Forge4D: Feed-Forward 4D Human Reconstruction and Interpolation from Uncalibrated Sparse-View Videos

Yingdong Hu*[1], Yisheng He*✉[2], Jinnan Chen[3], Weihao Yuan[2], Kejie Qiu[2], Zehong Lin[4], Siyu Zhu[5], Zilong Dong[2], Steven Hoi[2], and Jun Zhang[1]

[1] The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2] Tongyi Lab, Alibaba Group, Hangzhou, China
[3] National University of Singapore, Singapore
[4] Fudan University, Shanghai, China
[5] School of Data Science, Lingnan University, Hong Kong SAR, China
yingdong.hu@connect.ust.hk, ethanheysh@gmail.com

Abstract. Instant reconstruction of dynamic humans and 3D human-object interactions from uncalibrated sparse-view videos is critical for numerous downstream applications. Existing methods, however, are either limited by the slow reconstruction speeds or incapable of generating novel-time representations. To address these challenges, we propose *Forge4D*, a feed-forward 4D human-object reconstruction and interpolation model that efficiently reconstructs temporally aligned representations from uncalibrated sparse-view videos, enabling both novel-view and novel-time synthesis. Our model simplifies the 4D reconstruction and interpolation problem as a joint task of streaming 3D Gaussian reconstruction and dense motion prediction. For the task of streaming 3D Gaussian reconstruction, we introduce learnable state tokens to enforce temporal consistency in a memory-friendly manner. For novel-time synthesis, we design a novel motion prediction module to predict dense motions for each 3D Gaussian between two adjacent frames. To overcome the lack of the ground truth for dense motion supervision, we formulate dense motion prediction as a dense point matching task and introduce a self-supervised *retargeting loss* to optimize this module. An additional occlusion-aware *optical flow loss* is introduced to ensure motion consistency with plausible human and object movement, providing stronger regularization. Extensive experiments demonstrate the effectiveness of our model on both in-domain and out-of-domain datasets. Code and models will be made publicly available.

Keywords: Feed-forward 4D Human Reconstruction · 4D Gaussian Splatting · Novel View Synthesis

[1] * indicates equal contribution.
[2] ✉ Corresponding author. Contact:ethanheysh@gmail.com

1 Introduction

Instant 4D human-object reconstruction from uncalibrated sparse-view video streams is essential for various application scenarios, including real-time livestreaming [56], sports broadcasting, augmented/virtual reality (AR/VR) [2], articulation modeling [4,11,28,63], and immersive holographic communication [43,20,12]. However, this task remains challenging due to the inherent difficulty of simultaneously recovering accurate human-object geometry and dense motion trajectories from unposed sparse-view video streams, while maintaining the real-time interactivity required for practical applications. For example, holographic communication systems demand high interactability, while sports broadcasting requires the ability to present novel views at any time for enhanced viewing experiences and precise evaluation of athletic performance.

Existing works [66,24,17] typically rely on iterative optimization over entire dense-view video sequences for each scene. These approaches depend heavily on calibrated camera parameters and suffer from prolonged training durations required for 4D representation convergence. Meanwhile, recent visual geometry grounded models [45,47,49] have enabled intermediate 3D point cloud reconstruction and camera pose estimation from arbitrary long uncalibrated image sequences in a feed-forward manner. However, the inherent limitations of point cloud representations restrict their ability to synthesis photorealistic novel-view images. Subsequent works [15,60] have extended feed-forward reconstruction models to predict static 3D Gaussians (3DGS), enabling photorealistic novel-view synthesis. Nevertheless, these methods remain incapable of handling dynamic scenes and synthesizing novel-time images.

In this work, we propose Forge4D, the first *feed-forward* model for *metric-scale* 4D human-object reconstruction that enables *novel-view* and *novel-time* synthesis from *uncalibrated sparse-view* videos in an *efficient streaming manner*. Our framework enables: 1) efficient reconstruction of temporally consistent 3D Gaussian assets in real-world metric scale from streaming sparse-view video inputs, and 2) accurate frame-wise dense 3D motion prediction for human-object subjects and 4D interpolation for novel-time synthesis. To achieve these goals, we decompose the 4D reconstruction and interpolation problem into two tasks: streaming 3D Gaussian reconstruction and dense motion prediction. This design offers two advantages: 1) it simplifies the problem for feed-forward regression, and 2) the reconstructed streaming 3DGS provide visual supervision for accurate dense motion prediction. Specifically, for streaming 3D Gaussian reconstruction, we leverage the pretrained knowledge prior from the large 3D reconstruction model VGGT [45] and adapt it to predict streaming key-frame 3D Gaussian assets. This adaptation is non-trivial due to two major challenges. First, a scale discrepancy exists between VGGT's output and the real-world metric scale inherent in ground-truth camera extrinsics, causing fundamental misalignment and unstable optimization with novel-view photometric loss. Second, naively feeding VGGT with multiple video frames suffers from long reconstruction duration, low interactivity, and out-of-memory (OOM) issues due to increasing image tokens for global attention. To address the scale issue, we propose to maintain a *metric*

