.0861	24.44	0.5886	0.1132	30.14	0.7928	0.0479	28.73	0.7489	0.0647	31.09	0.8247	0.0364	30.29	0.8118	0.0498
MoB3R (Ours) + Opt	28.87	0.7038	0.0961	27.61	0.6735	0.1082	30.92	0.8124	0.0521	29.41	0.7698	0.0716	31.64	0.8512	0.0418	30.83	0.8219	0.0557
Feed-forward																		
WorldMirror (Liu et al., 2025)	25.83	0.5762	0.1247	24.02	0.5284	0.1489	25.47	0.5693	0.1206	25.12	0.5571	0.1338	25.62	0.5724	0.1254	25.41	0.5638	0.1302
NAFNet+WorldMirror (Chen et al., 2022; Liu et al., 2025)	26.27	0.5918	0.1083	24.51	0.5437	0.1398	25.93	0.5841	0.1189	25.58	0.5726	0.1194	26.14	0.5893	0.1061	25.87	0.5801	0.1227
Fine-tuned WorldMirror	26.96	0.6238	0.0968	24.92	0.5587	0.1316	26.27	0.5985	0.1097	25.86	0.5827	0.1188	26.31	0.6017	0.1054	25.98	0.5873	0.1186
MoB3R (Ours)	27.55	0.6719	0.0941	25.64	0.6031	0.1238	28.41	0.7362	0.0908	27.58	0.7168	0.1076	29.82	0.7883	0.0761	28.69	0.7481	0.0895

Table 2: Quantitative comparison on ExBluRF (Lee et al., 2023b). We evaluate deblurring (Deblur) and novel view synthesis (NVS) quality under 4, 12, and 32 input views. We mark the Best and Second best results within each method category.

Method	4 Views				12 Views				32 Views			
	Rot. Acc. ↑		Trans. Acc. ↑		Rot. Acc. ↑		Trans. Acc. ↑		Rot. Acc. ↑		Trans. Acc. ↑	
	R@10	R@30	t@0.1	t@0.3	R@10	R@30	t@0.1	t@0.3	R@10	R@30	t@0.1	t@0.3
WorldMirror (Liu et al., 2025)	67.4	88.7	54.8	76.5	71.2	90.3	58.9	79.6	73.0	91.1	61.2	81.4
NAFNet+WorldMirror (Chen et al., 2022; Liu et al., 2025)	68.6	89.4	56.1	77.6	72.5	91.0	60.3	80.8	74.2	91.7	62.4	82.5
Fine-tuned WorldMirror	70.7	90.9	58.4	77.2	75.6	90.7	64.8	82.6	76.1	91.5	65.2	85.1
MoB3R (Ours)	79.3	93.6	67.5	87.0	82.8	96.5	75.2	89.7	86.4	96.6	76.6	93.0

Table 3: Feed-forward camera pose accuracy on DL3DV (Ling et al., 2024) under 4, 12, and 32 input views. We mark the Best and Second best results.

2020), and HyperSim (Roberts et al., 2021), which together cover diverse indoor and outdoor scenes. Following the GoPro blur synthesis protocol (Nah et al., 2017), we use a video frame interpolation model (Zhong et al., 2024) to temporally upsample each sequence by 7×. We average seven consecutive frames in linear RGB to simulate exposure integration and convert the result back to sRGB. The center sharp frame serves as the reconstruction target.

Evaluation benchmarks. We evaluate on synthetic and real motion blur. For DL3DV (Ling et al., 2024), we reserve 10 indoor and 10 outdoor scenes with no overlap with training and apply the same blur synthesis pipeline. From the first 100 frames of each scene, we uniformly sample 32 candidate input views and eight fixed held-out views. The 4- and 12-view inputs are uniformly subsampled from the 32-view set. ExBluRF (Lee et al., 2023b) contains real handheld motion blur paired with sharp reference images. Following prior deblurring 3DGS methods (Zhao et al., 2024; Lee et al., 2025), we use its standard split and apply the same view-subsampling strategy. Both benchmarks are evaluated with 4, 12, and 32 input views. Additional results on the real-world DeblurNeRF benchmark (Ma et al., 2022) are provided in Appendix C.1.

Evaluation metrics. We evaluate deblurring at the input viewpoints and novel view synthesis (NVS) at held-out viewpoints using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). For NVS, we follow the two-pass protocol of WorldMirror (Liu et al., 2025) and use the test-time pose alignment of AnySplat (Jiang et al., 2025b) to decouple rendering quality from pose estimation. Following RayZer (Jiang et al., 2025a), we report the percentage of relative poses below rotation and translation error thresholds, denoted by R@θ and t@d, respectively.

Compared methods. We compare per-scene optimization and feed-forward methods. The former include BAD-Gaussians (Zhao et al., 2024), BAGS (Peng et al., 2024), CoMo-Gaussian (Lee et al., 2025), and CoherentGS (Xu et al., 2025b), which optimize the scene and blur model from COLMAP (Schönberger & Frahm, 2016) initialization. Feed-forward baselines include WorldMirror (Liu et al., 2025); NAFNet (Chen et al., 2022)+WorldMirror, which deblurs each input before reconstruction; and Fine-tuned WorldMirror, which directly adapts WorldMirror to our blurred training data without architectural changes. Fine-tuned WorldMirror follows the same geometry-to-appearance training order: it first fine-tunes geometry on blurred inputs and then trains the Gaussian branch with geometry frozen. To combine feed-forward prediction with per-scene optimization, we further report MoB3R+Opt. Specifically, we initialize CoMo-Gaussian (Lee et al., 2025) with the camera poses and point cloud predicted by MoB3R, and then run 1K per-scene refinement iterations.



Figure 4: Qualitative comparison under sparse views on DL3DV and ExBluRF. Weak multi-view constraints leave per-scene results blurred or oversmoothed (crops), whereas our initialization yields sharper MoB3R+Opt results.

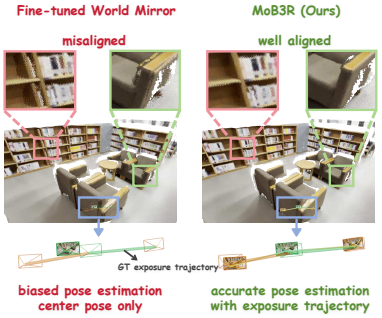

Figure 5: Geometry and pose comparison. Fine-tuned WorldMirror yields biased poses and misaligned geometry (left); MoB3R aligns its exposure trajectory and geometry with the ground truth (right).

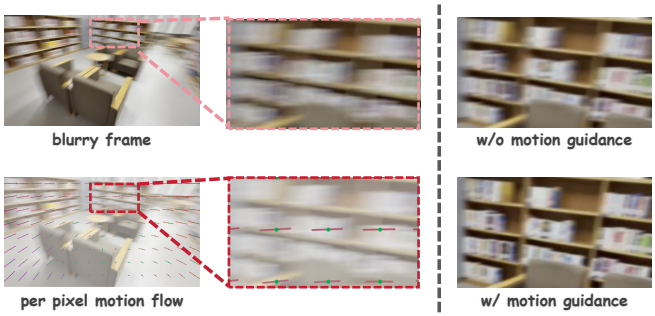

Figure 6: Effect of motion guidance. Left: a blurry frame and its per-pixel exposure motion derived from the recovered trajectory and depth. Right: without motion guidance, residual blur remains along the motion direction; with it, the Gaussian branch recovers sharper structures.

