No Pose, No Problem in 4D: Feed-Forward Dynamic Gaussians from Unposed Multi-View Videos

Matteo Balice[1] **Yanik Künzi**[2] **Chenyangguang Zhang**[2] **Matteo Matteucci**[1]
Marc Pollefeys[2] **Sungwhan Hong**[2,3†]

[1]Politecnico di Milano, [2]ETH Zürich [3]ETH AI Center

https://bralani.github.io/nopo4d_html/

Abstract

Recent feed-forward 3D gaussian splatting methods have made dramatic progress on individual aspects of 3D scene reconstruction, but no existing method jointly addresses dynamic content, multi-view input, and unknown camera poses in a single feed-forward pass. Methods that handle dynamics either require accurate camera poses or accept only monocular input; pose-free multi-view methods address only static scenes; and per-scene optimization methods bridge some of these gaps but at minutes-to-hours cost per scene. We introduce NoPo4D, the first feed-forward system that addresses this empty quadrant. Building on a pretrained geometry backbone and recent 4D Gaussian frameworks, NoPo4D introduces a velocity decomposition that splits Gaussian motion into per-pixel image-plane shifts and depth changes, allowing direct supervision from pseudo ground-truth optical flow on the 2D component. This sidesteps both the differentiable rendering that couples prior posed methods to pose accuracy and the 3D motion ground truth that prior pose-free methods require. The system is rounded out by a bidirectional motion encoder for cross-view and cross-frame feature aggregation, and view-dependent opacity that mitigates cross-view and cross-timestep Gaussian misalignments. On four multi-view dynamic benchmarks, NoPo4D consistently outperforms prior feed-forward baselines, and with an optional post-optimization stage surpasses per-scene optimization methods, while running orders of magnitude faster.

1 Introduction

Reconstructing dynamic 3D scenes from a handful of time-synchronized video streams, without known camera poses, without per-scene optimization, and in a single forward pass, would bring free-viewpoint rendering of everyday activities to practical scale. Applications range from sports and performance capture with minimal infrastructure, to egocentric scene understanding [11, 75, 52], AR [6], robotics [23], to immersive content creation outside controlled studio environments. Yet achieving this in practice demands solving three problems simultaneously: modeling dynamic content, fusing information across viewpoints, and estimating geometry without known camera poses.

Recent feed-forward methods have achieved remarkable progress along individual axes of this problem. Static reconstruction from unposed multi-view images can now be performed in under a second [28, 41, 56, 54], while feed-forward 4D methods have begun to tackle dynamic scenes from monocular video [64, 68]. However, each existing approach relaxes only a subset of the target constraints, whether by requiring known camera poses, accepting only monocular input, restricting to specific domains such as driving, or relying on per-scene optimization that takes minutes to hours [62, 57]. As shown in Tab. 1, no existing feed-forward method jointly handles dynamic scenes from unposed, multi-view video of general environments.

†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: **Capability comparison of related methods.** NoPo4D is the first feed-forward method that jointly handles dynamic scenes from unposed, multi-view video of general environments. **FF**: feed-forward inference. **Unposed**: no ground-truth camera poses required. **MV**: supports multi-view input. **Dyn.**: models dynamic scenes. **General**: not restricted to a specific domain.

Method	Type	FF	Unposed	MV	Dyn.	General
NoPoSplat [71]	Static	✓	✓	✓	✗	✓
FLARE [77]	Static	✓	✓	✓	✗	✓
AnySplat [32]	Static	✓	✓	✓	✗	✓
4DGT [64]	Dyn.	✓	✗	✗	✓	✓
NeoVerse [68]	Dyn.	✓	✓	✗	✓	✓
DGGT [8]	Dyn.	✓	✓	✓	✓	✗
Shape of Motion [57]	Opt.	✗	✗	✗	✓	✓
MonoFusion [62]	Opt.	✗	✗	✓	✓	✓
NoPo4D (Ours)	**Dyn.**	✓	✓	✓	✓	✓

We present **NoPo4D**, the first feed-forward system that addresses this empty quadrant. Filling it requires solving three problems that none of the constituent settings face in isolation. First, motion supervision must operate without GT poses or 3D motion annotations, neither of which is available at scale. Prior feed-forward dynamic methods sidestep one constraint at the cost of the other: posed approaches supervise through differentiable rendering [64], while pose-free approaches supervise velocities directly against 3D motion ground truth [68]. We avoid both by decomposing Gaussian motion into per-pixel image-plane shifts and depth changes, supervising the image-plane component directly with pseudo ground-truth optical flow. Second, motion must remain consistent across both cameras and consecutive frames, which we address with a bidirectional motion encoder that performs joint self-attention across views and timesteps. Third, geometric inconsistencies in pose-free reconstruction now compound across cameras *and* timesteps, a more severe problem than either monocular dynamic or static multi-view methods face; we mitigate this with view-dependent opacity.

We evaluate NoPo4D on four multi-view dynamic benchmarks: ExoRecon [62], Immersive Light Field [4], Kubric [18], and N3DV [39]. Across all four, NoPo4D consistently outperforms prior feed-forward baselines. With optional post-optimization, NoPo4D surpasses per-scene optimization methods. Ablations validate that each design choice contributes meaningfully. In summary, our contributions are as follows:

- We identify the setting of feed-forward, pose-free, multi-view dynamic scene reconstruction and show via systematic comparison that no prior method addresses it.
- We introduce NoPo4D, the first feed-forward system that operates in this setting. Our key technical contribution is a velocity decomposition that splits Gaussian motion into per-pixel image-plane shifts and depth changes, allowing optical flow to supervise the image-plane component directly. We complement this with a bidirectional motion encoder that aggregates features across views and frames, and adopt view-dependent opacity to handle cross-view and cross-timestep Gaussian misalignments.
- We evaluate on four multi-view dynamic benchmarks, demonstrating that NoPo4D consistently outperforms feed-forward baselines and, with optional post-optimization, surpasses per-scene optimization methods while running orders of magnitude faster.

2 Related Work

Static Feed-Forward Reconstruction. The DUSt3R family [60, 38] pioneered feed-forward 3D reconstruction by regressing dense point maps from image pairs, bypassing classical multi-stage pipelines. This idea was extended to dynamic point maps [76, 21], incremental multi-view processing [55, 59], and parallel many-view inference [65]. VGGT [56] scaled the paradigm to hundreds of views with joint prediction of cameras, depth, and point tracks, while Depth Anything 3 (DA3) [41] showed that a plain DINOv2 [44] backbone with alternating attention achieves spatially consistent multi-view geometry with fewer parameters. A parallel thread predicts Gaussian primitives directly from images. PixelSplat [7] and MVSplat [9] demonstrated feed-forward Gaussian prediction

but require known poses. Removing this assumption, NoPoSplat [71] fine-tunes a MASt3R-style architecture for pose-free prediction, C3G [2] builds on top of the existing foundation models [56], and subsequent methods generalize to cascaded pose-geometry pipelines [77], unconstrained views [32, 50], variable-length sequences [10, 27], and unified pose-free and pose-dependent settings [72]. All these methods are restricted to static scenes.

Feed-Forward Dynamic Reconstruction. L4GM [48] proposed the first feed-forward 4D model but is limited to object-centric scenes with per-frame Gaussians. 4DGT [64] introduced temporally coherent 4D Gaussians with lifespan and velocity attributes, trained on monocular posed video. NeoVerse [68] and MoVieS [40] extend this to unposed and motion-aware monocular settings respectively, but both remain restricted to single-camera input. DGGT [8] operates in the unposed multi-view regime but is tailored to driving scenarios. UFO-4D [30] achieves unposed feed-forward 4D reconstruction from a stereo pair. NoPo4D is, to the best of our knowledge, the first method to jointly handle dynamic scenes, multi-view input, unknown camera poses, and general scene content in a single feed-forward pass.