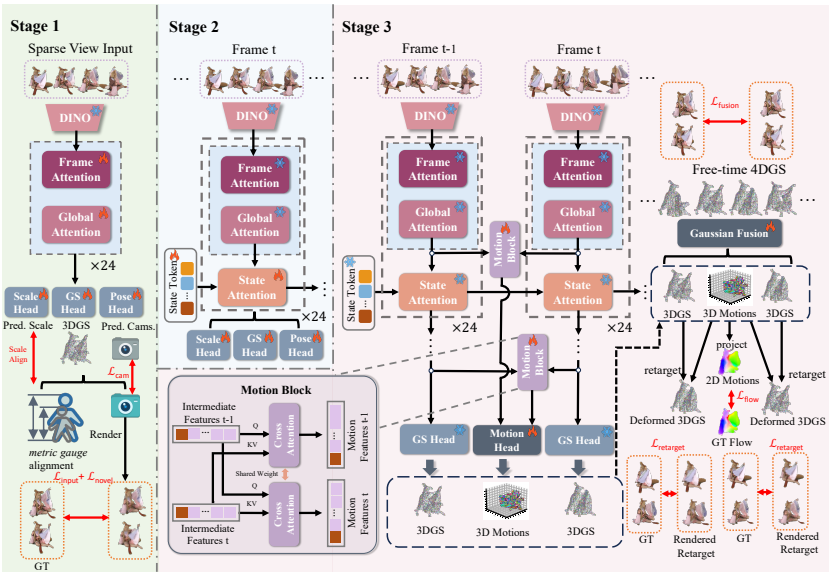

Fig. 1. The overall pipeline of *Forge4D*. It is trained in three stages: (1) static feed-forward 3D Gaussian reconstruction stage; (2) a streaming stage temporally aligned via state tokens; and (3) a feed-forward 4D reconstruction stage that predicts dense motion for each 3D Gaussian and interpolates free-time 3DGS using an occlusion-aware fusion process.

gauge and force the model to generate a temporally consistent scale in Sec. 3.2, which not only improves the stability under novel-view supervision, but also enables metric measurement. For the efficiency and OOM problem, we decompose the spatial and temporal dimensions of sparse-view videos and propose state tokens in Sec. 3.3 to iteratively incorporate temporal information in a streaming manner.

For dense motion prediction and novel-time synthesis, we propose a dense motion prediction module in Sec. 3.4 to facilitate 3D representation synthesis at arbitrary intermediate timestamps. In contrast to prior approaches that merely leverage middle-frame photometric supervision, our key insight is to formulate the task of dense motion prediction as a 3D Gaussian point-matching problem. However, there is no ground truth dense 3D motion for supervision of this module. Therefore, we propose a novel *retargeting loss* that projects current 3DGS to adjacent frames with the predicted dense motion and supervises the rendered images against ground truth. This regularization optimizes the dense motion in a self-supervised manner. For a stronger regularization, we also propose an *occlusion-aware optical flow loss* to enhance the plausibility of predicted motions by supervise them with 2D optical flows from a prior model. Given the dense motion between two timestamps, we deform the dynamic 3DGS from the two nearest frames under a constant velocity assumption for novel-time synthesis.

These deformed representations are then merged using an occlusion-aware fusion MLP to reduce 3DGS numbers and avoid flickering artifacts. Experimental results on benchmark datasets demonstrate the efficiency and effectiveness of the proposed framework.

The main contributions of this work are summarized as:

- We propose the first feed-forward model for metric-scale 4D human-object reconstruction from uncalibrated sparse-view videos, enabling novel-view synthesis and novel-time 4D interpolation in an efficient streaming manner.
- Our model simplifies this task by decomposing it into sequential streaming 3D Gaussian prediction and dense motion estimation. The novel *metric gauge* regularization, *retargeting loss*, and *occlusion-aware optical flow loss* stabilize the optimization and significantly improve motion prediction, photorealistic novel-view and novel-time synthesis.
- We introduce a novel motion-guided, occlusion-aware Gaussian fusion method for 3D Gaussian interpolation, enabling novel-time synthesis and effectively mitigating flickering and jittering artifacts caused by temporal redundancy in dynamic 3D Gaussian representations.