Implementation details. MoB3R is initialized from the pretrained WorldMirror (Liu et al., 2025) model. We train on four NVIDIA RTX PRO 6000 Blackwell GPUs with 96 GB memory each. The geometry stage first learns single-frame trajectories for 5K steps and then trains on multiple views for 20K steps; the appearance stage is trained for 10K steps. We use AdamW with cosine decay after a 1K-step linear warm-up. The peak learning rates are 5×10^{-6} for the ViT backbone and 5×10^{-5} for the downstream heads. We use gradient checkpointing and bfloat16 mixed-precision training to reduce memory consumption. During multi-view training, the number of input views is uniformly sampled from 2 to 24, with at most 24 views per GPU. Each image is randomly resized with an aspect ratio in $[0.5, 2.0]$ and a pixel count in $[10^5, 2.5 \times 10^5]$; both dimensions are aligned to the ViT patch size of 14. We use a fixed random seed for all experiments.

4.2 Quantitative Results

Tables 1 and 2 report rendering quality under 4, 12, and 32 input views. MoB3R leads all feed-forward baselines across datasets, view counts, and metrics. At 12 views, it improves deblurring/NVS PSNR over Fine-tuned WorldMirror by 2.31/1.40 dB on DL3DV and 2.14/1.72 dB on ExBluRF. The margin widens with more views, showing that the recovered trajectories effectively use cross-view evidence and generalize to real blur. MoB3R+Opt leads PSNR and SSIM in all but one per-scene setting, showing that our poses and point cloud provide a strong initialization. The gains extend to perceptual quality: MoB3R has the lowest LPIPS among feed-forward methods in every setting, while MoB3R+Opt ranks top two in 11 of 12 per-scene settings. CoherentGS obtains lower LPIPS in several cases with a diffusion prior, whereas ours leads PSNR and SSIM without diffusion. For pose estimation, Table 3 shows that MoB3R performs best at every threshold and view count; at 12 views, it improves R@10 and t@0.1 over Fine-tuned WorldMirror by 7.2 and

Method	Setting	Time
BAD-Gaussians (Zhao et al., 2024)	Per-scene	∼53 min
BAGS (Peng et al., 2024)	Per-scene	∼42 min
CoMo-Gaussian (Lee et al., 2025)	Per-scene	∼1.6 h
CoherentGS (Xu et al., 2025b)	Per-scene	∼1.1 h
MoB3R (Ours) + Opt	Per-scene	∼6 min
WorldMirror (Liu et al., 2025)	Feed-forward	seconds
MoB3R (Ours)	Feed-forward	seconds

Table 4: Reconstruction time. Per-scene methods report optimization; feed-forward methods report one forward pass.

Traj.	RInv.	Motion Guid.	Reblur	Pose Acc. ↑				Deblur ↑	
				R@10	R@30	t@0.1	t@0.3	PSNR	SSIM
–	–	–	–	75.6	90.7	64.8	82.6	23.77	0.4759
✓	–	–	–	76.3	92.1	65.9	83.0	23.61	0.4802
✓	✓	–	–	82.8	96.5	75.2	89.7	24.19	0.5187
✓	✓	✓	–	82.8	96.5	75.2	89.7	25.47	0.5869
✓	✓	✓	✓	82.8	96.5	75.2	89.7	26.08	0.6154

Table 5: Component ablation on DL3DV with 12 views, starting from directly fine-tuned WorldMirror. Traj., RInv., Motion Guid., and Reblur denote trajectory modeling, reversal-invariant pose loss, motion guidance, and the reblur objective, respectively.

10.4 points. This confirms the observation in Fig. 2(d): unlike direct fine-tuning, trajectory recovery reduces the motion-aligned pose bias. Together with the rendering gains, this validates explicit motion recovery for both geometry and sharp appearance reconstruction. ExBluRF pose results are provided in Appendix C.2.

4.3 Qualitative Results

As shown in Fig. 4, BAD-Gaussians retains severe blur, while CoMo-Gaussian and CoherentGS lose fine details. Initialized by our recovered poses and point cloud, MoB3R+Opt produces sharper structures and details closer to the ground truth. The gain is most visible around thin edges and repeated textures, where competing methods retain blur or oversmooth details. Figure 5 further connects trajectory accuracy to geometry: center-pose bias causes cross-view misalignment, whereas the recovered exposure trajectory yields consistent poses and scene structure.

4.4 Efficiency

Table 4 shows that per-scene methods require 42 minutes to 1.6 hours, whereas MoB3R runs in seconds. MoB3R+Opt takes approximately six minutes, remaining substantially faster than optimization from scratch. Thus, MoB3R preserves feed-forward efficiency and provides an efficient initialization for optional refinement.

4.5 Ablation Studies

We ablate the main components of MoB3R on DL3DV with 12 input views, starting from directly fine-tuned WorldMirror. Table 5 progressively adds each component. Trajectory modeling improves R@10 and t@0.1 by 0.7 and 1.1 points, while the reversal-invariant loss further adds 6.5 and 9.3 points. Motion guidance and the reblur objective further add 1.28 and 0.61 dB, confirming that recovered motion guides appearance reconstruction and cycle consistency improves deblurring. Fig. 6 visualizes this effect: without motion guidance, residual blur remains along the exposure motion. As expected, the last two components leave pose accuracy unchanged because they act only in the appearance stage. This validates the two-stage design.

5 Limitations

MoB3R assumes a static scene; independently moving objects may therefore cause inaccurate trajectories or reconstruction artifacts. It also treats input views as unordered and does not exploit temporal cues from sequential captures, which could reduce the forward–backward ambiguity.

6 Conclusion

We presented MoB3R, a feed-forward framework for sharp 3D Gaussian reconstruction from motion-blurred multi-view images. By treating blur as a cue to camera motion, MoB3R recovers intra-exposure trajectories and uses them to guide sharp appearance reconstruction. Experiments show that this design reduces motion-aligned pose bias and improves both pose accuracy and rendering quality in a single forward pass.

AI USE STATEMENT

We used generative AI tools only to polish the writing of this paper. All research ideas, methods, experiments, and results are the work of the authors, who reviewed all AI-assisted text and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The architecture, losses, and training stages of MOB3R are described in Section 3, with additional implementation details in Appendix A. The construction of the training data, the evaluation protocols, and the metrics are described in the experimental setup and in Appendix B.

REFERENCES

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1050–1060, 2025.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pp. 370–386. Springer, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3d reconstruction at scale. *arXiv preprint arXiv:2602.23361*, 2026.

Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. PF3plat: Pose-free feed-forward 3d gaussian splatting for novel view synthesis. In *International Conference on Machine Learning*, pp. 23662–23681, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, and Georgios Pavlakos. Rayzer: A self-supervised large view synthesis model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4929, 2025a.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025b.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8183–8192, 2018.

