CoSTA4D: Consistent Spatio-Temporal Aggregation for Streaming 4D Human Reconstruction from Uncalibrated Sparse-View Videos

Anonymous authors
Paper under double-blind review



Figure 1: Overview of CoSTA4D for streaming 4D human reconstruction. Given sequential multi-view observations from fixed, uncalibrated sparse cameras, CoSTA4D uses only current and past information to progressively refine the shared camera state and recursively maintain the dynamic Gaussian representation, enabling stable and temporally consistent novel-view reconstruction.

Abstract

Streaming human reconstruction from sparse, uncalibrated multi-view videos requires both geometric accuracy and temporal consistency. While human motion and visibility vary over time, the capture cameras remain fixed; however, changing visual evidence can still cause unstable camera and depth predictions, leading to geometric jitter and rendering flicker. This suggests that historical observations should not only provide temporal context for current reconstruction, but also accumulate evidence about the invariant capture geometry. We present **CoSTA4D**, a streaming feed-forward framework built on this principle. CoSTA4D combines anchor-context temporal aggregation with reliability-aware camera refinement, enabling historical context to stabilize geometry inference while repeated camera observations progressively improve the shared camera estimates and back-projected geometry. Correspondence-guided Gaussian propagation further encourages continuity of the dynamic representation across frames. The framework uses only current and past observations without per-sequence optimization. Experiments on DNA-Rendering and self-captured videos demonstrate state-of-the-art reconstruction quality and improved temporal consistency.

1 Introduction

Reconstructing dynamic humans from sparse multi-view videos supports free-viewpoint video and immersive communication (Lin et al., 2022; Zheng et al., 2024). Recent advances in visual geometry enable dense reconstruction from sparse and uncalibrated observations (Wang et al., 2024b; 2025a; Zhang et al., 2025; Xu et al., 2025b; Huang et al., 2026; An et al., 2026), while feed-forward human rendering enables high-quality novel-view synthesis without per-subject optimization (Zheng et al., 2024). However, streaming reconstruction requires not only accurate results at individual timestamps, but also stable geometry over time. Frame-to-frame fluctuations in camera and depth

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

predictions can propagate through back-projection into rendering flicker, making stable geometric inference essential for temporal consistency.

Recent streaming reconstruction methods exploit historical features, spatiotemporal modeling, or cross-frame representation propagation to improve temporal consistency (Jin et al., 2025; Zhi et al., 2025; Lyu et al., 2026; Chen et al., 2025b;a; Shao et al., 2025). For example, Forge4D compresses historical information into recurrent states and combines temporal scale constraints with motion-guided Gaussian fusion for streaming reconstruction (Hu et al., 2026). However, such generic historical states do not explicitly distinguish information with different temporal validity: past human poses, occlusions, and local dynamics can quickly become stale, whereas the fixed capture geometry remains valid throughout the sequence. Continuously mixing these heterogeneous histories may entangle outdated dynamic cues with current observations rather than consistently benefiting reconstruction. In contrast, optimization-based multi-frame geometry can explicitly integrate evidence across observations, but its iterative test-time processing is less compatible with causal, feed-forward streaming. This raises a fundamental question: *what information should persist in streaming reconstruction, and at what temporal scale?*

Our key observation is that different states in streaming human reconstruction have fundamentally different temporal persistence: human geometry, pose, and visibility continuously change, whereas the capture cameras remain physically invariant throughout the sequence. Consequently, camera predictions at different timestamps should not be treated as independent final estimates, but rather as repeated observations of the same latent camera state with unequal reliability. This reframes camera consistency as a *learned online state-estimation problem*: reliable evidence should accumulate over time, while transient unreliable estimates should not readily overwrite the accumulated capture geometry. More broadly, historical information in streaming reconstruction should be organized according to its physical persistence—keeping dynamic context fresh while preserving long-lived geometric state—rather than uniformly compressing all past observations.

We propose **CoSTA4D**, a streaming feed-forward framework that progressively exploits historical observations for 4D human reconstruction with a dynamic Gaussian representation (Kerbl et al., 2023; Wu et al., 2024). We introduce anchor-context temporal aggregation to leverage temporal context for more stable camera and depth inference. Exploiting the multi-view capture configuration, reliability-aware camera refinement recursively fuses camera predictions over time, with the refined parameters used for depth back-projection during both training and inference. Based on this geometry, correspondence-guided gaussian propagation selectively propagates reliable historical Gaussians to the current frame while retaining current predictions where correspondences are unreliable, establishing explicit continuity across timestamps. Using only current and past observations without per-sequence optimization, CoSTA4D achieves state-of-the-art reconstruction quality and improved temporal consistency on DNA-Rendering (Cheng et al., 2023) and self-captured videos.

Our main contributions are as follows:

- We introduce CoSTA4D, a streaming feed-forward framework for human reconstruction from sparse, uncalibrated multi-view videos that uses causal temporal context to infer current geometry while maintaining a shared estimate of the multi-view capture cameras.
- We propose reliability-aware sequence-level camera refinement that treats temporal predictions as repeated observations of each fixed camera and recursively fuses them to obtain stable camera estimates for 3D geometry construction.
- We design correspondence-guided gaussian propagation that combines motion-compensated historical Gaussians with current predictions according to correspondence reliability, improving representation continuity across consecutive timestamps.

