POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in High Performance Computing Engineering

Author: Matteo Balice

Advisor: Prof. Matteo Matteucci

Co-advisors: Dr. Sunghwan Hong, Prof. Marc Pollefeys

Academic year: 2025–2026

---

1. Introduction

The ability to reconstruct dynamic three-dimensional scenes from ordinary video and navigate them freely from any viewpoint stands among the most compelling goals of computer vision. Today, compelling free-viewpoint video requires a controlled studio environment, expensive calibrated camera arrays, and hours of per-scene processing. Recent progress in 3D Gaussian Splatting [1] and feed-forward geometry models has removed many of these barriers for *static* scenes, but extending these advances to dynamic phenomena introduces three compounding challenges.

1. **Modeling dynamic content.** Dynamic scenes require a 4D representation that captures how the scene evolves over time, together with mechanisms for tracking motion and supervising predicted dynamics with appropriate training signals.
2. **Fusing multi-view information.** A single camera leaves depth and occlusion ambiguities that only additional views can resolve, requiring attention mechanisms that aggregate information across cameras while preserving per-view spatial structure.
3. **Estimating geometry without poses.** In casual multi-camera setups (smartphones around a table, action cameras worn by athletes), calibration is unavailable, so poses must be inferred jointly with the scene geometry.

Each prior approach solves at most two of these three problems simultaneously (Table 1). Static pose-free methods achieve instant novel-view synthesis from unposed images but cannot model motion. Feed-forward dynamic methods such as 4DGT [6] model temporally coherent Gaussians with velocity attributes but require known poses and monocular input. NeoVerse [7] removes the pose requirement but relies on 3D scene-flow ground truth unavailable in real captures and remains restricted to a single camera. DGGT operates in the unposed multi-view dynamic regime but is restricted to driving scenarios.

This thesis introduces **NoPo4D**, the first feed-forward system that jointly addresses all four requirements: feed-forward inference, no camera pose requirement, multi-view input, and dynamic scene modeling in general environments. The key contributions are:

1. Identification of the previously unaddressed

Table 1: **Capability comparison of feed-forward and optimization methods.** NoPo4D is the first feed-forward system that jointly handles dynamic scenes from unposed multi-view video in general environments. **FF**: feed-forward; **Unposed**: no GT poses; **MV**: multi-view input; **Dyn.**: models dynamics; **General**: not domain-restricted.

Method	Type	FF	Unposed	MV	Dyn.	General
NoPoSplat	Static	✓	✓	✓	✗	✓
AnySplat	Static	✓	✓	✓	✗	✓
4DGT [6]	Dyn.	✓	✗	✗	✓	✓
NeoVerse [7]	Dyn.	✓	✓	✗	✓	✓
DGGT	Dyn.	✓	✓	✓	✓	✗
Shape of Motion	Opt.	✗	✗	✗	✓	✓
MonoFusion [5]	Opt.	✗	✗	✓	✓	✓
NoPo4D (Ours)	**Dyn.**	✓	✓	✓	✓	✓

problem setting and a systematic comparison demonstrating that no prior method fills it.

2. **NoPo4D**: a velocity decomposition that splits Gaussian motion into per-pixel image-plane shifts and depth changes, enabling direct optical-flow supervision without rendering or 3D annotations; complemented by a bidirectional motion encoder providing globally consistent velocities and view-dependent opacity handling multi-view geometric misalignments.
3. Comprehensive evaluation on four benchmarks demonstrating consistent superiority over feed-forward baselines and, with optional post-optimization, superiority over calibration-requiring per-scene optimization methods.

2. Method

2.1. System Overview

NoPo4D takes as input C uncalibrated, time-synchronized video streams $\mathcal{I} = \{(\mathbf{I}_t^c)_{t=1}^T\}_{c=1}^C$ and produces a collection of $G = C \cdot T \cdot H \cdot W$ 4D Gaussians in a single forward pass, with one Gaussian per pixel per frame per camera. Each Gaussian carries static attributes (mean $\boldsymbol{\mu}_g$, covariance $\boldsymbol{\Sigma}_g$, view-dependent opacity α_g, color $\mathbf{c}_g$) and dynamic attributes (temporal center τ_g, lifespan l_g, and forward/backward linear and angular velocities $\mathbf{v}_g^{\pm}, \boldsymbol{\omega}_g^{\pm}$ [6]). The position of Gaussian g at time t is:

$$\boldsymbol{\mu}_g(t) = \boldsymbol{\mu}_g + \begin{cases} \mathbf{v}_g^+ \cdot (t - \tau_g) & t > \tau_g, \\ \mathbf{v}_g^- \cdot (\tau_g - t) & t \leq \tau_g. \end{cases} \quad (1)$$

The asymmetric forward/backward parameterization allows independent modeling of approach and departure, enabling abrupt direction changes and transient objects.