Byeonghyeon Lee, Howoong Lee, Usman Ali, and Eunbyung Park. Sharp-nerf: Grid-based fast deblurring neural radiance fields using sharpness prior. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3709–3718, 2024.

Dogyoon Lee, Minhyeok Lee, Chajin Shin, and Sangyoun Lee. Dp-nerf: Deblurred neural radiance field with physical scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12386–12396, 2023a.

Dongwoo Lee, Jeongtaek Oh, Jaesung Rim, Sunghyun Cho, and Kyoung Mu Lee. Exblurf: Efficient radiance fields for extreme motion blurred images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17639–17648, 2023b.

Jungho Lee, Donghyeong Kim, Dogyoon Lee, Suhwan Cho, Minhyeok Lee, Wonjoon Lee, Taeoh Kim, Dongyoon Wee, and Sangyoun Lee. Comogaussian: Continuous motion-aware gaussian splatting from motion-blurred images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26415–26424, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, June 2024.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

Yiren Lu, Yunlai Zhou, Disheng Liu, Tuo Liang, and Yu Yin. Bard-gs: Blur-aware reconstruction of dynamic scenes via gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16532–16542, 2025.

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12861–12870, 2022.

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

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

Jeongtaek Oh, Jaeyoung Chung, Dongwoo Lee, and Kyoung Mu Lee. Deblurgs: Gaussian splatting for camera motion blur. *arXiv preprint arXiv:2404.11358*, 2024.

Cheng Peng and Rama Chellappa. Pdrf: progressively deblurring radiance field for fast scene reconstruction from blurry images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2029–2037, 2023.

Cheng Peng, Yutao Tang, Yifan Zhou, Nengyu Wang, Xijun Liu, Deming Li, and Rama Chellappa. Bags: Blur agnostic gaussian splatting through multi-scale kernel modeling. In *European Conference on Computer Vision*, pp. 293–310. Springer, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Changyue Shi, Wangbo Yu, Chaoran Feng, and Li Yuan. Deblurnvs: Geometric latent diffusion for novel view synthesis from sparse motion-blurred images. *arXiv preprint arXiv:2606.01315*, 2026.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026.

Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. Bad-nerf: Bundle adjusted deblur neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4170–4179, 2023.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*, 2025b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European conference on computer vision*, pp. 456–473. Springer, 2024.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, and Xiaowei Zhou. Scal3R: Scalable test-time training for large-scale 3d reconstruction. *arXiv preprint arXiv:2604.08542*, 2026.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16453–16463, 2025a.

Zhankuo Xu, Chaoran Feng, Yingtao Li, Jianbin Zhao, Jiashu Yang, Wangbo Yu, Li Yuan, and Yonghong Tian. Breaking the vicious cycle: Coherent 3d gaussian splatting from sparse and motion-blurred views. *arXiv preprint arXiv:2512.10369*, 2025b.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, 2025.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19447–19456, 2024.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21936–21947, 2025.

Lingzhe Zhao, Peng Wang, and Peidong Liu. Bad-gaussians: Bundle adjusted deblur gaussian splatting. In *European Conference on Computer Vision*, pp. 233–250. Springer, 2024.

Zhihang Zhong, Gurunandan Krishnan, Xiao Sun, Yu Qiao, Sizhuo Ma, and Jian Wang. Clearer frames, anytime: Resolving velocity ambiguity in video frame interpolation. In *European Conference on Computer Vision*, pp. 346–363. Springer, 2024.

APPENDIX

A ADDITIONAL METHOD DETAILS

A.1 MOTION-FROM-BLUR TRAJECTORY RECOVERY

Motion-token construction. WorldMirror (Liu et al., 2025) maintains two learnable camera-token embeddings: one for the first frame and one shared by all remaining frames. We preserve this distinction for both endpoint tokens. Before fine-tuning, the start and end embeddings are initialized from their corresponding pretrained camera embeddings with small independent perturbations:

$$z^k_{\text{first}} \leftarrow z^m_{\text{first}} + \varepsilon^k_{\text{first}}, \qquad z^k_{\text{rest}} \leftarrow z^m_{\text{rest}} + \varepsilon^k_{\text{rest}}, \qquad k \in \{s, e\}. \tag{3}$$

Here, m, s, and e denote the midpoint, start, and end of exposure. This initialization is performed once. All three token sets remain trainable during geometry fine-tuning and are fixed model parameters at inference. Their output features nevertheless remain image-dependent after frame-wise and cross-view attention.

Camera prediction and trajectory. The three output tokens share one CameraHead initialized from WorldMirror and jointly fine-tuned with the tokens. They predict the same camera encoding $c = (\mathbf{t}, \mathbf{q}, \mathbf{f})$, consisting of translation, rotation quaternion, and field of view. Thus, the motion tokens predict absolute endpoint cameras rather than pose offsets. We convert their outputs to camera poses and sample seven poses by linear translation interpolation and quaternion SLERP from start to midpoint and then from midpoint to end. The fourth sample is the shared midpoint. Since the intrinsics remain fixed during exposure, the same ground-truth intrinsics supervise all three camera predictions, and the predicted midpoint intrinsics are used for all seven poses at inference.

Reversal-invariant camera loss. The three camera predictions use the original WorldMirror camera loss. For a prediction $\widehat{c}$ and its ground truth c, we write $\ell_{\text{cam}}(\widehat{c}, c) = \|\widehat{c} - c\|_\epsilon$, where $\|\cdot\|_\epsilon$ is the Huber loss. A blurred image does not reveal the temporal order of its exposure trajectory: exchanging the start and end cameras produces the same set of sharp renderings. We therefore define

$$\begin{aligned} \mathcal{L}_{\text{fwd}} &= \ell_{\text{cam}}(\widehat{c}_i^s, c_i^s) + \ell_{\text{cam}}(\widehat{c}_i^e, c_i^e), \qquad \mathcal{L}_{\text{rev}} = \ell_{\text{cam}}(\widehat{c}_i^s, c_i^e) + \ell_{\text{cam}}(\widehat{c}_i^e, c_i^s), \\ \mathcal{L}_{\text{mid}} &= \ell_{\text{cam}}(\widehat{c}_i^m, c_i^m), \qquad\qquad\qquad \mathcal{L}_{\text{pose}} = \mathcal{L}_{\text{mid}} + \min(\mathcal{L}_{\text{fwd}}, \mathcal{L}_{\text{rev}}). \end{aligned} \tag{4}$$

These terms are computed independently for each image and then averaged over the training batch.

Geometry-stage objective. We retain the depth and point-map supervision of WorldMirror and optimize $\mathcal{L}_{\text{geo}} = \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{point}}\mathcal{L}_{\text{point}}$, where $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{point}}$ follow the confidence-weighted regression and gradient losses of WorldMirror, including their original weights. We encourage readers to refer to the WorldMirror technical report for further implementation details. We use the same objective for single-frame trajectory learning and multi-view geometry training. The latter retains the per-image trajectory loss while allowing frame-wise predictions to use cross-view attention.