Optimization-Based Dynamic Reconstruction. Per-scene optimization has been approached via canonical deformation fields [46, 45, 53] and explicit spatiotemporal factorizations [5, 16]. Within Gaussian splatting, one family attaches per-Gaussian motion models: deformation MLPs [70, 63], per-timestep parameters with rigidity constraints [43], sparse control points [29], motion scaffolds [37], and learned trajectories [51]. Another lifts Gaussians to native 4D primitives by conditioning position and opacity on time [69, 15]. For sparse-view input, Shape-of-Motion [57] proposes SE(3) motion bases, an idea also adopted by MonoFusion [62], which reconstructs each view independently with monocular depth priors [66, 67] before aligning them into a shared 4D representation. All these methods require known camera poses and minutes-to-hours of computation per scene.

3 Method

3.1 Problem Formulation

Consider C uncalibrated, time-synchronized video streams of a scene, given as images $\mathcal{I} = \left\{(\mathbf{I}_t^c)_{t=1}^T\right\}_{c=1}^C$, where $\mathbf{I}_t^c \in \mathbb{R}^{H\times W\times 3}$ is the frame from camera c at timestep t, and each stream contains T frames. We assume the cameras form a static rig, as in capture setups such as Ego-Exo4D [17]. NoPo4D jointly predicts: 1) per-camera intrinsics $\hat{\mathbf{K}}^c$ and extrinsics $(\hat{\mathbf{R}}^c, \hat{\mathbf{t}}^c)$, obtained by averaging the backbone's per-frame predictions across t for each camera c, and 2) a collection of G 4D Gaussians

$$\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \alpha_g, \mathbf{c}_g, \tau_g, l_g, \mathbf{v}_g^+, \mathbf{v}_g^-, \boldsymbol{\omega}_g^+, \boldsymbol{\omega}_g^-\right)_{g=1}^G, \tag{1}$$

where each Gaussian carries static attributes mean $\boldsymbol{\mu}_g$, covariance $\boldsymbol{\Sigma}_g$, opacity α_g, and color $\mathbf{c}_g$. Following [64, 68], dynamic attributes include temporal center τ_g, lifespan l_g, and forward/backward linear and angular velocities $\mathbf{v}_g^\pm, \boldsymbol{\omega}_g^\pm$ that permit asymmetric motion around the temporal center.

3.2 Model Architecture

Fig. 1 illustrates the architecture: a pretrained geometry backbone extracts multi-view features that are routed to frozen camera and depth heads for per-frame geometry and to trainable Gaussian and motion heads for the remaining static and dynamic attributes.

Feature Backbone. We employ the pretrained transformer backbone $\mathcal{T}$ from Depth Anything 3 [41]. Given input frames $\mathcal{I}$, the backbone processes them through a stack of frozen within-view self-attention layers. We inject sinusoidal temporal tokens encoding each frame's timestamp before the trainable alternating within-view and cross-view attention layers, providing the backbone with an explicit temporal signal. This yields multi-view features $\mathcal{F} = \left\{(\mathbf{F}_t^c)_{t=1}^T\right\}_{c=1}^C = \mathcal{T}(\mathcal{I})$, where $\mathbf{F}_t^c \in \mathbb{R}^{N\times D}$ has N tokens per frame and embedding dimension D. Frozen pretrained heads then predict per-frame depth maps $\hat{\mathbf{D}}_t^c \in \mathbb{R}_+^{H\times W}$, intrinsics $\hat{\mathbf{K}}_t^c$, and extrinsics $(\hat{\mathbf{R}}_t^c, \hat{\mathbf{t}}_t^c)$; under our static-rig assumption (Sec. 3.1) these are averaged across t to yield the per-camera estimates. Keeping these heads frozen prevents overfitting and empirically yields more stable training (see supplementary material).

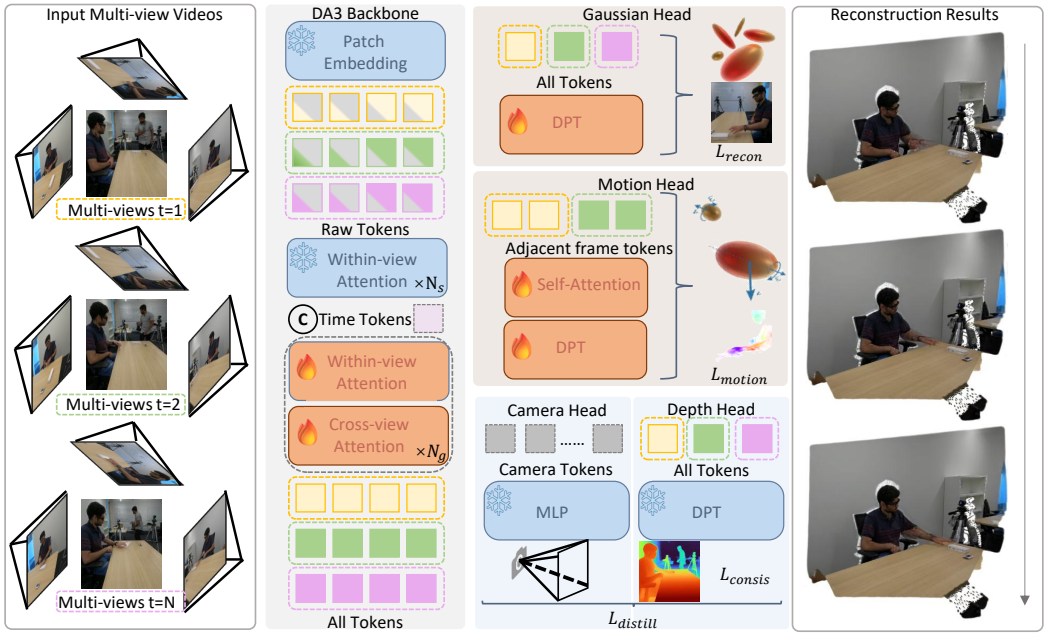


Figure 1: **Architecture overview.** Given C streams of time-synchronized video, DA3 [41] first extracts multi-view features through alternating within-view and cross-view attention layers. Pretrained, frozen depth and camera heads then recover per-frame geometry, which is unprojected into Gaussian means $\boldsymbol{\mu}$. Subsequently, two trainable heads decode the remaining attributes: a Gaussian head predicts static parameters $(\mathbf{R}, \mathbf{s}, \alpha, \mathbf{c})$, while a bidirectional motion branch processes consecutive-frame features to produce forward and backward linear and angular velocities $(\mathbf{v}^{\pm}, \boldsymbol{\omega}^{\pm})$, a per-pixel motion gate ρ, and the temporal covariance σ_g. All outputs are composed into a single 4D Gaussian representation and rendered via differentiable rasterization.

Gaussian Means. Each depth map $\hat{\mathbf{D}}_t^c$ is unprojected into a world-space point map $\hat{\mathbf{P}}_t^c \in \mathbb{R}^{H\times W\times 3}$ via the inverse projection Π^{-1} that lifts each pixel into a 3D point using its predicted depth and the camera parameters:

$$\hat{\mathbf{P}}_t^c = \Pi^{-1}(\hat{\mathbf{D}}_t^c, \hat{\mathbf{R}}^c, \hat{\mathbf{t}}^c, \hat{\mathbf{K}}^c). \tag{2}$$

We instantiate one Gaussian per pixel per frame: each Gaussian is uniquely indexed by $g = (c, t, x, y)$, and its mean $\boldsymbol{\mu}_g$ is set to the corresponding point-map entry $\hat{\mathbf{P}}_t^c(x, y)$.

Static Attributes. A DPT decoder [47] applied to the backbone features $\mathbf{F}_t^c$ predicts the remaining static parameters per pixel. The covariance $\boldsymbol{\Sigma}_g$ is parameterized by a unit quaternion $\mathbf{q}_g \in \mathbb{R}^4$ encoding rotation and a scale vector $\mathbf{s}_g$, following the standard 3DGS factorization [33]. Color $\mathbf{c}_g$ is decoded by the same head. We parameterize opacity with spherical harmonic coefficients $\alpha_g \in \mathbb{R}^{(k+1)^2}$ of order k rather than the scalar opacity used in standard 3DGS, making opacity view-dependent. This is critical in our pose-free multi-view dynamic setting: the SH coefficients act as a learned confidence metric that compensates for cross-view and cross-timestep Gaussian misalignments arising from imperfect geometry prediction, allowing unreliable Gaussians to become transparent from problematic viewpoints.