2 Related Work

2.1 Feed-forward Sparse-view Human Reconstruction

Canonical formulations of NeRF (Mildenhall et al., 2022) and 3D Gaussian Splatting (Kerbl et al., 2023) rely on per-scene optimization, while recent feed-forward human reconstruction methods directly predict 3D representations from input images. GPS-Gaussian (Zheng et al., 2024), GPS-Gaussian+ (Zhou et al., 2025), RoGSplat (Xiao et al., 2025), and GBC-Splat (Tu et al., 2025) im-

prove sparse-view human reconstruction through pixel-aligned Gaussians (Charatan et al., 2024; Xu et al., 2025a), cross-view modeling (Chen et al., 2024a;b; Xie et al., 2026; Lou et al., 2025), or human priors. Recently, visual geometry models such as DUSt3R (Wang et al., 2024b), MASt3R (Leroy et al., 2024), VGGT (Wang et al., 2025a) and VGGT-Ω (Wang et al., 2026) further enable reconstruction from uncalibrated images, motivating pose-free methods including NoPoSplat (Ye et al., 2025) and AnySplat (Jiang et al., 2025). HiReFF (Jiang et al., 2026) extends this direction to high-resolution uncalibrated human reconstruction with scale-synchronized camera supervision and high-resolution side-tuning. In contrast, CoSTA4D exploits sequential observations to stabilize camera and depth inference and recursively refine fixed camera parameters during streaming reconstruction.

2.2 Dynamic Gaussian Representations

Dynamic Gaussian (Luiten et al., 2024; Yang et al., 2024a; Wu et al., 2024; Yang et al., 2024b; Huang et al., 2024; Mihajlovic et al., 2024; Wang et al., 2025c; Zhou et al., 2026; Lu et al., 2025b) methods represent time-varying scenes through persistent primitives, motion trajectories, or deformation models. Recent feed-forward approaches instead predict dynamic representations directly from videos. 4DGT (Xu et al., 2025c) predicts 4D Gaussians using rolling windows, while DGS-LRM (Lin et al., 2025) predicts per-pixel deformable Gaussians. Both take posed monocular videos as input and therefore do not directly address camera estimation from uncalibrated captures. Recent human synthesis methods (Jin et al., 2025; Zhi et al., 2025) further model cross-view and temporal coherence with spatiotemporal generative priors, but focus on view synthesis rather than causal geometric reconstruction from uncalibrated streams. For uncalibrated sparse-view human videos, Forge4D (Hu et al., 2026) combines recurrent states and temporal scale constraints for aligned streaming reconstruction, with motion prediction and occlusion-aware Gaussian fusion supporting novel-time synthesis. Our propagation mechanism targets the current observed timestamp: backward motion correspondences selectively carry historical Gaussians into the current representation, while current predictions are retained where correspondence is unreliable. This explicit representation continuity complements temporal aggregation and camera refinement.

2.3 Temporal Modeling and Camera Consistency

Streaming geometry models (Lu et al., 2025a; Hu et al., 2025; Ke et al., 2025; Ha et al., 2025) incorporate historical observations through recurrent states or causal attention. CUT3R (Wang et al., 2025b) updates a persistent latent state to predict geometry in a shared coordinate system, while StreamVGGT (Zhuo et al., 2026) and STream3R (Lan et al., 2026) access cached historical features through causal attention. TTT3R (Chen et al., 2026) regulates latent-state updates using memory–observation alignment confidence to improve length generalization.

Alternatively, systems (He et al., 2024; Wang et al., 2024a; Elflein et al., 2025; Rockwell et al., 2024; Murai et al., 2025) combine learned geometric priors with camera tracking, loop closure, and global optimization for consistent camera trajectories and scene geometry. CoSTA4D instead exploits the temporal invariance of multi-view cameras, treating predictions across frames as repeated observations of the same camera state. Reliability-Aware Camera Refinement recursively updates shared camera parameters for depth back-projection and geometry construction.

3 Method

3.1 Overview

Given synchronized images $\mathcal{I}_t = \{I_t^n\}_{n=1}^N$ from $N = 4$ fixed, uncalibrated cameras, CoSTA4D reconstructs a dynamic human causally using current and past observations. Camera intrinsics and extrinsics remain fixed within each sequence, while their estimates are progressively refined. We maintain a historical state $\mathcal{H}_{t-1} = (\mathcal{M}_{t-1}, \mathcal{G}_{t-1}, \mathcal{S}_{t-1})$, containing feature memory, the previous Gaussian representation, and the shared camera state.