The system has four components, illustrated in Figure 1: (1) a pretrained DA3 [3] transformer backbone with frozen depth and camera heads that extract multi-view features and predict per-frame depth maps and camera parameters; before the alternating cross-view attention layers, sinusoidal temporal tokens encoding each frame's normalized timestamp are injected, making the backbone aware of temporal ordering; (2) back-projection of each depth pixel into a world-space Gaussian mean; (3) a trainable DPT Gaussian head decoding static attributes from backbone features; (4) a bidirectional motion encoder predicting velocity and temporal parameters.

2.2. Velocity Decomposition

The central technical contribution is a velocity decomposition that avoids both the rendering dependence of posed methods and the 3D annotation requirement of unposed methods.

Prior feed-forward dynamic methods predict 3D velocities directly and supervise them either by rendering 2D optical flow (which ties supervision quality to pose accuracy) or by comparison to 3D scene flow (unavailable at scale in real data). The proposed decomposition avoids both paths. For each pixel (x, y) in frame t, a DPT head predicts a 2D image-plane shift $(\Delta x, \Delta y)$ and a

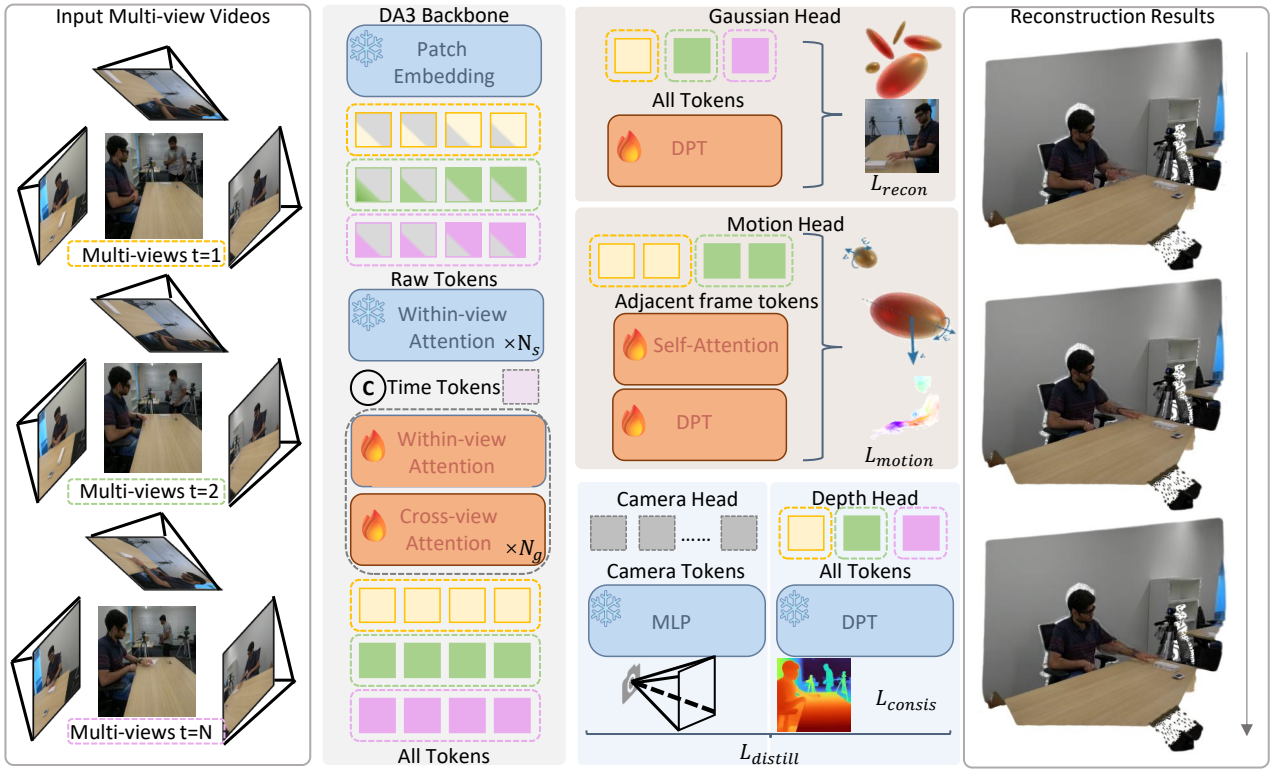

Figure 1: **NoPo4D architecture.** Given C synchronized video streams, DA3 [3] extracts multi-view features through alternating within-view and cross-view attention, injected with sinusoidal temporal tokens. Frozen depth and camera heads recover per-frame geometry, which is unprojected into Gaussian means. A trainable Gaussian head decodes static parameters $(\mathbf{R}, \mathbf{s}, \alpha, \mathbf{c})$, while a bidirectional motion branch processes consecutive-frame features to produce velocities $(\mathbf{v}^{\pm}, \boldsymbol{\omega}^{\pm})$, motion gate ρ, and temporal variance σ_g.

scalar depth displacement Δd:

$$x' = x + \Delta x, \quad y' = y + \Delta y,$$
$$\hat{D}^c_{t+1}(x, y) = \hat{D}^c_t(x, y) + \Delta d. \tag{2}$$

Using predicted intrinsics $\hat{\mathbf{K}}^c$, source and displaced pixels are unprojected into camera-space coordinates and differenced to obtain the displacement $\Delta\mathbf{X}_{\text{cam}} \in \mathbb{R}^3$. The forward linear velocity is then obtained by rotating this displacement into world coordinates:

$$\mathbf{v}_g^+ = \frac{\hat{\mathbf{R}}^c \, \Delta\mathbf{X}_{\text{cam}}}{\Delta t}. \tag{3}$$

The key property is that $(\Delta x, \Delta y)$ encodes the Gaussian's projected displacement between consecutive frames *without any rendering step*, making it directly comparable to pseudo-ground-truth optical flow from SEA-RAFT [4]. Motion supervision is therefore completely decoupled from pose estimation accuracy, sidestepping the circular dependency that plagues posed dynamic methods.

2.3. Bidirectional Motion Encoder

Image-plane shifts, depth changes, angular velocities, and temporal parameters are produced by a bidirectional motion encoder purpose-built to aggregate temporal context (what changed between frames) and cross-view context (how the same motion appears from different cameras) simultaneously.

For each consecutive frame pair $(t, t+1)$, the encoder concatenates feature tokens from all C cameras at both timesteps ($2CN$ tokens total), adds learnable temporal embeddings distinguishing source- from neighbor-frame tokens, and applies joint multi-head self-attention in a single pass. This allows every spatial location in every camera to attend to every other location across both frames and viewpoints, resolving motion ambiguities unresolvable from a sin-

gle stream. The encoder additionally outputs a per-pixel motion gate $\rho_g \in (0, 1)$ that suppresses velocity predictions in static regions, preventing flickering artifacts caused by spurious background velocities.

2.4. View-Dependent Opacity

Opacity is parameterized by spherical harmonic (SH) coefficients rather than a scalar. In the multi-view pose-free setting, Gaussians unprojected from different cameras are inevitably slightly misaligned in 3D space due to pose estimation errors. A scalar opacity forces a global keep-or-discard decision, whereas SH-parameterized opacity allows the model to learn a per-direction confidence: the Gaussian remains fully opaque from the camera that unprojected it correctly and fades for cameras where its reprojection is geometrically inconsistent. This soft learned selection mechanism handles the inherent imprecision of pose-free reconstruction without any explicit outlier detection or multi-view consistency loss.