Decomposed Motion Parameterization. We decompose Gaussian motion into per-pixel image-plane shifts and a depth change, which allows direct supervision from 2D optical flow without requiring 3D motion ground truth or a pose-dependent rendering step. This contrasts with prior feed-forward dynamic methods [64, 68], which predict 3D Gaussian velocities directly. A DPT head decodes per-pixel 2D shifts $(\Delta x, \Delta y)$ and a depth displacement Δd, with pixel shifts bounded via $\tanh$ and scaled to image dimensions (W, H). For pixel (x, y) in frame t with depth $\hat{\mathbf{D}}_t^c(x, y)$, these define the projected location and depth at $t + 1$:

$$x' = x + \Delta x, \quad y' = y + \Delta y, \quad \hat{\mathbf{D}}_{t+1}^c(x, y) = \hat{\mathbf{D}}_t^c(x, y) + \Delta d. \tag{3}$$

Using the predicted intrinsics $\hat{\mathbf{K}}^c$ with focal lengths (f_x, f_y) and principal point (c_x, c_y), we unproject the source and displaced pixels into camera space and take their difference:

$$\Delta\mathbf{X}_{\text{cam}} = \begin{pmatrix} \dfrac{(x' - c_x)\,\hat{\mathbf{D}}^c_{t+1} - (x - c_x)\,\hat{\mathbf{D}}^c_t}{f_x} \\ \dfrac{(y' - c_y)\,\hat{\mathbf{D}}^c_{t+1} - (y - c_y)\,\hat{\mathbf{D}}^c_t}{f_y} \\ \hat{\mathbf{D}}^c_{t+1} - \hat{\mathbf{D}}^c_t \end{pmatrix}. \tag{4}$$

The forward linear velocity $\mathbf{v}_g^+$ for the Gaussian at this pixel is obtained by rotating into world coordinates and scaling by the temporal interval Δt:

$$\mathbf{v}_g^+ = \frac{\hat{\mathbf{R}}^c\,\Delta\mathbf{X}_{\text{cam}}}{\Delta t}. \tag{5}$$

The backward velocity $\mathbf{v}_g^-$ is computed symmetrically from features of the preceding frame.

Motion Encoder. The displacements above are produced by a bidirectional motion encoder $\mathcal{M}$. For each consecutive pair of frames $(t, t{+}1)$, tokens from all C cameras at frame t are concatenated with tokens from all C cameras at frame $t{+}1$, with two learnable embeddings tagging the source frame versus the neighbour. Joint self-attention over this combined set of $2CN$ tokens aggregates information across views and time. The resulting tokens are split back into forward and backward halves and decoded by a shared DPT head. Per Gaussian, the head outputs the 2D pixel shifts $(\Delta x, \Delta y)$ and depth displacement Δd above, angular velocities $\boldsymbol{\omega}_g^{\pm}$, a per-pixel motion gate $\rho_g \in (0, 1)$ that suppresses motion in static regions, a temporal covariance σ_g, and a temporal center τ_g anchored to the source frame's timestamp.

Post-Optimization. We include an optional test-time refinement stage following AnySplat [32]. After the feed-forward pass, we prune low-opacity Gaussians, render from input views, and minimize photometric losses between rendered and input images. Since the feed-forward output provides a strong initialization, convergence is rapid.

3.3 Training

Losses. We optimize using a weighted combination of objectives:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{motion}}\mathcal{L}_{\text{motion}} + \lambda_{\text{consis}}\mathcal{L}_{\text{consis}} + \mathcal{L}_{\text{distill}}. \tag{6}$$

The reconstruction loss $\mathcal{L}_{\text{recon}} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}$ is evaluated exclusively on rendered target frames. Since the model is pose-free, rendering target views during training requires aligning predicted poses into a common coordinate frame; we use a two-pass strategy with closed-form Sim(3) alignment (details in supplementary).

The motion loss supervises Gaussian dynamics by aligning the predicted pixel shifts $(\Delta x, \Delta y)$ from Eq. (3) with pseudo ground-truth optical flow $\mathbf{f}^*$ from SEA-RAFT [61]. A natural alternative is to render 2D flow from the predicted Gaussian velocities and compare it to $\mathbf{f}^*$, but rendering introduces dependence on Gaussian alignment and amplifies pose-induced errors. Since each Gaussian corresponds one-to-one with an input pixel, the predicted pixel shift directly encodes the source Gaussian's 2D motion, allowing us to bypass rendering entirely. We supervise only pixels with ground-truth shifts exceeding 2 pixels ($\Omega' = \{p \in \Omega \mid \|\mathbf{f}^*(p)\|_2 > 2\}$):

$$\lambda_{\text{motion}}\mathcal{L}_{\text{motion}} = \lambda_{\text{flow}}\frac{1}{|\Omega'|}\sum_{p\in\Omega'}\left\|(\Delta x, \Delta y)(p) - \mathbf{f}^*(p)\right\|_1. \tag{7}$$

The depth consistency loss ensures rendered depth matches backbone priors:

$$\lambda_{\text{consis}}\mathcal{L}_{\text{consis}} = \lambda_{\text{consis}}\frac{1}{|\Omega|}\sum_{p\in\Omega}\|\hat{\mathbf{D}}(p) - \mathbf{D}(p)\|_2^2. \tag{8}$$

To preserve pretrained knowledge, we distill from the frozen DA3 Giant model using pseudo-labels $\{\mathbf{T}^*, \mathbf{D}^*, \nabla\mathbf{D}^*\}$:

$$\mathcal{L}_{\text{distill}} = \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}}, \tag{9}$$

where $\mathcal{L}_{\text{pose}}$ is a Huber loss between predicted and teacher poses, and $\mathcal{L}_{\text{depth}} = \lambda_{\text{depth}}\|\mathbf{D} - \mathbf{D}^*\|_2^2$ and $\mathcal{L}_{\text{normal}} = \lambda_{\text{normal}}\|\nabla\mathbf{D} - \nabla\mathbf{D}^*\|_2^2$.

Training Curriculum. We train in two stages. The first stage trains on single-camera viewpoints, optimizing static scene representations and initializing the motion heads. The second stage extends to multi-camera batches, sampling frames distributed across multiple synchronized cameras. Within the second stage, we apply a curriculum on the temporal stride between context frames, linearly increasing it from a small initial value. Starting with closely spaced frames allows the model to reliably learn from optical flow supervision before tackling wider temporal baselines.

4 Experiments

4.1 Implementation Details

We train on ~2,900 Ego-Exo4D [17] scenes, with frames undistorted and downsampled to 796×448 (maximum 448 on the longer side); following AnySplat [32], we randomize aspect ratios in $[0.5, 1.0]$ and apply random center-cropping (77–100%) and horizontal flipping. Training runs in two stages of 20,000 steps each: single-camera viewpoints, then 16-frame batches across 1–4 cameras with temporal stride sampled from 4–8 (warmed up over the first 2,000 steps). We use Depth Anything 3 Large as backbone. We use AdamW [42] with learning rates 5×10^{-6} (backbone alternating attention), 5×10^{-5} (motion encoder), and 1×10^{-4} (velocity DPT and Gaussian head), with 1,000-step linear warmup and cosine annealing to $\frac{1}{10}$ of the initial rate. Loss weights are $\lambda_{\text{MSE}} = 1$, $\lambda_{\text{LPIPS}} = 0.5$, $\lambda_{\text{SSIM}} = 0.05$, $\lambda_{\text{consis}} = 0.1$, $\lambda_{\text{flow}} = 0.02$, $\lambda_{\text{pose}} = 1.0$, $\lambda_{\text{depth}} = 0.1$, $\lambda_{\text{normal}} = 0.1$. Post-optimization runs 100 steps after pruning Gaussians with opacity below 0.01, taking under 5 seconds per scene. Training uses batch size 1 per GPU at 448×448 on 4 NVIDIA GH200s.