A.2 Motion-Guided Appearance Reconstruction

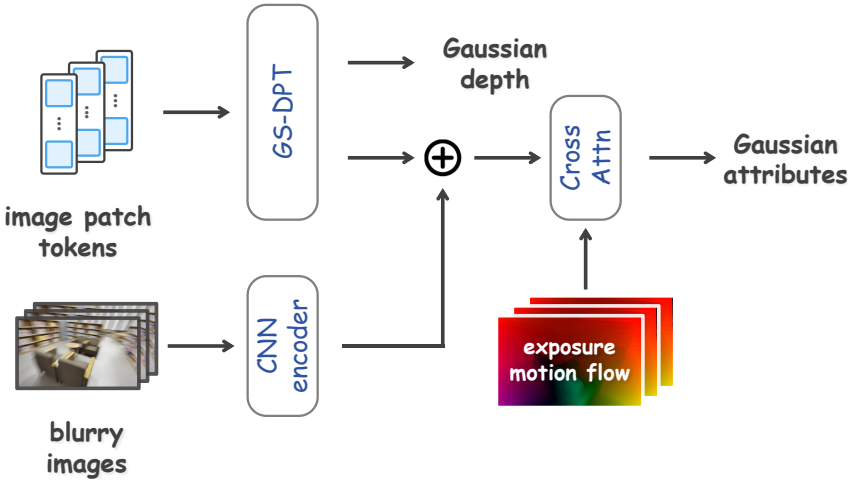

Figure 7: Detailed appearance branch. GS-DPT decodes the image patch tokens into Gaussian depth and features, while a shallow CNN extracts image features from the blurry inputs. The fused feature is conditioned on the exposure motion field through cross-attention before the Gaussian attributes are predicted.

Gaussian branch. Figure 7 details the appearance stage. Following WorldMirror (Liu et al., 2025), GS-DPT decodes multi-level image patch tokens into a Gaussian depth map $\widehat{D}_i^g$ and a 128-channel feature map F_i^{DPT}. A shallow CNN maps the blurry image B_i to an image feature F_i^{img} of the same size. The two features are added as $F_i = F_i^{\text{DPT}} + F_i^{\text{img}}$. The Gaussian depth is back-projected with the predicted midpoint camera to obtain one Gaussian center per context-view pixel.

Motion-guided cross-attention. The four-channel exposure motion field M_i concatenates the midpoint-to-start and midpoint-to-end flows defined in the main paper. We encode M_i into motion tokens U_i and reshape the fused feature F_i into Gaussian feature tokens Z_i. Both token sets preserve the same spatial order. Within each frame, Z_i provides the queries and U_i provides the keys and values, $\widetilde{Z}_i = Z_i + \mathrm{CA}(Z_i, U_i)$, where CA denotes multi-head cross-attention. The residual connection retains the original Gaussian feature while adding the local motion information. We reshape $\widetilde{Z}_i$ back to a feature map before attribute prediction.

Gaussian attributes. The attribute decoder consists of a 3×3 convolution from 128 to 256 channels, ReLU, and a 1×1 convolution with 12 output channels. For each Gaussian, these channels predict a rotation quaternion, anisotropic scale, opacity, residual degree-zero spherical-harmonic color, and a fusion weight, with dimensions 4, 3, 1, 3, and 1, respectively. Together with the centers obtained from $\widehat{D}_i^g$, they form the sharp 3D Gaussians. As in WorldMirror, overlapping Gaussians from different context views are merged by voxelization using their predicted fusion weights.

Appearance-stage objective. Let ℓ_{img} be the MSE–LPIPS image loss defined in the main paper. The sharp rendering loss $\mathcal{L}_{\mathrm{sharp}}$ compares renderings of the predicted Gaussians with the corresponding sharp target images. For reblur supervision, we render at the seven poses sampled along each predicted exposure trajectory, average the renderings, and compare the result with the blurry input using the same image loss. This gives $\mathcal{L}_{\mathrm{reblur}}$. We further retain the Gaussian-depth loss $\mathcal{L}_{\mathrm{gsdepth}}$ and the rendered-depth gradient-consistency loss $\mathcal{L}_{\mathrm{consis}}$ from WorldMirror. The complete appearance objective is

$$\mathcal{L}_{\mathrm{app}} = \lambda_{\mathrm{sharp}}\mathcal{L}_{\mathrm{sharp}} + \lambda_{\mathrm{reblur}}\mathcal{L}_{\mathrm{reblur}} + \lambda_{\mathrm{gsdepth}}\mathcal{L}_{\mathrm{gsdepth}} + \lambda_{\mathrm{consis}}\mathcal{L}_{\mathrm{consis}}. \tag{5}$$

Here, $\mathcal{L}_{\mathrm{gsdepth}}$ supervises $\widehat{D}_i^g$ with the same confidence-weighted depth and gradient terms as WorldMirror, while $\mathcal{L}_{\mathrm{consis}}$ matches the gradients of the rendered depth and the depth-head prediction in high-confidence regions. We use the WorldMirror settings for the original terms and the reblur setting reported in the main paper. During this stage, the geometry network is frozen, while the complete Gaussian branch, motion-field encoder, and cross-attention module are fine-tuned. Further details of the original Gaussian losses and voxel merging follow the WorldMirror technical report.

B Evaluation Details

DL3DV. For each test scene, we use the first 100 frames and uniformly select 32 as the maximum input set. We select eight fixed views from the remaining frames for NVS evaluation. The 4- and 12-view inputs are uniformly subsampled from the 32-view set, while the same eight held-out views are used in all settings. Deblurring metrics are computed at the input views, and NVS metrics are computed at the held-out views. Camera pose accuracy is evaluated against the COLMAP poses provided by DL3DV after global alignment.

ExBluRF. We follow the standard evaluation protocol established by DeblurNeRF (Ma et al., 2022) and ExBluRF (Lee et al., 2023b): frames whose indices are divisible by the hold number are treated as sharp test views for NVS evaluation, while the remaining blurred frames form the input candidate set. We uniformly subsample this set to obtain 4, 12, and 32 input views. Deblurring is evaluated at input views with paired references, while NVS is evaluated only at the held-out sharp views.

DeblurNeRF. We use the standard train–test split and report NVS metrics on the held-out sharp views. Since its blurred input views do not have paired sharp references, we do not report deblurring metrics on this benchmark.

Metrics. We report PSNR, SSIM, and LPIPS for image quality. For camera poses, R@θ and t@d denote the percentages of relative rotations and translations below the corresponding thresholds. All methods use the same input views, held-out views, and pose-alignment protocol within each benchmark.