At timestamp t, Anchor-Context Temporal Aggregation predicts depth $\mathcal{D}_t$ and current camera parameters $\widehat{\mathcal{C}}_t$ from current and historical observations. Reliability-Aware Camera Refinement updates the shared estimate $\overline{\mathcal{C}}_t$, which is used to construct the current geometry:

$$\mathcal{X}_t = \mathrm{BackProj}(\mathcal{D}_t, \overline{\mathcal{C}}_t). \tag{1}$$

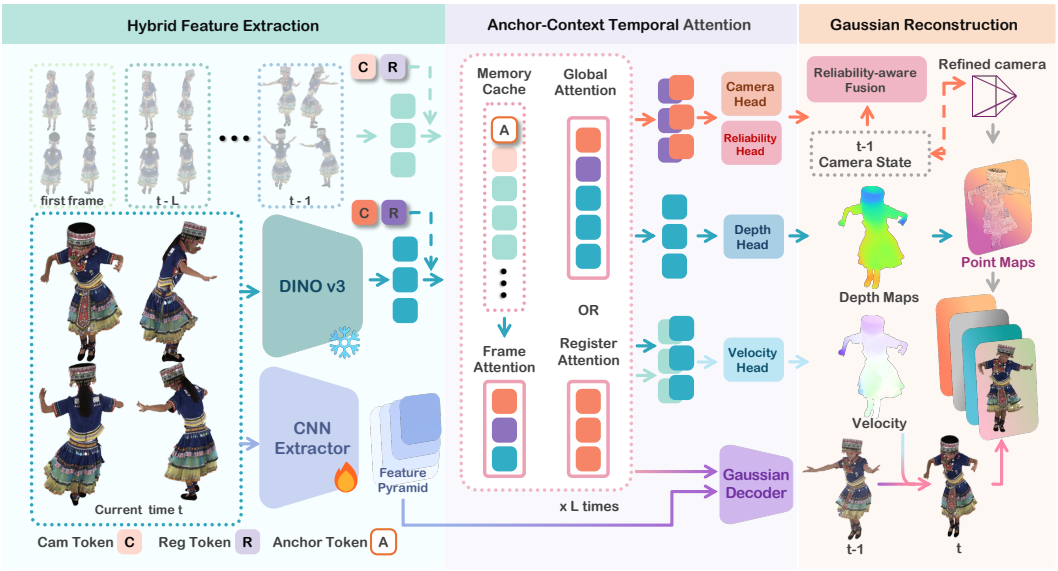

Figure 2: Overview of CoSTA4D. Current and historical sparse-view observations are first encoded by the hybrid feature extractor and integrated through Anchor-Context Temporal Aggregation. Camera, depth, and motion branches then predict the current geometry, while Reliability-Aware Camera Refinement recursively updates the shared camera state from repeated observations. Finally, the Gaussian decoder combines the current geometry and cross-frame motion to produce a temporally consistent dynamic Gaussian representation.

The Gaussian head then predicts a complete current representation $\widehat{\mathcal{G}}_t$, and Correspondence-Guided Gaussian Propagation combines current and historical Gaussians to obtain $\mathcal{G}_t$. The first timestamp uses current observations only and initializes all historical states. Camera translations, reconstructed geometry, and 3D motion share a fixed sequence-level coordinate and scale convention throughout the sequence. The overall pipeline is illustrated in Fig. 2.

3.2 Anchor-Context Temporal Aggregation

We build on VGGT-Ω (Wang et al., 2026), retaining its within-timestamp multi-view interaction while introducing causal temporal aggregation independently for each physical view. For each view, we maintain two complementary memories: the first-frame register tokens are retained as a persistent *anchor*, while camera and patch tokens from the most recent L timestamps form the *recent context*. Temporal interaction is restricted to the history of the same view; cross-view information exchange remains handled by the original multi-view processing of the backbone. The first frame initializes both memories, after which the recent context is updated with a sliding window. The persistent anchor provides a stable long-term reference, whereas the recent context captures local changes in geometry and visibility. Together, they provide causal temporal evidence for current camera and depth inference. At each timestamp, the current tokens attend to the persistent anchor and recent context of the same view.

3.3 Reliability-Aware Camera Refinement

Anchor-Context Temporal Aggregation conditions each camera prediction on historical context, while the multi-view capture setup provides a stronger sequence-level constraint: the parameters of each physical camera remain constant throughout the sequence. We therefore treat predictions across timestamps as repeated observations of the same camera state and recursively refine a shared estimate according to their reliability.

The following updates are applied independently to each camera. We use camera-to-world poses and denote the current prediction by $(\widehat{T}_t, \widehat{\mathbf{f}}_t)$, where $\widehat{T}_t \in SE(3)$ and $\widehat{\mathbf{f}}_t = [f_t^x, f_t^y]^\top$ contains the horizontal and vertical field-of-view angles in radians. A lightweight head $\mathcal{U}$ takes the current

camera token z_t as input and predicts separate reliability scores for rotation, translation, and FoV:

$$\boldsymbol{\ell}_t = \mathcal{U}(\mathbf{z}_t), \qquad \rho_t^\gamma = \exp(-\ell_t^\gamma), \qquad \gamma \in \{R, T, F\}. \tag{2}$$

A larger ρ_t^γ indicates a more reliable observation for the corresponding parameter group. For each group, we maintain a scalar fusion state A_t^γ and compute the update weight as

$$\beta_t^\gamma = \frac{(A_{t-1}^\gamma + \delta)\rho_t^\gamma}{1 + (A_{t-1}^\gamma + \delta)\rho_t^\gamma + \epsilon}, \qquad A_t^\gamma = (1 - \beta_t^\gamma)(A_{t-1}^\gamma + \delta). \tag{3}$$

Thus, more reliable observations contribute more strongly to the shared camera estimate. We update the pose in the tangent space of $SE(3)$ and the FoV by weighted averaging:

$$\overline{T}_t = \mathrm{Exp}\Big(B_t \,\mathrm{Log}(\widehat{T}_t \overline{T}_{t-1}^{-1})\Big)\,\overline{T}_{t-1}, \qquad \overline{\mathbf{f}}_t = (1 - \beta_t^F)\overline{\mathbf{f}}_{t-1} + \beta_t^F \widehat{\mathbf{f}}_t, \tag{4}$$

where $B_t = \mathrm{diag}(\beta_t^R I_3, \beta_t^T I_3)$. The refined cameras are used for depth back-projection and geometric supervision throughout training and inference, directly affecting geometry construction rather than merely smoothing frame-wise camera predictions.