4.2 Experimental Setup

Datasets and protocol. We evaluate on four multi-view dynamic benchmarks. *ExoRecon* [62] provides sparse multi-view captures derived from Ego-Exo4D and serves as our in-distribution benchmark; we use the same test scenes as ExoRecon to allow direct comparison with MonoFusion. *Immersive Light Field* [4] contains real-world dynamic scenes from a multi-camera rig, *Kubric* [18] is a synthetic dataset with rendered scenes of varying complexity, and *N3DV* [39] provides multi-view dynamic videos with denser camera coverage that allow evaluation under varying camera counts and from novel viewpoints. The latter three are out-of-distribution relative to our Ego-Exo4D training data, allowing us to assess generalization. We adopt a chunk-based evaluation protocol on all four benchmarks: each scene is partitioned into chunks of 5 consecutive timestamps, where 4 surrounding frames serve as input context and the middle frame is held out as the target. Each method synthesizes the target frame at each input camera viewpoint, and we report PSNR, SSIM, and LPIPS averaged across cameras, chunks, and test scenes. We cap evaluation at 60 chunks per scene for ExoRecon and at 100 chunks per scene for the others. On N3DV, we additionally evaluate from novel viewpoints at moderate ($5.7°$–$27.7°$) and extreme ($34.6°$–$71.9°$) rotations from the input cameras.

Baselines. We compare NoPo4D against feed-forward and per-scene optimization baselines. The feed-forward baselines include *NeoVerse* [68] and *MoVieS* [40], both originally designed for monocular video; we extend them to multi-view input by independently reconstructing each camera stream and aligning per-view outputs into a shared coordinate frame. *DGGT* [8] is the closest existing method, operating in the unposed multi-view dynamic regime. The optimization baselines include *MonoFusion* [62] and *MV-SOM* [58] and *Dyn3D-GS* [43]; all require known camera poses, in contrast to NoPo4D. For fair comparison on ExoRecon, we retrain DGGT on the same Ego-Exo4D-derived training data as NoPo4D, eliminating any distribution-shift advantage. The remaining feed-forward baselines are designed for monocular input and cannot be straightforwardly retrained on the multi-view setup; we evaluate them as released. On Immersive Light Field, Kubric, and N3DV, NoPo4D operates outside its training distribution. Note that NeoVerse was trained on Kubric and is therefore in-distribution on that benchmark; all other feed-forward baselines are also out-of-distribution on these test sets.

4.3 Quantitative Evaluation

In-distribution performance on ExoRecon. Table 2 reports performance on ExoRecon, our in-distribution benchmark. To enable a fair comparison, we also retrain DGGT [8] on the same Ego-Exo4D-derived training data as ours. Overall, we find that NoPo4D substantially outperforms

other feed-forward baselines. With an optional 100-step post-optimization, NoPo4D surpasses the per-scene optimization SOTA MonoFusion, and further broadens the gap. This demonstrates that a strong feed-forward initialization, combined with brief refinement, can match or exceed methods that rely on full per-scene optimization with known calibration.

Table 2: **Held-out view synthesis on ExoRecon [62]. FF** denotes feed-forward inference.

Method	FF	PSNR ↑	SSIM ↑	LPIPS ↓
Dyn3D-GS [43]	✗	24.28	0.692	0.539
MV-SOM [58]	✗	26.91	0.890	0.138
MonoFusion [62]	✗	30.43	**0.930**	**0.060**
NoPo4D (Ours, post-opt.)	✗	**31.95**	0.928	0.075
MoVies [40]	✓	18.78	0.725	0.288
DGGT [8]	✓	19.67	0.645	0.417
DGGT (EgoExo4D) [8]	✓	20.44	0.704	0.418
NeoVerse [68]	✓	20.03	0.714	0.354
NoPo4D (Ours)	✓	**29.15**	**0.886**	**0.125**

Generalization across distributions. Tables 3 and 4 report zero-shot performance on three benchmarks outside NoPo4D's Ego-Exo4D training distribution. All feed-forward baselines, except for Neoverse, also operate out-of-distribution on these test sets, so the gaps reflect how well each model's learned priors transfer; the comparison to DGGT, trained on the same data as NoPo4D, further isolates the architectural contribution. NoPo4D consistently outperforms the strongest feed-forward baseline across all three benchmarks. On N3DV in particular, it leads by nearly +4 PSNR from input cameras and remains competitive at moderate ($5.7°$–$27.7°$) novel viewpoints. At extreme ($34.6°$–$71.9°$) viewpoints, pixel-aligned metrics begin to misrepresent perceptual quality, rewarding methods that hedge with blurry, low-confidence predictions over those that produce sharper but slightly-shifted output. As shown in Fig. 3, NoPo4D produces visually cleaner reconstructions than DGGT under such viewpoints despite a small numerical gap, with NeoVerse and MoVieS showing more severe artifacts. These results indicate that NoPo4D's design transfers across capture modalities (egocentric, real multi-camera, synthetic, and dense multi-view) without any domain-specific tuning. Additional analysis of robustness to varying input view counts is provided in the supplementary material.

Table 3: **Generalization to out-of-distribution datasets.** NoPo4D was not trained on either benchmark. Note that NeoVerse was trained on Kubric and is therefore in-distribution on that benchmark.

	Immersive Light Field [4]			Kubric [18]		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MoVies [40]	16.12	0.650	0.410	14.49	0.720	0.260
NeoVerse [68]	19.83	0.680	0.360	16.86	0.500	0.520
DGGT [8]	20.65	0.700	0.360	20.14	0.530	0.430
DGGT (EgoExo4D) [8]	20.23	0.685	0.453	18.22	0.551	0.522
NoPo4D (Ours)	21.63	0.740	0.280	23.61	0.739	0.256
NoPo4D (Ours, post-opt.)	**24.04**	**0.817**	**0.245**	**28.17**	**0.838**	**0.159**

Table 4: **View synthesis on N3DV [39].** We report under three viewpoint configurations: original input cameras, moderate novel views ($5.7°$–$27.7°$ rotation), and extreme novel views ($34.6°$–$71.9°$ rotation).

	Input Cameras			Novel Cameras (Moderate)			Novel Cameras (Extreme)		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MoVies [40]	14.10	0.556	0.505	15.49	0.457	0.502	10.42	0.156	0.492
NeoVerse [68]	24.50	0.789	0.313	25.34	0.833	0.274	17.11	0.659	0.221
DGGT [8]	23.08	0.749	0.268	22.92	0.735	0.260	17.12	0.590	0.212
DGGT (EgoExo4D) [8]	22.49	0.713	0.356	23.10	0.751	0.239	16.82	0.582	0.206
NoPo4D (Ours)	28.43	0.905	0.144	26.79	0.899	0.146	**19.36**	**0.771**	**0.130**
NoPo4D (Ours, post-opt.)	**31.79**	**0.956**	**0.089**	**27.66**	**0.909**	**0.102**	19.18	0.759	0.132