C Additional Quantitative Results

C.1 Novel-View Synthesis on DeblurNeRF

DeblurNeRF (Ma et al., 2022) contains real motion-blurred inputs and sharp held-out views. Following its standard split, we evaluate novel-view synthesis on the held-out views. We compare with BAD-Gaussians (Zhao et al., 2024), BAGS (Peng et al., 2024), CoMo-Gaussian (Lee et al., 2025), CoherentGS (Xu et al., 2025b), WorldMirror (Liu et al., 2025), and NAFNet (Chen et al., 2022). Table 6 reports the results; deblurring metrics are unavailable because the blurry input views have no paired sharp references. WM denotes WorldMirror.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Per-scene optimization			
BAD-Gaussians⋆	21.69	0.6471	0.1262
BAGS⋆	26.70	0.8237	0.0956
CoMo-Gaussian⋆	27.85	0.8431	0.0822
CoherentGS (Xu et al., 2025b)	27.48	0.8367	0.0694
MoB3R (Ours)+Opt	28.03	0.8514	0.0795
Feed-forward			
WorldMirror	23.84	0.7012	0.1583
NAFNet+WM	24.37	0.7198	0.1426
Fine-tuned WM	24.52	0.7263	0.1371
MoB3R (Ours)	26.14	0.8086	0.1004

Table 6: Novel-view synthesis on real-world scenes from DeblurNeRF (Ma et al., 2022). Since blurred input views have no paired sharp references, only held-out-view NVS is evaluated. ⋆ Results are taken from CoMo-Gaussian (Lee et al., 2025).

C.2 Camera Pose Accuracy on ExBluRF

Table 7 complements the DL3DV pose evaluation in the main paper with results on real motion blur. MoB3R consistently improves both rotation and translation accuracy across all input-view settings, showing that explicit trajectory recovery also reduces pose bias on real captures.

Method	4 Views				12 Views				32 Views			
	Rot. Acc. ↑		Trans. Acc. ↑		Rot. Acc. ↑		Trans. Acc. ↑		Rot. Acc. ↑		Trans. Acc. ↑	
	R@10	R@30	t@0.1	t@0.3	R@10	R@30	t@0.1	t@0.3	R@10	R@30	t@0.1	t@0.3
WorldMirror (Liu et al., 2025)	64.3	86.4	51.9	73.8	68.1	88.2	56.2	77.1	69.8	89.0	58.1	78.9
NAFNet+WorldMirror (Chen et al., 2022; Liu et al., 2025)	65.5	87.1	53.2	75.0	69.4	89.0	57.5	78.4	71.0	89.7	59.3	80.2
Fine-tuned WorldMirror	67.8	86.9	54.6	74.7	70.5	89.8	57.2	79.6	70.8	90.4	60.9	82.1
MoB3R (Ours)	76.1	91.7	63.9	84.5	79.6	94.9	72.1	87.6	83.0	94.8	73.4	91.2

Table 7: Feed-forward camera pose accuracy on ExBluRF (Lee et al., 2023b) under 4, 12, and 32 input views. We mark the Best and Second best results.

C.3 Depth Accuracy on DL3DV

Both the exposure motion field and the Gaussian centers depend on the depth predicted by the geometry stage, so we evaluate it directly on blurred inputs. DL3DV consists of real captures without sensor depth. We therefore use as pseudo ground truth the depth maps reconstructed by COLMAP multi-view stereo (Schönberger & Frahm, 2016) from the sharp frames. Since multi-view stereo leaves holes in textureless, sky, and occluded regions, we evaluate only on pixels with valid pseudo ground-truth depth. For each input view, the predicted depth is aligned to the pseudo ground truth of the corresponding sharp midpoint frame by per-image median scaling. We report the absolute relative error (AbsRel) and the fraction of pixels with $\max(\hat{d}/d, d/\hat{d}) < 1.25$ ($\delta_{1.25}$).

Table 8 shows that blur substantially degrades the depth of the pretrained WorldMirror. Deblurring the inputs with NAFNet or fine-tuning on blurred data only partially recovers the accuracy, whereas MoB3R achieves the lowest error in all settings. Since pixels with invalid pseudo ground truth are excluded, these numbers reflect relative accuracy among methods rather than absolute depth quality.

Method	4 Views		12 Views		32 Views	
	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑
WorldMirror (Liu et al., 2025)	0.094	0.912	0.089	0.921	0.091	0.917
NAFNet+WorldMirror (Chen et al., 2022; Liu et al., 2025)	0.087	0.924	0.083	0.931	0.082	0.934
Fine-tuned WorldMirror	0.079	0.937	0.074	0.945	0.071	0.948
MoB3R (Ours)	0.066	0.953	0.063	0.960	0.062	0.962

Table 8: Depth accuracy from blurred inputs on DL3DV (Ling et al., 2024) under 4, 12, and 32 input views. Predicted depth at each input view is compared with the pseudo ground-truth depth of the corresponding sharp midpoint frame after per-image median scale alignment, using only pixels with valid pseudo ground-truth depth.

C.4 Exposure Trajectory Accuracy on DL3DV

Per-scene methods such as CoMo-Gaussian (Lee et al., 2025) estimate the exposure trajectory jointly with the scene by minimizing a photometric reblur loss from COLMAP initialization. Under sparse blurry views, this non-convex objective easily converges to local minima. We therefore evaluate the full exposure trajectory rather than only its midpoint. Since ExBluRF and DeblurNeRF do not provide ground-truth exposure trajectories, we conduct this evaluation on DL3DV, whose synthesized blur provides the ground-truth camera pose of every averaged frame. For each input view, we compare all $S = 7$ poses sampled along the predicted trajectory with the corresponding ground-truth exposure poses, and compute R@θ and t@d over all trajectory poses after the same global alignment as in Table 3. The numbers are therefore not directly comparable to the midpoint-only evaluation.

Table 9 shows that CoMo-Gaussian recovers markedly less accurate trajectories, especially with 4 input views, where weak multi-view constraints leave the optimization trapped near its erroneous initialization. The gap narrows as more views constrain the optimization, but MoB3R remains clearly ahead even with 32 views while predicting the trajectories in a single forward pass. Running the same CoMo-Gaussian optimization from our prediction (MoB3R+Opt) further refines them; the gain is small with 4 views and grows with denser views, where the photometric objective is better constrained. This indicates that the gap stems from the initialization and the optimization landscape rather than from the trajectory model itself.

Method	4 Views				12 Views				32 Views			
	Rot. Acc. ↑		Trans. Acc. ↑		Rot. Acc. ↑		Trans. Acc. ↑		Rot. Acc. ↑		Trans. Acc. ↑	
	R@10	R@30	t@0.1	t@0.3	R@10	R@30	t@0.1	t@0.3	R@10	R@30	t@0.1	t@0.3
CoMo-Gaussian (Lee et al., 2025)	46.3	70.8	35.2	59.1	59.6	79.4	47.9	68.7	67.8	86.3	54.4	76.2
MoB3R (Ours)	76.7	92.5	64.4	85.2	80.9	95.9	72.8	88.6	85.0	96.2	74.6	92.1
MoB3R (Ours) + Opt	77.3	92.7	65.5	85.6	82.4	96.3	74.8	89.6	86.9	96.7	77.3	93.4

Table 9: Exposure-trajectory accuracy on DL3DV (Ling et al., 2024) under 4, 12, and 32 input views. Unlike Table 3, which evaluates only the midpoint pose, all $S = 7$ poses along each exposure trajectory are compared with the ground-truth exposure poses. MoB3R+Opt runs the same CoMo-Gaussian optimization initialized with MoB3R.