3.4 Correspondence-Guided Gaussian Propagation

To establish explicit representation continuity across timestamps, we introduce lightweight historical propagation on top of per-frame Gaussian prediction. At each timestamp, the Gaussian head first predicts a complete current representation $\widehat{\mathcal{G}}_t$, such that the current reconstruction does not rely on successful historical propagation. For adjacent timestamps, a motion head predicts a backward 3D displacement $\mathbf{v}_t$ and correspondence confidence r_t from temporal features of the same physical view. These predictions establish correspondences with the final previous Gaussian state $\mathcal{G}_{t-1}$. For valid and confident correspondences, the motion-compensated historical Gaussian is propagated to the current representation; otherwise, the current prediction $\widehat{\mathcal{G}}_t$ is retained. The resulting $\mathcal{G}_t$ is stored as the state for the next timestamp, providing explicit temporal continuity without increasing the number of Gaussian primitives. Historical sampling, propagation gating, and confidence learning are detailed in Appendix A.3.

3.5 Training Objectives and Two-Stage Optimization

We adopt a two-stage training strategy. In **Stage I**, we train the backbone, camera, depth, and camera-reliability heads for stable geometric inference. The geometry objective is:

$$\mathcal{L}_{\mathrm{geo}} = \lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cam}} + \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}} + \lambda_{\mathrm{point}}\mathcal{L}_{\mathrm{point}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}, \tag{5}$$

where the four terms supervise camera parameters, depth, back-projected 3D geometry, and camera reliability, respectively. Since camera refinement relies on observation reliability, we learn it directly from camera prediction errors. For camera n, timestamp t, and parameter group $\gamma \in \{R, T, F\}$, let $e_{n,t}^\gamma$ denote the corresponding rotation, translation, or FoV error. We use:

$$\mathcal{L}_{\mathrm{rel}} = \sum_{n,t,\gamma} \left[\exp(-\ell_{n,t}^\gamma)\,\mathrm{sg}(e_{n,t}^\gamma) + \ell_{n,t}^\gamma\right], \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Since $\rho_{n,t}^\gamma = \exp(-\ell_{n,t}^\gamma)$, more accurate camera observations are encouraged to receive higher reliability and therefore contribute more during recursive refinement.

In **Stage II**, we introduce Gaussian and motion heads while further optimizing the geometry branch:

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{render}}\mathcal{L}_{\mathrm{render}} + \lambda_{\mathrm{motion}}\mathcal{L}_{\mathrm{motion}}. \tag{7}$$

The rendering loss combines photometric and perceptual supervision:

$$\mathcal{L}_{\mathrm{render}} = \lambda_{\mathrm{rgb}}\|\widehat{I} - I^{\mathrm{gt}}\|_1 + \lambda_{\mathrm{perc}}\,\mathrm{LPIPS}(\widehat{I}, I^{\mathrm{gt}}). \tag{8}$$

For cross-frame motion, a frozen SEA-RAFT (Wang et al., 2024c) provides backward correspondences during training. The predicted 3D displacement is jointly supervised by image-space flow consistency, pseudo-3D motion, and correspondence confidence:

$$\mathcal{L}_{\mathrm{motion}} = \lambda_{\mathrm{2D}}\|\widehat{\mathbf{b}}_t - \mathbf{b}_t\|_{1,\Omega_t} + \lambda_{\mathrm{3D}}\|\mathbf{v}_t - \mathbf{v}_t^*\|_{1,\Omega_t} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}}, \tag{9}$$

where $\widehat{\mathbf{b}}_t$ is the image-space motion induced by the predicted 3D displacement, $\mathbf{v}_t^*$ is a pseudo-3D target constructed from teacher correspondences and reconstructed geometry, and Ω_t denotes reliable supervision locations. Their construction, confidence supervision, and uncertainty filtering are detailed in Appendix A.4. SEA-RAFT is used only for training and removed at inference.



Figure 3: Qualitative novel-view comparisons on DNA-Rendering. Rows (a)–(e) show five examples; enlarged regions highlight faces, garment patterns, and local object details. Columns follow the method labels in the figure, with GT denoting the reference image. Our method preserves clearer local structures in the displayed examples.

4 Experiments

We evaluate CoSTA4D on novel-view reconstruction, progressive camera estimation, and temporal consistency. Quantitative comparisons and visual results assess reconstruction quality, while camera-error trajectories examine the use of accumulated observations for refining camera estimates.

4.1 Experimental Setup

Datasets. We train CoSTA4D on DNA-Rendering (Cheng et al., 2023), using 153 identities with 439 motion sequences for training and 30 unseen identities with 30 sequences for testing. Each training clip contains five synchronized timestamps from four fixed cameras sampled around an approximately 90° surrounding layout with random angular perturbations. The same four cameras are retained within each clip. Human masks are used to extract the foreground and composite the background to white. We further collect two self-captured multi-view sequences of approximately 300 and 1200 frames for generalization evaluation, without fine-tuning or test-time optimization.