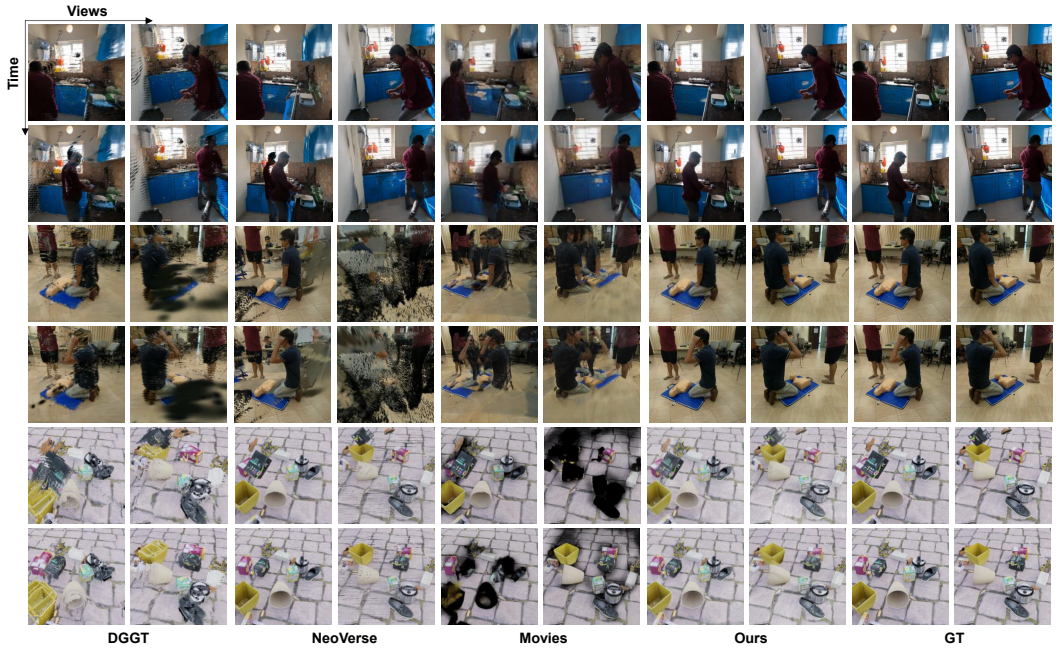


Figure 2: **Qualitative comparison on ExoRecon [62] and Kubric [18].**



Figure 3: **Qualitative comparison on N3DV [39] under extreme viewpoint changes.**

4.4 Qualitative Evaluation

Figures 2 and 3 compare rendering results across feed-forward methods. NoPo4D recovers fine-scale structure with sharp dynamic foregrounds, while baselines exhibit blurring, geometric distortion, or missing motion. At extreme viewpoints, DGGT shows semi-transparent ghosting and NeoVerse and MoVieS over-smooth, while NoPo4D preserves recognizable geometry and texture, supporting our reading of Table 4: pixel-aligned metrics under-credit confident-but-shifted predictions relative to blurry-but-aligned ones. We provide video comparisons in the supplementary material.

4.5 Ablation Study

We conduct ablations to validate each design choice on ExoRecon. Architectural components are isolated in Table 5a, auxiliary losses in Table 5b. Backbone fine-tuning strategies are analyzed in the supplementary material.

Architectural components. Table 5a isolates four architectural choices. Removing the bidirectional motion encoder $\mathcal{M}$ and feeding raw backbone tokens directly to the velocity DPT head (*No motion branch*) drops performance by 6.2 PSNR, confirming that explicit cross-frame feature aggregation is essential for predicting consistent motion. Replacing view-dependent opacity with a single scalar (*No opacity SH*) causes a comparable 7.5 PSNR drop, validating that opacity SHs serve as a learned confidence metric, masking misaligned Gaussians from viewpoints where the underlying geometry disagrees across cameras. Replacing the 2D pixel-shift + depth decomposition with direct 3D velocity prediction (*No decomposition*) drops 1.5 PSNR. In this configuration, optical-flow supervision must be applied through differentiable rendering rather than directly on the predicted 2D shifts. The drop confirms that the decomposed parameterization, by allowing pixel-level flow supervision without

Table 5: **Ablation studies on ExoRecon.** *Left:* architectural component ablation, where each row removes or replaces one component of the full NoPo4D pipeline. *Right:* auxiliary loss ablation, where each column indicates whether a loss is included (✓) or removed (✗) from the training objective; the reconstruction loss $\mathcal{L}_{\text{recon}}$ is always active.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
No opacity SH	21.61	0.758	0.247
No motion branch	22.97	0.856	0.161
No decomposition	27.67	0.865	0.129
No temporal encoding	27.69	0.861	**0.125**
NoPo4D (Ours)	**29.15**	**0.886**	**0.125**

(a) Architectural components.

Recon.	Distill.	Consis.	Flow	PSNR ↑	SSIM ↑	LPIPS ↓
✓	✗	✗	✗	25.09	0.829	0.162
✓	✗	✓	✗	3.51	0.238	0.803
✓	✗	✓	✓	25.91	0.815	0.159
✓	✓	✗	✓	28.02	0.867	**0.122**
✓	✓	✓	✗	28.19	0.869	0.125
✓	✓	✓	✓	**29.15**	**0.886**	0.125

(b) Auxiliary losses.

rendering, provides a cleaner training signal than the rendered alternative. Removing both the decomposition and the optical-flow supervision simultaneously (*No decomposition + no flow loss*) causes training to fail outright: predicted velocity magnitudes grow without bound and the loss diverges within the first few thousand steps, indicating that the decomposition and flow loss are jointly necessary to constrain motion predictions in our pose-free multi-view setting. Finally, removing the sinusoidal temporal tokens from the backbone (*No temporal encoding*) drops 1.7 PSNR, indicating that the alternating attention layers benefit from an explicit temporal signal when aggregating features across frames.

Auxiliary losses. Table 5b ablates the three auxiliary losses on top of reconstruction. Two findings stand out. First, the consistency loss alone is catastrophic (recon + consis only, 3.51 PSNR). Without geometric grounding from distillation or flow, consistency constrains rendered depth to match backbone-predicted depth but provides no signal anchoring the geometry to the target; the model converges to a degenerate solution where Gaussians collapse together, satisfying the constraint trivially. This confirms that consistency functions as a regularizer, not a primary supervisory signal. Second, the leave-one-out comparisons reveal complementary roles: removing distillation causes the largest drop (-3.24 PSNR), as the alternating attention layers drift away from the pretrained DA3 representations; removing flow (-0.96 PSNR) costs the model dynamic awareness, with predicted velocities collapsing toward zero; and removing consistency (-1.13 PSNR) allows multi-view geometric disagreements to persist. The full configuration achieves the best result by combining all three, each addressing a complementary failure mode: distillation preserves priors, flow grounds dynamics, and consistency enforces multi-view coherence.

5 Conclusion

We presented NoPo4D, a feed-forward system that fills a previously empty quadrant in 4D scene reconstruction: dynamic scenes captured by multiple synchronized cameras with unknown poses, in general environments, reconstructed in a single forward pass. The system rests on three design choices: a velocity decomposition that allows direct optical-flow supervision without rendering, a bidirectional motion encoder that aggregates features across views and time, and view-dependent opacity that mitigates the cross-view and cross-timestep Gaussian misalignments inherent to pose-free multi-view dynamic capture. Across four multi-view dynamic benchmarks NoPo4D consistently outperforms feed-forward baselines, and with brief post-optimization surpasses per-scene optimization methods that require known poses while running orders of magnitude faster.

Limitations and future work. NoPo4D assumes a static camera rig, which holds for capture setups such as Ego-Exo4D but degrades in quality when cameras themselves move. The model relies on pseudo ground-truth signals from pretrained models; as these teachers improve, NoPo4D benefits automatically without architectural changes, but training quality is bounded by their accuracy. The model also requires multiple synchronized streams; extending the same parameterization to monocular or asynchronous capture remains open. Finally, the velocity decomposition is currently first-order in time; capturing higher-order motion within the same feed-forward framework is a natural next direction.

Acknowledgement This work was supported by the Swiss AI Initiative through a grant from the Swiss National Supercomputing Centre (CSCS) under project ID ab036 on the Alps infrastructure; the Swiss National Science Foundation (Advanced Grant 216260: "Beyond Frozen Worlds: Capturing Functional 3D Digital Twins from the Real World"); the European Union's Horizon Europe research and innovation programme under grant agreement No. 101214398 (ELLIOT); the Horizon Europe programme under grant agreement No. 101298421 (GRAIL); and the ETH AI Center through an ETH AI Center postdoc fellowship and an ETH Zurich Career Seed Award awarded to Sunghwan Hong.