D Additional Analysis and Ablations

D.1 Quantifying Motion-Aligned Pose Bias

Protocol of the preliminary analysis. We fit a vanilla 3DGS (Kerbl et al., 2023) to the sharp views of each scene in the DL3DV (Ling et al., 2024) test set. Starting from this reconstruction, we manually specify an exposure trajectory around each test camera and average seven sharp renderings to synthesize a blurred frame. Varying the trajectory length changes the blur magnitude without changing scene content. For the directional analysis, we avoid translation along and rotation about the optical axis, as both produce spatially varying flow directions. This restriction is used only for this analysis. We apply fine-tuned WorldMirror (Liu et al., 2025) to the synthesized blurred views and globally align its predictions to the sharp reference cameras. We assign every pixel the same depth before applying either camera motion, so the resulting flow depends only on camera motion rather than scene-depth variation. We use the endpoint flow to represent the blur direction and the residual-pose flow to represent the predicted-pose bias. Their dominant directions are computed

with a magnitude-weighted mean modulo $180°$, and blur magnitude is the mean endpoint-flow magnitude. Each point in the top plots of Fig. 2(d) compares the two directions for one view, while the bottom plot reports translational relative pose error as blur magnitude increases.

Correlation analysis. We further quantify the relation between exposure motion and pose prediction. In this controlled set, the exposure trajectories are randomly sampled around the reference cameras. To obtain a clear image-space motion direction, we sample only camera translations parallel to the image plane and keep the camera orientation fixed. The controlled set therefore contains neither translation along the optical axis nor camera rotation.

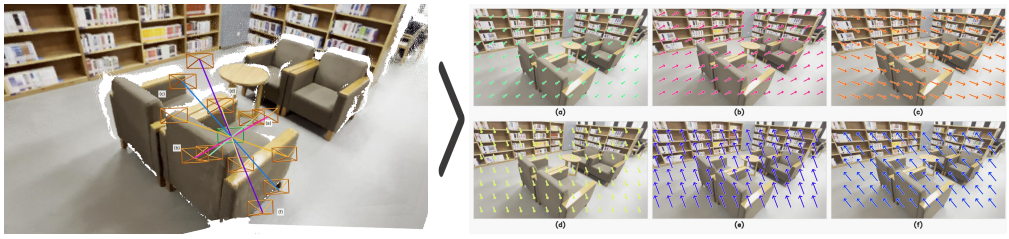

Figure 8: **Visualization of sampled exposure trajectories and their induced image-plane motion.** Left: We sample six camera trajectories on the image plane around a shared middle-exposure pose T_{mid}. For each trajectory, the translation direction and magnitude are randomly sampled, while the camera orientation remains fixed and T_{mid} is constrained to be the midpoint between T_{start} and T_{end}. Right: We visualize the corresponding pose variations as camera-motion-induced optical flow by projecting pixels with a fixed depth from T_{start} to T_{end}. Arrow directions indicate the induced 2D motion directions, while arrow lengths reflect the relative magnitudes of the sampled camera translations.

Figure 8 shows the sampled trajectories and their pose-induced 2D flows. To isolate camera motion from scene depth, we assign the same depth to all pixels when computing the flow. We summarize each exposure motion and aligned pose-residual flow field by its magnitude-weighted dominant direction. A blurred image does not distinguish the forward and reverse directions of its exposure trajectory: swapping T_{start} and T_{end} produces the same observation. We therefore treat directions separated by $180°$ as equivalent, map them to the doubled-angle circle, and report their axial circular correlation ρ_{axial}. A larger value indicates a stronger alignment between motion blur and pose bias.

Method	ρ_{axial}
Fine-tuned WorldMirror	0.79
MoB3R (Ours)	0.32

Table 10: Correlation between the motion-blur direction and the predicted pose-bias direction on the controlled DL3DV test set.

The correlation in Table 10 is consistent with Fig. 2(d) in the main paper: directly adapting a single-pose model does not remove its motion-aligned bias, whereas explicit exposure-trajectory modeling substantially weakens this correlation.

D.2 Base Model Selection

We compare three feed-forward models as candidates for our geometry backbone: AnySplat (Jiang et al., 2025b), Depth Anything 3 (Lin et al., 2025), and WorldMirror (Liu et al., 2025). We evaluate their pretrained camera predictions and rendered images on blurred DL3DV inputs with 12 views, using the same pose alignment and evaluation protocol as in the main paper. As shown in Table 11, WorldMirror ranks first on four of the seven metrics and second on the remaining three, providing the strongest overall performance under blur. We therefore adopt it as the base model for MoB3R.

Base model	Camera Pose			
	R@10↑	R@30↑	t@0.1↑	t@0.3↑
AnySplat (Jiang et al., 2025b)	59.8	83.7	46.3	69.5
Depth Anything 3 (Lin et al., 2025)	70.4	91.0	57.6	80.3
WorldMirror (Liu et al., 2025)	71.2	90.3	58.9	79.6

Base model	Deblur		
	PSNR↑	SSIM↑	LPIPS↓
AnySplat (Jiang et al., 2025b)	20.91	0.3897	0.1956
Depth Anything 3 (Lin et al., 2025)	22.54	0.4458	0.1509
WorldMirror (Liu et al., 2025)	22.87	0.4413	0.1482

Table 11: Base-model comparison on blurred DL3DV inputs with 12 views. We report camera pose accuracy and input-view deblurring. We mark the Best and Second best results.

D.3 Multi-Stage Geometry Training

Table 12 compares progressive training with directly optimizing the single-frame trajectory and multi-view geometry objectives together. Both settings use the same architecture, losses, training budget, and appearance stage. The progressive setting first learns exposure motion from individual frames and then introduces multi-view geometry, while retaining the per-frame trajectory loss. The joint setting optimizes both objectives from the start.

Schedule	Camera Pose			
	R@10↑	R@30↑	t@0.1↑	t@0.3↑
Joint training	80.8	94.7	72.8	88.1
Progressive training	82.8	96.5	75.2	89.7

Schedule	Deblur		
	PSNR↑	SSIM↑	LPIPS↓
Joint training	25.76	0.5817	0.1134
Progressive training	26.08	0.6154	0.1027

Table 12: Progressive versus joint geometry training on DL3DV with 12 input views. We report camera pose accuracy and deblurring quality. Both settings use the same training budget and appearance stage.

The results show that learning exposure motion before introducing multi-view geometry provides a better initialization than optimizing both objectives jointly from the start, improving both pose accuracy and deblurring quality.

D.4 Full Fine-Tuning versus LoRA

We compare full geometry fine-tuning with rank-32 LoRA applied to the transformer attention layers. Both settings add motion tokens and geometry heads to model exposure-time motion, and these new parameters are trainable in both cases. LoRA updates the pretrained backbone only through low-rank adapters, whereas full fine-tuning updates all pretrained parameters; the comparison therefore measures adaptation to the new exposure-trajectory representation rather than fine-tuning on blurred images alone. As shown in Table 13, full fine-tuning improves pose accuracy and deblurring quality, supporting broader adaptation of the pretrained representation when replacing a single-pose output with an exposure trajectory.