Implementation Details. We train on eight NVIDIA A100 GPUs using Adam. The learning rate is cosine-decayed from 1×10^{-5} to 1×10^{-7}. Stage I is trained for 50K iterations with a per-GPU batch size of 3, and Stage II for 250K iterations with a per-GPU batch size of 1. The temporal context length is $L = 3$, and both input images and rendering supervision use 2048×2048 resolution. Automatic mixed precision is used throughout training.

Evaluation Protocol. Four fixed input views are used throughout each test sequence, with the remaining synchronized views serving as novel-view references. We run the released implementations of all baselines and re-evaluate their outputs under a unified evaluation protocol, while preserving each method's native camera-input requirements and output resolution. To avoid introducing interpolation artifacts into model predictions, the corresponding ground-truth image and human mask

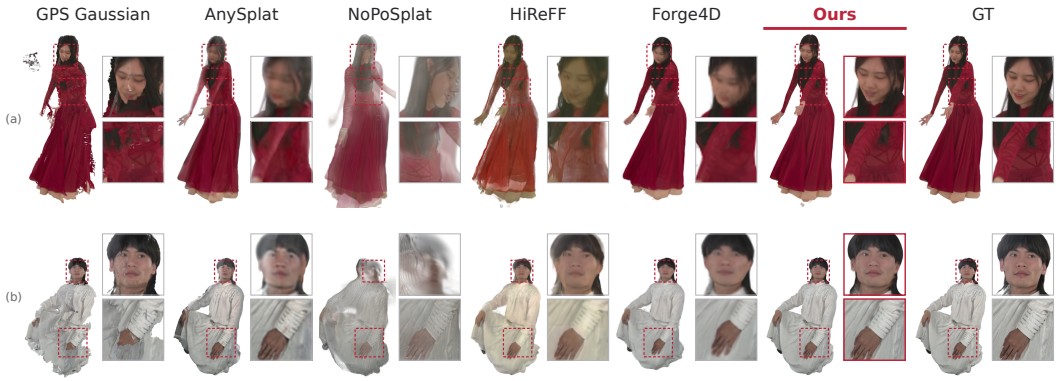

Figure 4: Qualitative comparisons on self-captured videos without adaptation. Enlarged regions highlight facial detail, garment appearance, and the hand–sleeve region. CoSTA4D more closely reproduces these local structures and colors in the displayed examples.

Table 1: Novel-view evaluation on DNA-Rendering. **K** and **T** denote required input intrinsics and extrinsics; ∅ denotes no supplied camera parameters. † marks the posed-video reference with a different input protocol. **Best** and Second-best results are bold and underlined, respectively.

Method	Cam. input	PSNR↑	SSIM↑	LPIPS↓	tOF↓	tLP(×100) ↓
GPS-Gaussian (Zheng et al., 2024)	**K**, **T**	29.7887	0.9729	0.0376	3.2130	0.7423
4DGT† (Xu et al., 2025c)	**K**, **T**	18.9478	0.9333	0.0542	4.7361	0.8567
NoPoSplat (Ye et al., 2025)	**K**	26.9590	0.9690	0.0353	3.6735	0.2911
AnySplat (Jiang et al., 2025)	∅	24.8063	0.9654	0.0505	2.9902	0.3489
DepthAnything 3 (Lin et al., 2026)	∅	22.8522	0.9590	0.0506	4.8008	0.8345
HiReFF (Jiang et al., 2026)	∅	26.6225	0.9688	0.0422	3.1570	0.2278
Forge4D (Hu et al., 2026)	∅	28.8507	0.9721	0.0392	3.2886	0.2192
Ours	∅	**31.3429**	**0.9781**	**0.0289**	**1.7954**	**0.1405**

are resized to match each method's rendered output resolution. The same resized mask is applied to both prediction and ground truth, with backgrounds composited to white before computing PSNR, SSIM, and LPIPS. Temporal consistency is evaluated using tOF and tLP (Chu et al., 2020) over the same evaluation region. Metrics are averaged per sequence and then across the test set. CoSTA4D receives no camera calibration at inference, and historical states are maintained continuously over each sequence. When camera registration is needed for evaluation, a single sequence-level alignment is estimated and fixed for the entire sequence.

4.2 Novel-view Reconstruction and Generalization

We compare CoSTA4D with GPS-Gaussian (Zheng et al., 2024), 4DGT (Xu et al., 2025c), NoPoSplat (Ye et al., 2025), AnySplat (Jiang et al., 2025), Depth Anything 3 (Lin et al., 2026), HiReFF (Jiang et al., 2026) and Forge4D (Hu et al., 2026). Their different camera-input requirements and protocols are summarized in Tables 1 and 2.

DNA-Rendering. As shown in Table 1, CoSTA4D achieves the best performance across all reported metrics, demonstrating strong novel-view rendering quality and temporal consistency without camera calibration at inference. Figure 3 further shows that CoSTA4D better preserves fine-scale facial, garment, and local structural details, with fewer blur and surface artifacts than competing methods.