References

[1] Honggyu An, Jinhyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-View Completion Models are Zero-shot Correspondence Estimators, December 2024. URL http://arxiv.org/abs/2412.09072. arXiv:2412.09072.

[2] Honggyu An, Jaewoo Jung, Mungyeom Kim, Sunghwan Hong, Chaehyun Kim, Kazumi Fukuda, Minkyeong Jeon, Jisang Han, Takuya Narihira, Hyuna Ko, et al. C3g: Learning compact 3d representations with 2k gaussians. *arXiv preprint arXiv:2512.04021*, 2025.

[3] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-Controlled Generative Rendering from A Single Video, March 2025. URL http://arxiv.org/abs/2503.11647. arXiv:2503.11647 version: 1.

[4] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics*, 39(4), August 2020. ISSN 0730-0301, 1557-7368. doi: 10.1145/3386569.3392485. URL https://dl.acm.org/doi/10.1145/3386569.3392485.

[5] Ang Cao and Justin Johnson. HexPlane: A Fast Representation for Dynamic Scenes, March 2023. URL http://arxiv.org/abs/2301.09632. arXiv:2301.09632.

[6] Ata Çelen, Jaewoo Jung, Federico Tombari, Marc Pollefeys, Sunghwan Hong, Michael Niemeyer, and Daniel Barath. Genrec: Knowing where to reconstruct and where to generate. *arXiv preprint arXiv:2608.17832*, 2026.

[7] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction, December 2023. URL http://arxiv.org/abs/2312.12337. arXiv:2312.12337.

[8] Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, BING WANG, Guang Chen, et al. Feedforward 4d reconstruction for dynamic driving scenes using unposed images.

[9] 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, March 2024. URL http://arxiv.org/abs/2403.14627. arXiv:2403.14627.

[10] Zequn Chen, Jiezhi Yang, and Heng Yang. Pref3r: Pose-free feed-forward 3d gaussian splatting from variable-length image sequence. *arXiv preprint arXiv:2411.16877*, 2024.

[11] Zhiang Chen, Nahyuk Lee, Boyang Sun, Taein Kwon, Marc Pollefeys, Zuria Bauer, and Sunghwan Hong. Prose: Training-free egocentric scene registration with vision-language models. *arXiv preprint arXiv:2606.16569*, 2026.

[12] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021.

[13] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7174–7194, 2022.

[14] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.

[15] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4D-Rotor Gaussian Splatting: Towards Efficient Novel View Synthesis for Dynamic Scenes, July 2024. URL `http://arxiv.org/abs/2402.03307`. arXiv:2402.03307 [cs].

[16] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance, March 2023. URL `http://arxiv.org/abs/2301.10241`. arXiv:2301.10241.

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives, September 2024. URL `http://arxiv.org/abs/2311.18259`. arXiv:2311.18259.

[18] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti, Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator, March 2022. URL `http://arxiv.org/abs/2203.03570`. arXiv:2203.03570.

[19] Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sunghwan Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models. *arXiv preprint arXiv:2604.08456*, 2026.

[20] A Said Gurbuz, Sunghwan Hong, Ahmed Nassar, Marc Pollefeys, and Peter Staar. Moving beyond sparse grounding with complete screen parsing supervision. *arXiv preprint arXiv:2602.14276*, 2026.

[21] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. D^2USt3R: Enhancing 3D Reconstruction with 4D Pointmaps for Dynamic Scenes, April 2025. URL `http://arxiv.org/abs/2504.06264`. arXiv:2504.06264.

[22] Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. *arXiv preprint arXiv:2512.04012*, 2025.

[23] Jisang Han, Seonghu Jeon, Jaewoo Jung, René Zurbrügg, Honggyu An, Tifanny Portela, Marco Hutter, Marc Pollefeys, Seungryong Kim, and Sunghwan Hong. Geometric action model for robot policy learning. *arXiv preprint arXiv:2606.17046*, 2026.

[24] Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9907–9917, 2021.

[25] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022.

[26] Sunghwan Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seungryong Kim. Neural matching fields: Implicit representation of matching fields for visual correspondence. *Advances in Neural Information Processing Systems*, 35:13512–13526, 2022.

[27] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. PF3plat: Pose-Free Feed-Forward 3D Gaussian Splatting, October 2024. URL `http://arxiv.org/abs/2410.22128`. arXiv:2410.22128.

[28] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, and Chong Luo. Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20196–20206, 2024.

[29] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4220–4230, 2024.

[30] Junhwa Hur, Charles Herrmann, Songyou Peng, Philipp Henzler, Zeyu Ma, Todd Zickler, and Deqing Sun. Ufo-4d: Unposed feedforward 4d reconstruction from two images. *arXiv preprint arXiv:2602.24290*, 2026.

[31] Byungwoo Jeon, Sunghwan Hong, Dohyeon Kim, Jonghoon Lee, Jungwoo Park, Inhee Lee, Doosan Baek, Jinil Kim, Taehyun Lim, Hanbyul Joo, et al. Learning the 4d world through point trajectory forecasting. *arXiv preprint arXiv:2610.28594*, 2026.

[32] 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, May 2025. URL `http://arxiv.org/abs/2505.23716`. arXiv:2505.23716.

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering, August 2023. URL `http://arxiv.org/abs/2308.04079`. arXiv:2308.04079.

[34] Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. *arXiv preprint arXiv:2509.18096*, 2025.

[35] JoungBin Lee, Jaewoo Jung, Jisang Han, Takuya Narihira, Kazumi Fukuda, Junyoung Seo, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. 3d scene prompting for scene-consistent camera-controllable video generation. *arXiv preprint arXiv:2510.14945*, 2025.

[36] Nahyuk Lee, Zhiang Chen, Marc Pollefeys, and Sunghwan Hong. Tora: Topological representation alignment for 3d shape assembly. *arXiv preprint arXiv:2604.04050*, 2026.

[37] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds, November 2024. URL `http://arxiv.org/abs/2405.17421`. arXiv:2405.17421.

[38] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R, June 2024. URL `http://arxiv.org/abs/2406.09756`. arXiv:2406.09756 [cs].

[39] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5521–5531, 2022.

[40] Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Tao Hu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. Movies: Motion-aware 4d dynamic view synthesis in one second. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

[41] 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, November 2025. URL http://arxiv.org/abs/2511.10647. arXiv:2511.10647.

[42] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, November 2017. URL https://arxiv.org/abs/1711.05101v3.

[43] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis, August 2023. URL http://arxiv.org/abs/2308.09713. arXiv:2308.09713 [cs].

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, April 2023. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

[45] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields, September 2021. URL http://arxiv.org/abs/2011.12948. arXiv:2011.12948.

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes, November 2020. URL http://arxiv.org/abs/2011.13961. arXiv:2011.13961.

[47] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction, March 2021. URL http://arxiv.org/abs/2103.13413. arXiv:2103.13413.

[48] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, and Huan Ling. L4GM: Large 4D Gaussian Reconstruction Model, June 2024. URL http://arxiv.org/abs/2406.10324. arXiv:2406.10324.

[49] Heeseong Shin, Chaehyun Kim, Sunghwan Hong, Seokju Cho, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Towards open-vocabulary semantic segmentation without semantic labels. *Advances in Neural Information Processing Systems*, 37:9153–9177, 2024.

[50] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3R: Zero-shot Gaussian Splatting from Uncalibrated Image Pairs, August 2024. URL http://arxiv.org/abs/2408.13912. arXiv:2408.13912.

[51] 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.

[52] Boyang Sun, Jiajie Li, Yung-Hsu Yang, Chenyangguang Zhang, Tim Engelbracht, Sunghwan Hong, Cesar Cadena, Marc Pollefeys, and Hermann Blum. Lime: Learning intent-aware camera motion from egocentric video. *arXiv preprint arXiv:2607.02417*, 2026.

[53] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video, August 2021. URL http://arxiv.org/abs/2012.12247. arXiv:2012.12247.