D.5 Number of Exposure Samples

The number of exposure samples S appears twice in our pipeline: the blurred data are synthesized by averaging S_{data} interpolated frames within the exposure interval, and the reblur loss averages renderings at S_{reblur} poses sampled along the predicted trajectory. To study their interaction, we fix

	Camera Pose				Deblur	
Adaptation	R@10↑	R@30↑	t@0.1↑	t@0.3↑	PSNR↑	SSIM↑
LoRA	79.4	95.2	69.4	87.7	25.08	0.5842
Full fine-tuning	82.8	96.5	75.2	89.7	26.08	0.6154

Table 13: Geometry adaptation on DL3DV with 12 input views. We report camera pose accuracy and deblurring quality.

the exposure interval of each original frame i to $t \in [i - \frac{1}{2}, i + \frac{1}{2}]$ and re-synthesize the training and test sets by averaging $S_{\text{data}} \in \{5, 7, 9\}$ frames interpolated uniformly within this interval, then retrain the appearance stage with $S_{\text{reblur}} \in \{5, 7, 9\}$ on each dataset. The geometry stage is shared across all settings and is trained only on the $S_{\text{data}} = 7$ data. Table 14 reports feed-forward deblurring quality on DL3DV with 12 input views, without per-scene optimization. Since the exposure interval is fixed, all columns share the same blur extent, and a larger S_{data} only discretizes the same exposure more densely. As the test sets still differ in this discretization, the results should be compared mainly within each column.

For every dataset, the matched setting $S_{\text{reblur}} = S_{\text{data}}$ achieves the best PSNR and LPIPS, and the performance drops as the mismatch grows, indicating that the reblur objective is most effective when its discretization agrees with the blur formation. The only exception is SSIM at $S_{\text{data}} = 7$, where $S_{\text{reblur}} = 9$ is slightly higher (0.6181 vs. 0.6154). Along the diagonal, increasing S from 5 to 7 brings a clear gain, whereas the gain from 7 to 9 is marginal while the rendering cost of the reblur loss grows linearly with S. Given the cost of synthesizing the data and retraining for each value, we did not tune S exhaustively. We instead use $S = 7$, an odd value that includes the midpoint pose and is comparable to the number of virtual views in BAD-Gaussians (Zhao et al., 2024).

	$S_{\text{data}} = 5$			$S_{\text{data}} = 7$			$S_{\text{data}} = 9$		
S_{reblur}	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
5	25.86	0.6098	0.1089	25.79	0.6064	0.1125	25.63	0.6015	0.1158
7	25.68	0.6021	0.1116	26.08	0.6154	0.1027	25.99	0.6139	0.1039
9	25.51	0.5973	0.1142	25.96	0.6181	0.1046	26.14	0.6212	0.1012

Table 14: Sensitivity to the number of exposure samples on DL3DV with 12 input views (feed-forward, deblurring). Columns: samples averaged within the same fixed exposure interval to synthesize the training and test blur; rows: samples used in the reblur loss. The geometry stage is fixed across settings. Best marks the best result within each column; our default is $S_{\text{data}} = S_{\text{reblur}} = 7$.

E Additional Qualitative Results

Figures 9–14 show additional results on DL3DV, ExBluRF, and DeblurNeRF under sparse views and real motion blur. Per-scene methods retain blur or lose fine details, whereas MoB3R+Opt produces sharper boundaries and textures.

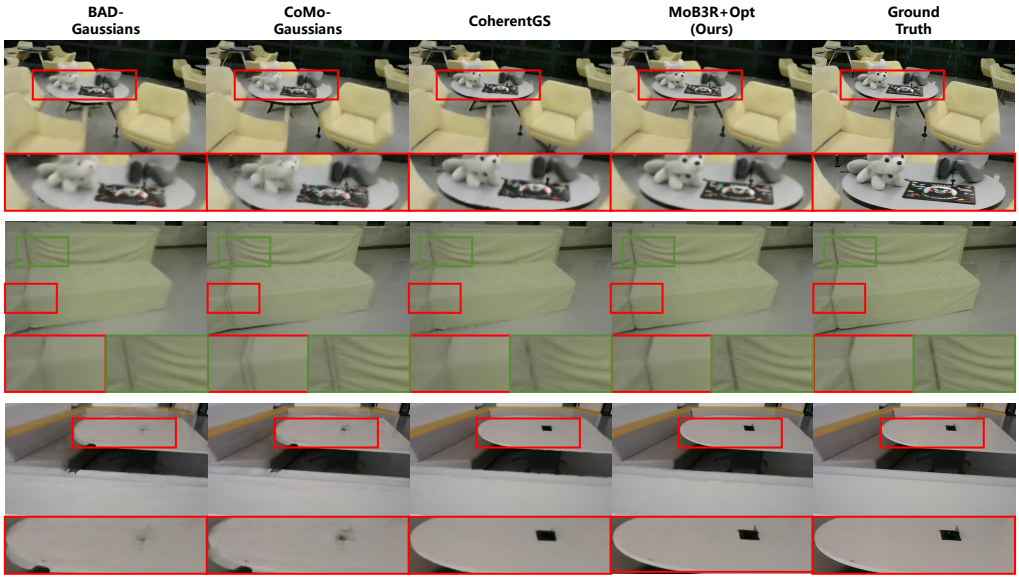

Figure 9: Additional qualitative comparison on DL3DV under sparse input views. BAD-Gaussians retains severe blur, while CoMo-Gaussian and CoherentGS lose fine details. MoB3R+Opt produces sharper boundaries and textures. Red and green boxes mark enlarged regions.

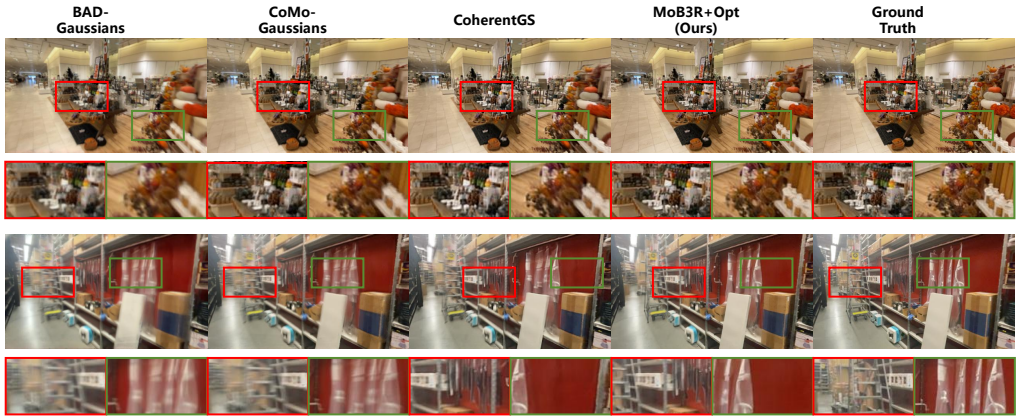

Figure 10: Additional comparison on DL3DV under sparse input views. MoB3R+Opt produces clearer leaf boundaries, shelf edges, and repeated textures than per-scene methods. Red and green boxes mark enlarged regions.