2.5. Training

The total training loss combines four terms:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{motion}}\mathcal{L}_{\text{motion}} + \lambda_{\text{consis}}\mathcal{L}_{\text{consis}} + \mathcal{L}_{\text{distill}}. \quad (4)$$

The *reconstruction loss* $\mathcal{L}_{\text{recon}}$ combines MSE, SSIM, and LPIPS on rendered target frames. The *motion loss* directly compares predicted pixel shifts to SEA-RAFT flow, applied only to pixels with flow magnitude >2 pixels (Ω'):

$$\mathcal{L}_{\text{motion}} = \frac{\lambda_{\text{flow}}}{|\Omega'|} \sum_{p \in \Omega'} \left\| (\Delta x, \Delta y)(p) - \mathbf{f}^*(p) \right\|_1. \quad (5)$$

The *distillation loss* $\mathcal{L}_{\text{distill}}$ preserves the pretrained geometric priors of the DA3 backbone by distilling from a frozen DA3 Giant teacher (Huber loss on poses, ℓ_2 on depth and normal maps). The *depth consistency loss* $\mathcal{L}_{\text{consis}}$ prevents Gaussians from drifting away from their initialized positions.

Training proceeds in two stages of 20 000 steps each: first on single-camera viewpoints to initialize the static representation and motion heads, then on 1–4 cameras simultaneously with temporal stride warmed up from 4 to 8. Only the 16 alternating cross-view attention layers of the DA3 backbone are fine-tuned; the within-view layers and pretrained depth and camera heads remain frozen throughout, since unfreezing the latter causes a 12.2 dB quality drop.

An optional test-time post-optimization applies 100 gradient steps of photometric refinement to the feed-forward output, converging in under 5 seconds per scene.

3. Experimental Results

3.1. Setup

NoPo4D is trained on approximately 2 900 Ego-Exo4D scenes of diverse everyday human activities captured by synchronized multi-camera rigs in kitchens, sports facilities, and medical settings. Evaluation covers four multi-view dynamic benchmarks. *ExoRecon* (in-distribution) derives sparse multi-view captures from Ego-Exo4D. *Immersive Light Field (ILF)* contains real-world scenes from a 46-camera array. *Kubric* is a synthetic dataset with large-displacement multi-object motion. *N3DV* offers 18–21 cameras, enabling evaluation at extreme novel viewpoints (34.6°–71.9° from input cameras). All benchmarks share a chunk-based evaluation protocol: 4 context frames predict a held-out middle frame; metrics (PSNR, SSIM, LPIPS) are averaged across cameras and scenes. Feed-forward baselines include NeoVerse, MoVieS, and DGGT (retrained on Ego-Exo4D data). Per-scene optimization baselines include MonoFusion [5], MV-SOM, and Dyn3D-GS, all requiring known camera calibration.

3.2. Quantitative Results

Table 2 shows results on the in-distribution ExoRecon benchmark. NoPo4D substantially outperforms all feed-forward baselines, leading the nearest competitor (NeoVerse) by 9.1 dB PSNR. The gap over DGGT retrained on the same data (8.7 dB) isolates architectural contributions from training-distribution effects. With optional 100-step post-optimization, NoPo4D surpasses MonoFusion, the current per-scene optimization state of the art, by 1.5 dB, despite requiring no camera calibration.

Table 3 reports out-of-distribution performance on ILF and Kubric. NoPo4D consistently leads the strongest feed-forward baseline (DGGT),

Table 2: **Held-out view synthesis on ExoRecon [5]. FF**: feed-forward inference. The DGGT baseline in the feed-forward section is retrained on the same Ego-Exo4D data as NoPo4D to eliminate distribution effects.

Method	FF	PSNR↑	SSIM↑	LPIPS↓
Dyn3D-GS	✗	24.28	0.692	0.539
MV-SOM	✗	26.91	0.890	0.138
MonoFusion [5]	✗	30.43	**0.930**	**0.060**
NoPo4D + post-opt.	✗	**31.95**	0.928	0.075
MoVieS	✓	18.78	0.725	0.288
DGGT (Ego-Exo4D)	✓	20.44	0.704	0.418
NeoVerse	✓	20.03	0.714	0.354
NoPo4D (Ours)	✓	**29.15**	**0.886**	**0.125**

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

	Immersive Light Field			Kubric		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MoVieS	16.12	0.650	0.410	14.49	0.720	0.260
NeoVerse	19.83	0.680	0.360	16.86	0.500	0.520
DGGT (Ego-Exo4D)	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 + post-opt.	**24.04**	**0.817**	**0.245**	**28.17**	**0.838**	**0.159**

with a particularly pronounced advantage on Kubric (+3.5 dB feed-forward over DGGT), where large-displacement motion benefits most from the explicit velocity parameterization. Post-optimization delivers further gains of 2–5 dB across both datasets.