[54] Alexander Veicht, Sunghwan Hong, Dániel Baráth, and Marc Pollefeys. Zipsplat: Fewer gaussians, better splats. *arXiv preprint arXiv:2606.05102*, 2026.

[55] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025.

[56] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer, March 2025. URL http://arxiv.org/abs/2503.11651. arXiv:2503.11651.

[57] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of Motion: 4D Reconstruction from a Single Video, July 2024. URL http://arxiv.org/abs/2407.13764. arXiv:2407.13764 [cs].

[58] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *International Conference on Computer Vision (ICCV)*, 2025.

[59] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[60] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy, December 2023. URL http://arxiv.org/abs/2312.14132. arXiv:2312.14132.

[61] Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow, May 2024. URL http://arxiv.org/abs/2405.14793. arXiv:2405.14793.

[62] Zihan Wang, Jeff Tan, Tarasha Khurana, Neehar Peri, and Deva Ramanan. MonoFusion: Sparse-View 4D Reconstruction via Monocular Fusion, July 2025. URL http://arxiv.org/abs/2507.23782. arXiv:2507.23782 [cs].

[63] 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, July 2024. URL http://arxiv.org/abs/2310.08528. arXiv:2310.08528 [cs].

[64] Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv. 4DGT: Learning a 4D Gaussian Transformer Using Real-World Monocular Videos, June 2025. URL http://arxiv.org/abs/2506.08015. arXiv:2506.08015 [cs].

[65] 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, January 2025. URL http://arxiv.org/abs/2501.13928. arXiv:2501.13928.

[66] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data, April 2024. URL http://arxiv.org/abs/2401.10891. arXiv:2401.10891.

[67] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2, June 2024. URL http://arxiv.org/abs/2406.09414. arXiv:2406.09414.

[68] Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4D World Model with in-the-wild Monocular Videos, January 2026. URL http://arxiv.org/abs/2601.00393. arXiv:2601.00393.

[69] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting, February 2024. URL http://arxiv.org/abs/2310.10642. arXiv:2310.10642 [cs].

[70] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction, November 2023. URL http://arxiv.org/abs/2309.13101. arXiv:2309.13101 [cs].

[71] 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, October 2024. URL http://arxiv.org/abs/2410.24207. arXiv:2410.24207.

[72] Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. Yonosplat: You only need one model for feedforward 3d gaussian splatting. *arXiv preprint arXiv:2511.07321*, 2025.

[73] Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.

[74] Yuanwen Yue, Damien Robert, Jianyuan Wang, Sunghwan Hong, Jan Dirk Wegner, Christian Rupprecht, and Konrad Schindler. Litept: Lighter yet stronger point transformer. *arXiv preprint arXiv:2512.13689*, 2025.

[75] Chenyangguang Zhang, Malgorzata Gwiazda, Guanlong Jiao, Yuanchen Ju, Federico Tombari, Koushil Sreenath, Marc Pollefeys, and Sunghwan Hong. Chronograph: Functional 4d scene graphs with vision-language models for interaction understanding and grounded planning. *arXiv preprint arXiv:2609.39665*, 2026.

[76] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion, October 2024. URL http://arxiv.org/abs/2410.03825. arXiv:2410.03825.

[77] 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, February 2025. URL http://arxiv.org/abs/2502.12138. arXiv:2502.12138.

Supplementary Material

This document provides additional analysis supporting the main paper. §A compares different backbone fine-tuning strategies. §B details the two-pass pose alignment used for target-view rendering. §C analyzes robustness to varying input view counts. §D presents additional qualitative comparisons across all four benchmarks. §E documents failure cases. §F discusses broader impacts.

A Backbone Fine-Tuning Strategy

The main paper notes that NoPo4D keeps the backbone's depth and camera prediction heads frozen while fine-tuning only the alternating attention layers. Here we justify this design by comparing against alternative fine-tuning configurations on ExoRecon.

Table 6: **Comparison of backbone fine-tuning strategies on ExoRecon.** Each row varies which components of the pretrained backbone are unfrozen during training. Fine-tuning all 16 alternating attention layers, while keeping the depth and camera heads frozen, achieves the best balance between preserving pretrained priors and adapting to the dynamic multi-view setting.

Method	Trainable params	PSNR ↑	SSIM ↑	LPIPS ↓
Unfreeze camera and depth heads	525 M	16.94	0.654	0.354
Full fine-tuning	621 M	25.09	0.829	0.162
Freeze backbone	266 M	25.68	0.804	0.151
Fine-tune last 4 layers	318 M	26.32	0.841	0.152
Fine-tune last 8 layers	369 M	27.28	0.853	0.140
Ours (fine-tune 16 alternating layers)	470 M	**29.15**	**0.886**	**0.125**

Table 6 reveals four findings. First, unfreezing the depth and camera heads (*Unfreeze camera and depth heads*) is catastrophic, dropping PSNR by 12.4 points compared to our configuration. The pretrained DA3 heads encode strong geometric priors that are easily destabilized by the noisy gradients flowing back from the dynamic Gaussian losses; once these priors are corrupted, the unprojected point map (and therefore Gaussian means) becomes unreliable. Second, fully fine-tuning the entire backbone (*Full fine-tuning*) also underperforms our configuration by 4.3 PSNR, indicating that the deeper backbone layers contain general-purpose representations that benefit from preservation. Third, completely freezing the backbone (*Freeze backbone*) under-fits, dropping 3.7 PSNR; the dynamic multi-view setting requires task-specific adaptation that the frozen backbone cannot provide. Fourth, progressive unfreezing from the back of the network (*Fine-tune last 4 / last 8 layers*) shows monotonic improvement as more layers are exposed, but our strategy of unfreezing all 16 alternating attention layers (while keeping the within-view layers and pretrained heads frozen) achieves the best result. This isolates the parameters that govern cross-view feature aggregation, where the dynamic multi-view setting most differs from DA3's pretraining objective.

B Pose Alignment for Target-View Rendering

Since NoPo4D is pose-free, rendering target views during training requires aligning predicted target poses with the reconstructed scene. We use a two-pass strategy: the first pass processes only context frames, producing Gaussians and predicted context poses that define the scene's coordinate frame. A second pass processes all frames under `no_grad`, yielding fresh context and target poses. Umeyama alignment between the two context-pose sets gives a closed-form Sim(3) mapping target poses into the scene's frame. Because the second pass is detached, the rendering loss supervises only the Gaussians from the first pass, not the encoder's pose predictions. The same two-pass strategy is used at inference, where no ground-truth poses are available.

C Robustness to Varying Input View Counts

NoPo4D is trained on multi-view sequences with 1–4 cameras but can be deployed at inference time with arbitrary view counts. We evaluate this generalization by varying the number of input cameras $C \in \{2, \ldots, 12\}$ on N3DV.