Generalization to Unseen Data. We directly apply the DNA-Rendering-trained model to self-captured sequences without adaptation. Table 2 shows that CoSTA4D maintains its advantage under unseen subjects and capture conditions, indicating strong generalization. The qualitative comparisons in Fig. 4 further show improved preservation of facial details, garment appearance, and local structures in real-world captures.

Table 2: Evaluation on self-captured videos with independent synchronized reference views. Camera-input notation and highlighting follow Table 1.

Method	Cam. input	PSNR↑	SSIM↑	LPIPS↓	tOF↓	tLP($\times$100) ↓
GPS-Gaussian (Zheng et al., 2024)	**K**, **T**	26.3889	0.9671	0.0521	2.9345	2.1834
4DGT† (Xu et al., 2025c)	**K**, **T**	19.0836	0.9317	0.0538	4.6814	0.8756
NoPoSplat (Ye et al., 2025)	**K**	23.4625	0.9558	0.0652	2.2858	1.8172
AnySplat (Jiang et al., 2025)	∅	28.7384	0.9736	0.0529	1.0663	0.6619
DepthAnything 3 (Lin et al., 2026)	∅	25.3402	0.9595	0.0652	1.3341	1.1288
HiReFF (Jiang et al., 2026)	∅	28.3025	0.9697	0.0618	1.2153	0.5140
Forge4D (Hu et al., 2026)	∅	27.9093	0.9739	0.0359	1.5666	0.3063
Ours	∅	**30.1472**	**0.9742**	**0.0351**	**0.6637**	**0.2795**

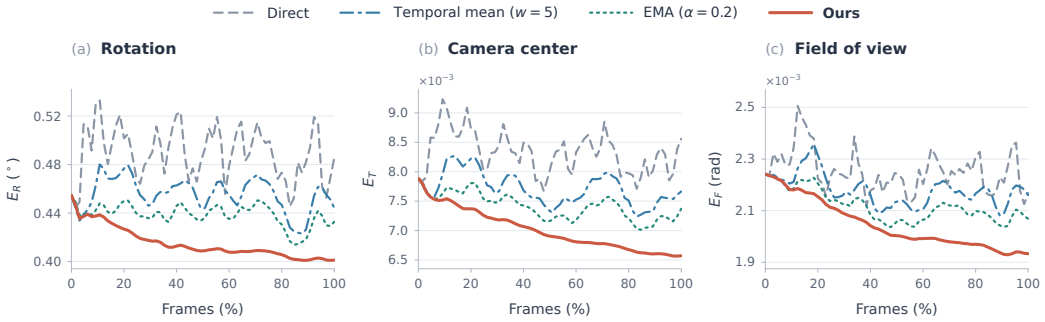

Figure 5: Camera errors over normalized sequence progress for four fixed cameras: (a) rotation, (b) camera center, and (c) FoV. Direct denotes the per-timestamp CameraHead prediction. Reliability-aware fusion exhibits a sustained overall decrease compared with the fluctuations of the evaluated temporal filters. FoV is shown in radians in panel (c).

4.3 Progressive Camera Estimation

We evaluate whether repeated observations improve estimates of the multi-view capture cameras. Here, $N = 4$ is the number of input cameras and k denotes the number of synchronized observations processed per camera. Camera evaluation is performed on 30 test sequences, using 150 synchronized timestamps per sequence; thus, $k_{\max} = 150$. We report the $SO(3)$ geodesic rotation error E_R in degrees, the camera-center error E_T normalized by the reference rig radius, and the mean absolute FoV error E_{FoV} in radians. Detailed metric definitions are provided in Appendix A.5.

We compare direct CameraHead predictions with a five-frame temporal mean, EMA ($\alpha = 0.2$), and our reliability-aware refinement, all using the same cached predictions and evaluation registration. As shown in Fig. 5, our method yields lower errors as observations accumulate and outperforms simple temporal filtering. Table 3 summarizes the initial and final errors. Relative to the direct CameraHead at $k_{\max}$, our method reduces rotation, camera-center, and FoV errors by 17.6%, 22.0%, and 12.0%, respectively.

4.4 Temporal Rendering Consistency

We evaluate temporal consistency from fixed held-out viewpoints using tOF and tLP (Chu et al., 2020), which measure motion and perceptual consistency between reconstructed and reference videos, respectively. As reported in Tables 1 and 2, CoSTA4D achieves the lowest temporal errors on both DNA-Rendering and the self-captured dataset, with substantial reductions in both tOF and tLP over competing methods. Combined with the strong PSNR and LPIPS results, these improvements demonstrate better temporal consistency while retaining high spatial reconstruction quality.

Table 3: Initial and final camera errors after k synchronized observations per camera, with $N = 4$ fixed. $k_{\max}$ is the common evaluated prefix length. Lower is better; bold and underlining mark the **Best** and Second-best final results.

Method	E_R (deg.)		E_T (normalized)		E_{FoV}	
	$k = 1$	$k_{\max}$	$k = 1$	$k_{\max}$	$k = 1$	$k_{\max}$
Direct CameraHead	0.4421	0.4856	0.00742	0.00856	0.00230	0.00217
Temporal mean ($w = 5$)	0.4421	0.4441	0.00742	0.00766	0.00230	0.00216
EMA ($\alpha = 0.2$)	0.4421	0.4327	0.00742	0.00737	0.00230	0.00207
Ours	0.4421	**0.4002**	0.00742	**0.00668**	0.00230	**0.00191**

Table 4: Component ablation on DNA-Rendering.