Table 4 shows results on N3DV, which provides enough cameras for evaluation at extreme novel viewpoints. NoPo4D leads by nearly 4 dB at input cameras and remains competitive at moderate novel views. At extreme viewpoints (34.6°–71.9°), NoPo4D produces visually cleaner reconstructions than all baselines, while DGGT exhibits semi-transparent ghosting and NeoVerse/MoVieS over-smooth.

3.3. Scalability and Ablations

NoPo4D generalizes to 12 cameras despite training with at most 4, losing only ∼4.7 dB PSNR (vs. ∼6.9 dB for NeoVerse), with memory and inference time scaling near-linearly (<27 GB, <5 s at 12 cameras), compared to DGGT's super-linear growth (∼50 GB, ∼15 s).

Architectural ablations. Removing the bidirectional motion encoder (−6.2 dB) and replacing SH opacity with scalar opacity (−7.5 dB) are the most damaging modifications: independent per-camera velocity estimation produces conflicting Gaussian velocities, while scalar opacity cannot suppress artifacts selectively per viewpoint. Replacing the velocity decomposition with direct 3D prediction costs −1.5 dB, and removing temporal tokens costs a further −1.5 dB.

Loss ablations. Distillation is the most critical auxiliary term (−3.2 dB), anchoring the attention layers to DA3's geometric priors during noisy dynamic training. Flow supervision prevents velocity collapse (−1.0 dB) and depth consistency reduces multi-view disagreements (−1.1 dB). Using consistency alone without distillation is catastrophic (3.51 dB), as the model collapses Gaussians onto a degenerate surface.

Backbone fine-tuning. Fine-tuning only the 16 alternating cross-view attention layers is optimal; unfreezing the depth/camera heads drops

Table 4: **View synthesis on N3DV [2],** evaluated at input cameras, moderate novel views (5.7°–27.7°), and extreme novel views (34.6°–71.9°).

	Input Cameras			Novel Cameras (Moderate)			Novel Cameras (Extreme)		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
MoVieS	14.10	0.556	0.505	14.54	0.435	0.582	9.63	0.162	0.716
NeoVerse	24.50	0.789	0.313	18.63	0.530	0.449	13.32	0.423	0.601
DGGT	23.08	0.749	0.268	17.93	0.532	0.435	15.87	0.449	0.557
DGGT (Ego-Exo4D)	22.49	0.713	0.356	6.84	0.223	0.626	10.43	0.419	0.631
NoPo4D (Ours)	28.43	0.905	0.144	**19.84**	**0.578**	**0.315**	15.69	**0.467**	**0.520**
NoPo4D + post-opt.	**31.79**	**0.956**	**0.089**	19.93	0.570	0.320	**15.91**	0.449	0.531

−12.2 dB, full fine-tuning −4.1 dB, and freezing the backbone entirely −3.5 dB.

4. Conclusions

NoPo4D is the first feed-forward system for 4D Gaussian reconstruction from unposed multi-view video in general environments. Its three core contributions (velocity decomposition for render-free optical-flow supervision, a bidirectional motion encoder for cross-view consistency, and SH-parameterized opacity for pose-free misalignment handling) jointly address a problem that no prior method solved simultaneously. Across four benchmarks, NoPo4D outperforms all feed-forward baselines and, with optional post-optimization, surpasses calibration-requiring per-scene methods while being significantly faster.

Open directions include support for moving camera rigs, reduced dependency on pretrained teachers (DA3, SEA-RAFT), higher-order motion modeling, and extension to monocular or asynchronous capture.

References

[1] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering, August 2023. arXiv:2308.04079.

[2] 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 (CVPR)*, pages 5521–5531, 2022.

[3] 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. arXiv:2511.10647.

[4] Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow, May 2024. arXiv:2405.14793.

[5] Zihan Wang, Jeff Tan, Tarasha Khurana, Neehar Peri, and Deva Ramanan. MonoFusion: Sparse-View 4D Reconstruction via Monocular Fusion, July 2025. arXiv:2507.23782.

[6] 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. arXiv:2506.08015.

[7] 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. arXiv:2601.00393.