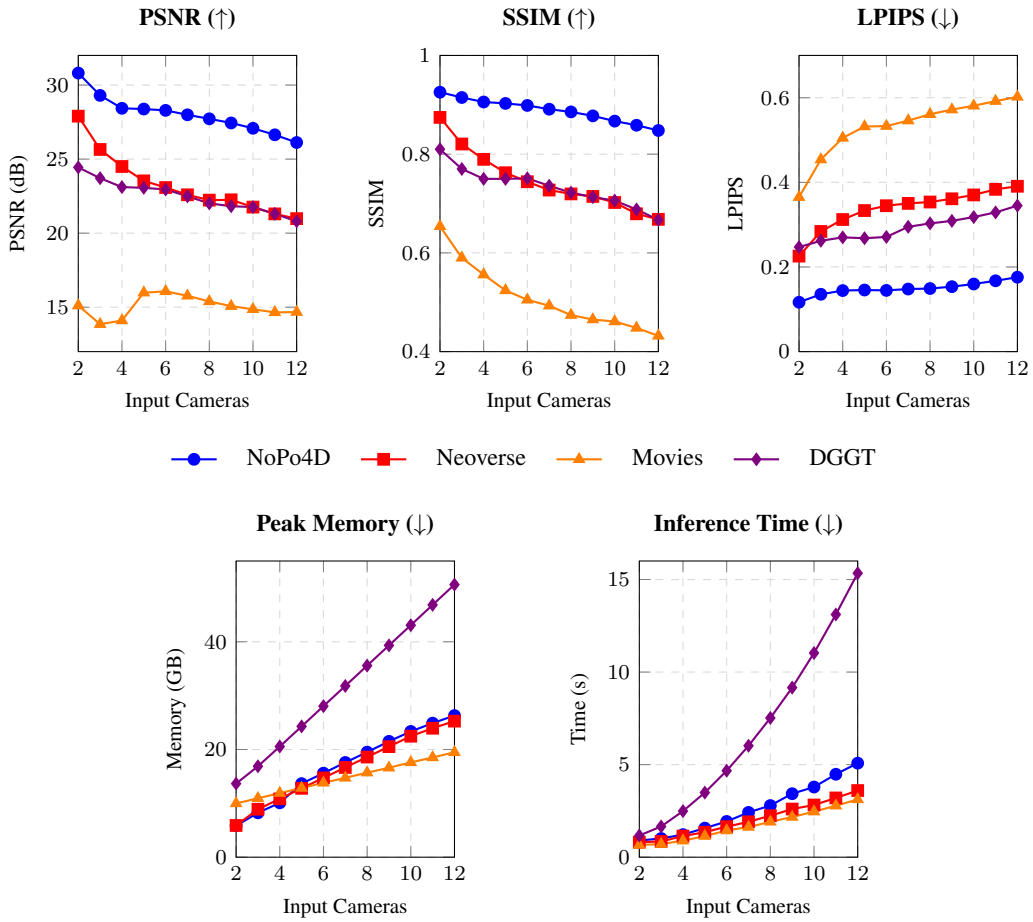


Figure 4: **Scalability Analysis.** Multi-view input density vs. rendering quality (top row) and computational cost (bottom row). DGGT inference time and scaling curves indicate distinct trade-offs against baseline models.

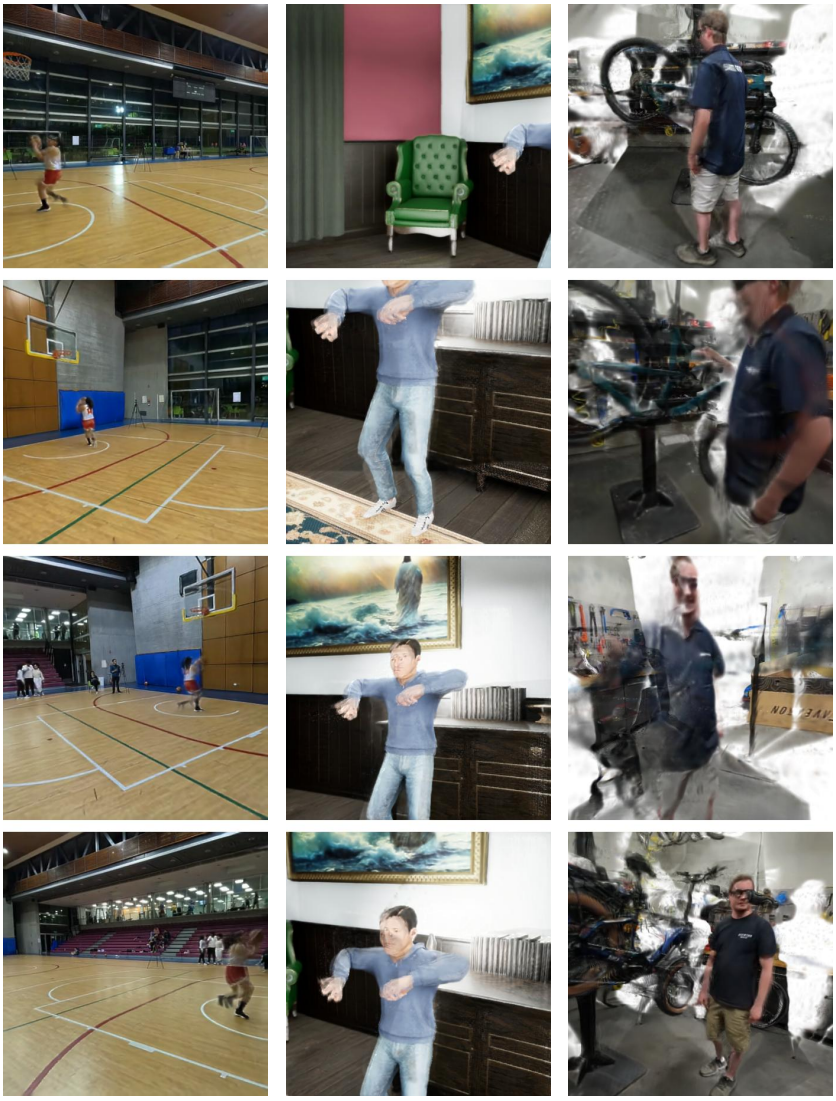
Figure 4 reveals several findings. First, NoPo4D maintains a substantial quality lead over all baselines across the full range of input camera counts, despite never having seen $C > 4$ during training. Second, all methods show a modest quality decrease as the camera count grows: with more cameras, the target views being evaluated are further from any input view, making the task harder. NoPo4D degrades most gracefully, losing only $\sim$4.7 PSNR from $C = 2$ to $C = 12$, compared to $\sim$6.9 for NeoVerse and $\sim$3.6 for DGGT. Third, computational cost scales near-linearly with input cameras for NoPo4D, NeoVerse, and MoVieS, while DGGT's memory and inference time grow super-linearly, reaching $\sim$50 GB and $\sim$15 seconds at 12 cameras. NoPo4D remains under 27 GB and 5 seconds at the same scale.

D Additional Qualitative Results

Static qualitative comparisons. We provide additional side-by-side comparisons against feed-forward baselines on each of the four benchmarks.

E Failure Cases

We document representative failure modes in Fig. 5.



(a) (b) (c)

Figure 5: **Failure cases.** (a) Cross-view misalignments and floating artifacts in fast-moving regions, where the motion encoder cannot fully resolve large inter-frame displacements. (b) Degradation under camera motion (RecamMaster synthetic dataset [3]), where the static-rig assumption causes the per-camera pose averaging to collapse distinct viewpoints into an incorrect single pose. (c) Floating Gaussian artifacts in scene regions far from input camera viewpoints, where the lack of cross-view geometric constraint leaves Gaussians at incorrect depths.

F Broader Impacts

NoPo4D advances feed-forward 4D scene reconstruction from unposed multi-view video, with potential applications [24–26, 12–14, 49, 28, 1, 34, 73, 19, 36, 35, 74, 22, 20, 31] in accessible content creation, point tracking, scene understanding, correspondence estimation, robotics, sports and performance capture, egocentric scene understanding, and immersive media. We note two responsible-use considerations specific to this capability. First, reconstructing dynamic scenes from arbitrary multi-camera arrays without requiring calibration lowers the barrier to surveillance applications; downstream deployments should consider whether captured subjects have consented to 3D reconstruction, not only 2D recording. Second, the capacity to render scenes from novel viewpoints raises concerns about synthetic media when combined with generative tools. NoPo4D itself produces only reconstructions of observed content (not synthesized scenes), but practitioners building on this work should be mindful of these dual-use possibilities. We follow standard practices in our own work: training data (Ego-Exo4D) was released under its original consent and license terms, and we do not collect new human-subjects data.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Table 1,2,3 and 4 support the claims in the introduction and abstract.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have dedicated section for limitation and failure cases.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: [N/A]

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The Method and Experiments section explain the details of the architecture and the experimental setup.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The Method and Experiments section explain the details of the architecture and the experimental setup

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The Method and Experiments section explain the details of the architecture and the experimental setup

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The Experiments section reports the performance under fair setup

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The Experiments section explains the details of the experimental setup

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: code of ethics is strictly conformed

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Supplementary material.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer:[N/A]

    Justification: [N/A]

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All cited

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.