Variant	Temporal Aggregation	Camera Refinement	Gaussian Propagation	PSNR↑	LPIPS↓	tOF↓	tLP($\times$100) ↓
Frame-wise	×	×	×	28.1842	0.0406	3.2928	0.3347
Temporal	✓	×	×	30.1684	0.0322	2.6472	0.2007
Temporal + Camera	✓	✓	×	31.4286	0.0294	2.0121	0.1968
Temporal + Gaussian	✓	×	✓	30.0821	0.0297	1.9416	0.1782
Full Model	✓	✓	✓	31.3429	0.0289	1.7954	0.1405

4.5 Component Ablations

We ablate the three main components of CoSTA4D on DNA-Rendering. As shown in Table 4, Anchor-Context Temporal Aggregation improves both spatial reconstruction and temporal consistency, demonstrating the benefit of historical context for current-frame inference. Reliability-Aware Camera Refinement further improves reconstruction through more stable geometric estimation, while Correspondence-Guided Gaussian Propagation mainly enhances temporal consistency with nearly unchanged spatial quality. Combining all three components yields the best temporal consistency while maintaining strong reconstruction quality, confirming their complementary roles.

5 Conclusion and Limitations

We presented CoSTA4D for streaming 4D human reconstruction from multi-view, uncalibrated sparse-view videos. By exploiting the temporal invariance of the capture cameras, CoSTA4D uses historical observations to stabilize current geometry inference, progressively refine the shared camera state, and improve cross-frame representation continuity. Experiments on DNA-Rendering and self-captured videos demonstrate strong spatial reconstruction quality and temporal consistency.

CoSTA4D currently focuses on multi-view cameras capture, where temporal camera invariance can be explicitly exploited. Extending the framework to more flexible camera configurations and broader real-world capture conditions is an interesting direction for future work.

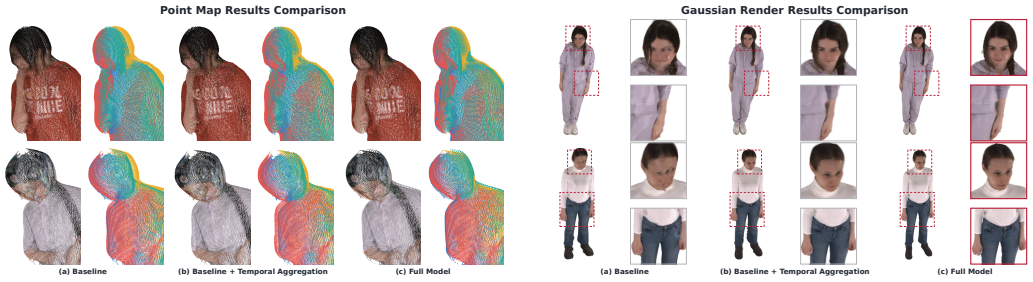

Figure 6: Qualitative component ablations on DNA-Rendering. We compare the frame-wise baseline, the variant with Anchor-Context Temporal Aggregation, and the full model. Left: point-map reconstructions. Right: Gaussian renderings with enlarged local details. Temporal aggregation improves geometric coherence, while the full model further refines geometry and rendering quality.

6 AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar, clarity, and readability of the manuscript. All AI-assisted revisions were reviewed by the authors to ensure accuracy and preserve the intended scientific meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Honggyu An, Jaewoo Jung, Mungyeom Kim, Chaehyun Kim, Minkyeong Jeon, Jisang Han, Kazumi Fukuda, Takuya Narihira, Hyunah Ko, Junsu Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. Learning compact 3D representations from feed-forward novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 363–373, 2026.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467, 2024.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22831–22840, 2025a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In *Computer Vision – ECCV 2024*, volume 15079 of *Lecture Notes in Computer Science*, pp. 370–386. Springer Nature Switzerland, 2024a.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. MVSplat360: Feed-forward 360 scene synthesis from sparse views. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107064–107086. Curran Associates, Inc., 2024b.

Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. LONG3R: Long sequence streaming 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5273–5284, 2025b.

Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. DNA-Rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19925–19936. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01829.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics*, 39(4):75, 2020.

Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3R-SfM: Towards feed-forward structure-from-motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16774–16784, 2025.

Hyunho Ha, Lei Xiao, Christian Richardt, Thu Nguyen-Phuoc, Changil Kim, Min H. Kim, Douglas Lanman, and Numair Khan. Geometry-guided online 3D video synthesis with multi-view temporal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11275–11285, 2025.

Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21594–21603, 2024.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2005–2015, 2025.

Yingdong Hu, Yisheng He, Jinnan Chen, Weihao Yuan, Kejie Qiu, Zehong Lin, Siyu Zhu, Zilong Dong, Steven Hoi, and Jun Zhang. Forge4D: Feed-forward 4D human reconstruction and interpolation from uncalibrated sparse-view videos. In *European Conference on Computer Vision (ECCV)*, 2026. Accepted at ECCV 2026.

Ranran Huang, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. From none to all: Self-supervised 3D reconstruction via novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37358–37369, June 2026.

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 (CVPR)*, pp. 4220–4230, 2024.