2 Related Work

Dynamic Scene Reconstruction and Streaming. Dynamic scene reconstruction from multi-view videos is crucial for numerous real-world applications. Prior methods primarily focus on optimizing a unified 4D representation to match dense multi-view 2D observations either by incorporating temporal dimensions into spatial coordinates [18,66,53,9,19] or by deforming 3D representations from keyframes using dynamic factors [31,27,17,48,40]. Another line of research assumes causal inputs and reconstructs per-frame 3D representations in a streaming manner [10,30,58,39]. However, these methods suffer from a limited reconstruction speed and are sensitive to the number of input views. In contrast to these iterative optimization-based approaches, *Forge4D* introduces an efficient feed-forward model that reconstructs the entire 4D scene in a single forward pass from uncalibrated sparse videos, significantly enhancing interactivity and applicability to downstream tasks.

Feed-Forward Reconstruction. Recent advances in visual geometry models have demonstrated the capability of deep neural networks for 3D reconstruction from multi-view images in a feed-forward manner. DUSt3R [47], VGGT [45], and π^3 [49] enable direct regression of camera poses and 3D point maps in the first frame's coordinate space. To enable photorealistic novel-view synthesis, another line of works [3,6,67,14,5,42,52] directly predict static 3DGS [51,38,61,55,64,29] or textured meshes [22] from calibrated multi-view images. To alleviate the reliance on camera calibration, NoPosplat [60] and AnySplat [15] propose to reconstruct 3DGS from uncalibrated multi-view images. However, all these methods are limited to per-timestamp static reconstruction and cannot synthesize novel-time 3D representations. Although L4GM [37] extends the static Gaussian reconstruction framework [41] to feed-forward 4D reconstruction, it suffers

from low-resolution reconstructions and poor generalization on real-world subjects. Concurrent to our work, recent methods [57,26,25] target 4D Gaussian reconstruction from monocular calibrated videos, and StreamSplat [54] further extends to uncalibrated ones. However, these methods neither explore the critical connection between 4D reconstruction and 3D point matching nor formulate motion learning as an explicit geometric correspondence problem, resulting in suboptimal performance. In contrast, our model is specifically designed to recover detailed geometry, appearance, and dense frame-wise motion for human performance from uncalibrated multi-view videos. By reformulating 4D reconstruction as a 3D Gaussian point matching task and introducing specialized losses for motion retargeting and occlusion-aware flow alignment, our approach achieves superior reconstruction quality and temporal consistency.

Template-Based Human Reconstruction and Animation. Several methods [34,12,21,62,13,33,35,16] reconstruct animatable avatars from single-view or sparse observations with parametric human priors like SMPL-X [32] and FLAME [23], which are then driven by estimated SMPL-X parameters from videos during inference. However, they fail to model the dynamic garments and human-object interaction and results in stiff and unnatural rendered videos due to the fundamental limitation of parametric human model. The drifting errors of SMPL-X parameters estimation from videos also hinder the stable performance. In contrast, *Forge4D* regresses pixel-aligned 3DGS representations with dense motion prediction to capture dynamic subjects, completely eliminating reliance on human templates while maintaining compatibility with dynamic clothing and complex human-object interactions.

3 Method

3.1 Overview

This work utilizes a transformer-based model $\mathcal{D}_{\text{4D}}$ for feed-forward reconstruction and novel-time interpolation of dynamic 4D Gaussian $\mathcal{G}_{4D}$ from n sparse uncalibrated videos $\{\boldsymbol{I}_i^t\}_{i=0,t=0}^{n,k}$ with a consistent video length of k, which can be expressed in the form of:

$$\mathcal{G}_{4D} = \mathcal{D}_{\text{4D}}(\{\boldsymbol{I}_i^t\}_{i=0,t=0}^{n,k}). \tag{1}$$