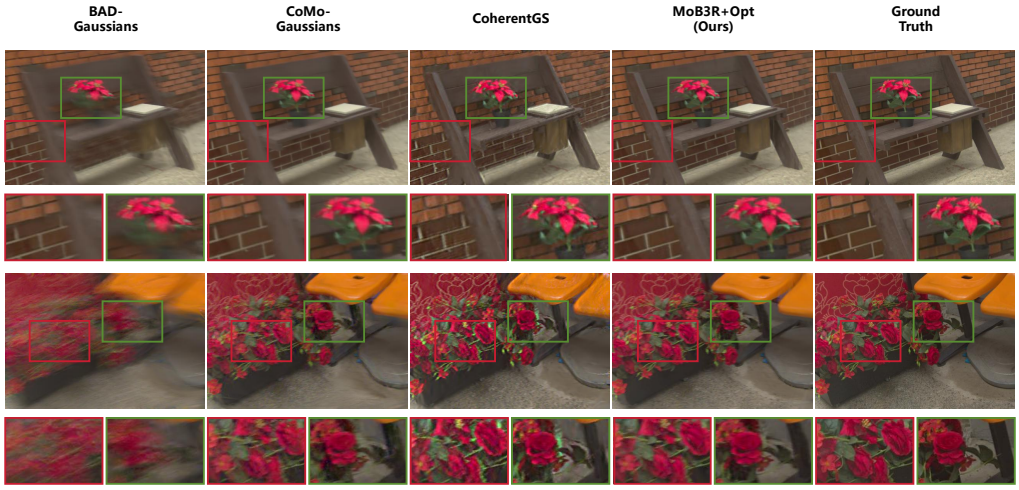

Figure 11: Qualitative comparison on ExBluRF under sparse input views. BAD-Gaussians retains strong blur, while CoMo-Gaussian and CoherentGS lose fine flower and table details. MoB3R+Opt produces sharper results. Red and green boxes mark enlarged regions.

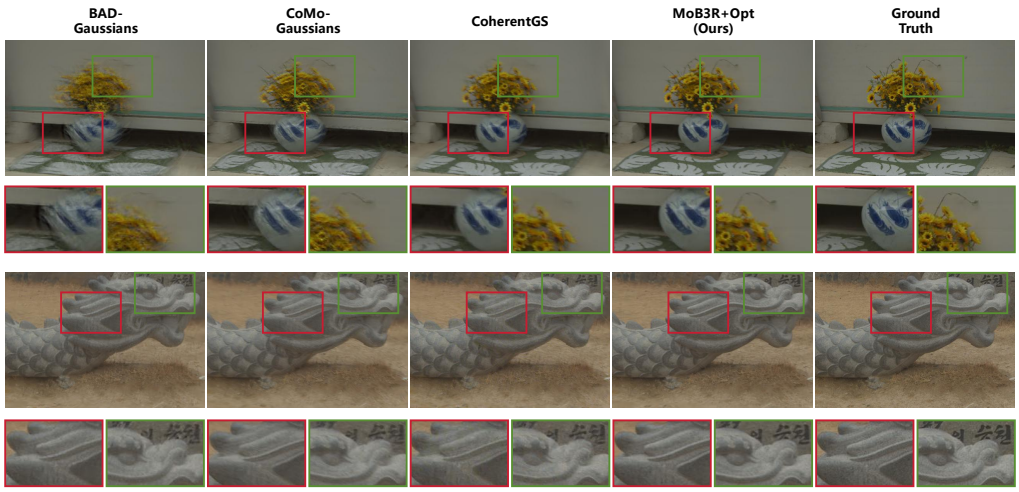

Figure 12: Additional qualitative comparison on ExBluRF. MoB3R+Opt recovers clearer edges and object details than the per-scene baselines. Red and green boxes mark enlarged regions.

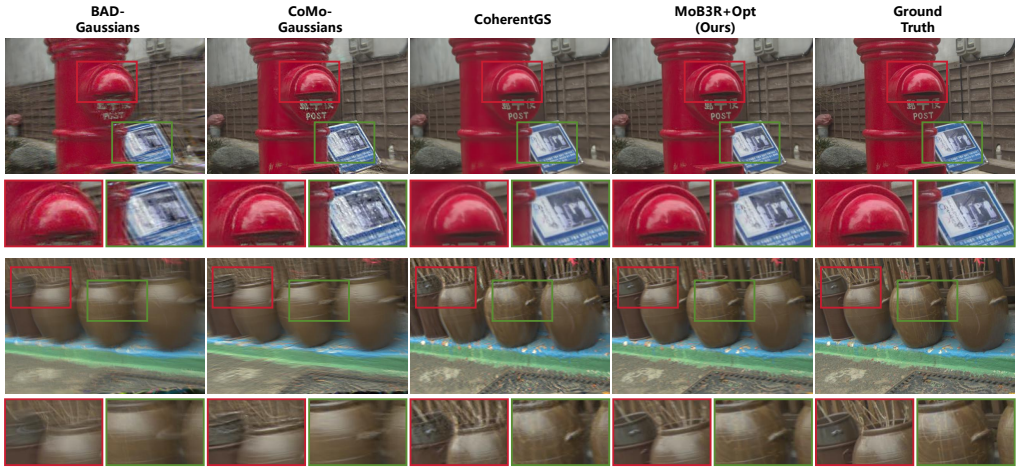

Figure 13: Additional qualitative comparison on ExBluRF. MoB3R+Opt produces sharper contours and textures, while the competing methods retain blur in the enlarged regions. Red and green boxes mark enlarged regions.

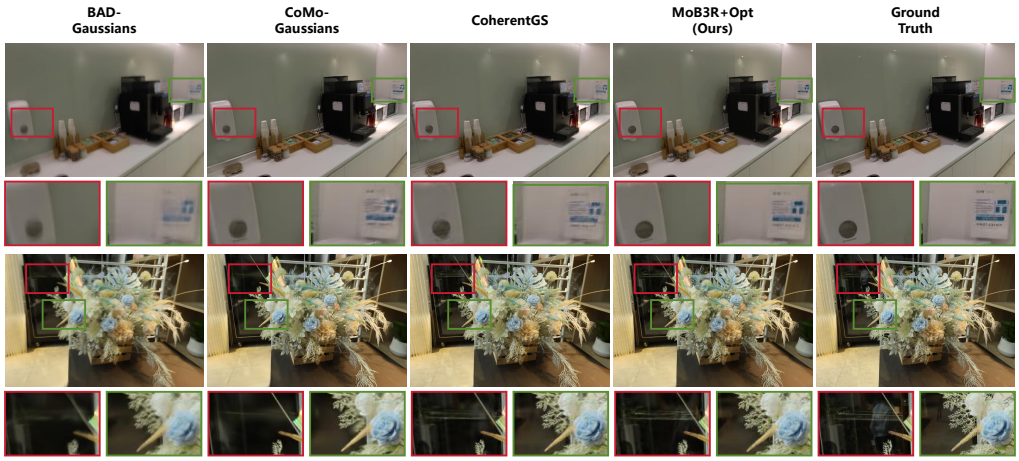

Figure 14: Qualitative comparison on DeblurNeRF with real motion-blurred inputs. MoB3R+Opt produces sharper object boundaries and fine details, while the competing methods retain blur or lose details. Red and green boxes mark enlarged regions.