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

Yiming Jiang, Hanzhang Tu, Wenfeng Song, Siyou Lin, Liang An, Shuai Li, Aimin Hao, and Yebin Liu. HiReFF: High-resolution feedforward human reconstruction from uncalibrated sparse-view video. In *European Conference on Computer Vision (ECCV)*, 2026.

Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. Diffuman4D: 4D consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11047–11057. IEEE, 2025.

Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7233–7243, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *International Conference on Learning Representations*, volume 2026, pp. 42746–42768, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 71–91. Springer Nature Switzerland, 2024.

Chieh Lin, Zhaoyang Lv, Songyin Wu, Zhen Xu, Thu Nguyen-Phuoc, Hung-Yu Tseng, Julian Straub, Numair Khan, Lei Xiao, Ming-Hsuan Yang, Yuheng Ren, Richard Newcombe, Zhao Dong, and Zhengqin Li. DGS-LRM: Real-time deformable 3D gaussian reconstruction from monocular videos. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46915–46936. Curran Associates, Inc., 2025.

Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–9, 2022. doi: 10.1145/3550469.3555376.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, volume 2026, pp. 141261–141285, 2026.

Yaopeng Lou, Liao Shen, Tianqi Liu, Jiaqi Li, Zihao Huang, Huiqiang Sun, and Zhiguo Cao. MuGS: Multi-baseline generalizable gaussian splatting reconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25583–25593. IEEE, 2025.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3R: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22820–22830, 2025a.

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

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pp. 800–809. IEEE, 2024.

Xiaoyang Lyu, Muxin Liu, Xiaoshan Wu, Ruicheng Wang, Yi-Hua Huang, Yang-Tian Sun, Shaoshuai Shi, and Xiaojuan Qi. Stabilizing streaming video geometry via dynamic feature normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7577–7587, 2026.

Marko Mihajlovic, Sergey Prokudin, Siyu Tang, Robert Maier, Federica Bogo, Tony Tung, and Edmond Boyer. SplatFields: Neural gaussian splats for sparse 3D and 4D reconstruction. In *Computer Vision – ECCV 2024*, volume 15060 of *Lecture Notes in Computer Science*, pp. 313–332. Springer Nature Switzerland, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, January 2022. doi: 10.1145/3503250.

Riku Murai, Eric Dexheimer, and Andrew J Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16695–16705. IEEE, 2025.

Chris Rockwell, Nilesh Kulkarni, Linyi Jin, Jeong Joon Park, Justin Johnson, and David F. Fouhey. FAR: Flexible, accurate and robust 6DoF relative camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19854–19864, 2024.

Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22841–22852, 2025.

Hanzhang Tu, Zhanfeng Liao, Boyao Zhou, Shunyuan Zheng, Xilong Zhou, Liuxin Zhang, QianYing Wang, and Yebin Liu. GBC-Splat: Generalizable gaussian-based clothed human digitalization under sparse RGB cameras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26377–26387. IEEE, 2025.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21686–21697, 2024a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, 2026.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024b.

Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou. FreeTimeGS: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21750–21760, 2025c.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, efficient, accurate RAFT for optical flow. In *Computer Vision – ECCV 2024*, volume 15065 of *Lecture Notes in Computer Science*, pp. 36–54. Springer Nature Switzerland, 2024c.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D gaussian splatting for real-time dynamic scene rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320. IEEE, 2024.

Junjin Xiao, Qing Zhang, Yongwei Nie, Lei Zhu, and Wei-Shi Zheng. RoGSplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5980–5990. IEEE, 2025.

Yusen Xie, Zhenmin Huang, Jianhao Jiao, Dimitrios Kanoulas, and Jun Ma. UniGS: Unified geometry-aware gaussian splatting for multimodal rendering. In *Advances in Neural Information Processing Systems*, 2026.

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

Jiale Xu, Shenghua Gao, and Ying Shan. FreeSplatter: Pose-free gaussian splatting for sparse-view 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25442–25452, 2025b.

Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv. 4DGT: Learning a 4D gaussian transformer using real-world monocular videos. In *Advances in Neural Information Processing Systems*, volume 38, pp. 42606–42632. Curran Associates, Inc., 2025c.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4D gaussian splatting. In *International Conference on Learning Representations*, volume 2024, pp. 9142–9159, 2024a.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20331–20341, 2024b.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, volume 2025, pp. 54009–54033, 2025.

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

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. GPS-Gaussian: Generalizable pixel-wise 3D gaussian splatting for real-time human novel view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19680–19690. IEEE, 2024.

Yihao Zhi, Chenghong Li, Hongjie Liao, Xihe Yang, Zhengwentai Sun, Jiahao Chang, Xiaodong Cun, Wensen Feng, and Xiaoguang Han. MV-Performer: Taming video diffusion model for faithful and synchronized multi-view performer synthesis. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–14, 2025. doi: 10.1145/3757377.3763935.

Boyao Zhou, Shunyuan Zheng, Hanzhang Tu, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. GPS-Gaussian+: Generalizable pixel-wise 3D gaussian splatting for real-time human-scene rendering from sparse views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2025. doi: 10.1109/TPAMI.2025.3561248.

Junsheng Zhou, Zhifan Yang, Liang Han, Wenyuan Zhang, Kanle Shi, Shenkun Xu, and Yu-Shen Liu. 4C4D: 4 camera 4D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11829–11839, 2026.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pp. 88055–88072, 2026.