However, the strong entanglement between object geometry and motion makes the direct regression of 4D Gaussians challenging. To address this issue, we propose to decompose the problem into a streaming 3D Gaussian reconstruction task and a dense motion prediction task, and develop a progressive training pipeline. As shown in Fig. 1, the proposed pipeline is composed of three stages: 1) a feed-forward static 3D reconstruction stage to reconstruct static 3DGS in the real-world metric scale, 2) a streaming dynamic reconstruction stage to reconstruct streaming 4D Gaussians in an efficient and memory-friendly way, and 3) a dense motion prediction and Gaussian fusion stage to enable novel-time synthesis. In the first stage, a novel metric gauge calculation method is

proposed to align the backbone output scale with the real-world scale, which is critical to more stable supervision of novel views. While directly applying the 3D Gaussian reconstruction pipeline to each time stamp suffers from scale misalignment between different times, the main purpose of the second stage is to align different scales across different time stamps. To this end, we propose a state token to encode information from former frames and interact with the immediate frame efficiently, while being memory-friendly. In the final stage, a novel motion prediction module is proposed, together with a novel dual frame *retargeting loss* and *occlusion-aware optical flow loss* that marry the task of dynamic Gaussian motion prediction to the task of point matching. To enable novel-time synthesis, an additional occlusion-aware Gaussian fusion procedure is proposed for better dynamic 3D Gaussian interpolation.

3.2 Stage 1: Feed-forward 3D Gaussian Reconstruction

Recent advances in large 3D reconstruction models have demonstrated remarkable capabilities in recovering colored point maps and camera poses from a set of uncalibrated images. However, the point cloud representation limits the capability for photorealistic synthesis. To address this issue, we propose a feed-forward 3D Gaussian reconstruction model $\mathcal{D}_{\text{3D}}$ that leverages the geometry prior within these foundations, while introducing a 3D Gaussian prediction branch for photorealistic rendering. Specifically, we use a pre-trained VGGT as our backbone and predict pixel-aligned 3DGS with an additional DPT [36] head as $\mathcal{G}^t_{3D} = \mathcal{D}_{\text{3D}}(\{\boldsymbol{I}^t_i\}^n_{i=0})$, where $\mathcal{G}^t_{3D} = \{\boldsymbol{P}^t_i, \boldsymbol{O}^t_i, \boldsymbol{C}^t_i, \boldsymbol{Q}^t_i, \boldsymbol{S}^t_i\}^n_{i=0}$, with $\boldsymbol{P}^t_i \in \mathbb{R}^{3\times H\times W}$, $\boldsymbol{O}^t_i \in \mathbb{R}^{1\times H\times W}_i$, $\boldsymbol{C}^t \in \mathbb{R}^{3\times H\times W}$, $\boldsymbol{Q}^t_i \in \mathbb{R}^{4\times H\times W}$, and $\boldsymbol{S}^t_i \in \mathbb{R}^{3\times H\times W}$ representing the Gaussian position, opacity, color, rotation, and scale maps, respectively, from view i of size $H \times W$ at time t.

To supervise this branch, photometric losses (e.g., L2, SSIM, LPIPS) are applied between the rendered and ground-truth (GT) images. However, a fundamental scale ambiguity occurs between the output scale of VGGT (i.e., normalized point clouds) and the real-world metric scale. Directly applying photometric supervision without addressing this scale discrepancy results in an unstable optimization trajectory and fails to converge to a coherent 3D structure, as further evaluated in Sec. 4.3. To resolve this issue, we introduce a *metric gauge* regularization term p_{gauge} to align the scale of the GT novel-view camera extrinsics with the model's internal coordinate system, thereby stabilizing training. This approach is grounded in the key insight that if the model's predicted camera poses and intrinsics are accurate, their difference from the GT poses should be primarily a consistent translation scaling factor, as visualized in Fig. 2. Formally, for n input cameras, we calculate the ratio

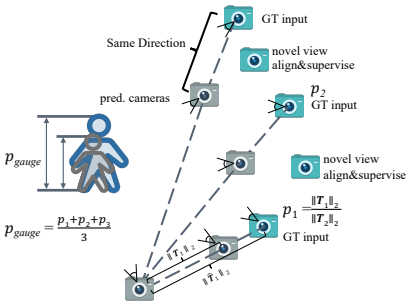

Fig. 2. Gauge illustration.

$p_i = \|\boldsymbol{T}_i\|_2 / \|\hat{\boldsymbol{T}}_i\|_2$ of translation magnitudes between each predicted camera and its GT counterpart. The novel-view cameras are then scaled using the mean of these ratios, $p_{\text{gauge}} = \frac{1}{n}\sum_{i=1}^{n} p_i$. This factor is also utilized for scale head supervision, which predicts $\hat{p}_{\text{gauge}}$ for metric scale recovery during evaluation. To ensure that the metric gauge accurately represents the scale difference and to simultaneously refine the predicted camera parameters, we propose a comprehensive camera loss:

$$\begin{aligned}\mathcal{L}_{\text{cam}} = &\sum_{i=0}^{n} \|\boldsymbol{q}_i - \hat{\boldsymbol{q}}_i\|_2 + \sum_{i=1}^{n} \left\| \frac{\boldsymbol{T}_i}{\|\boldsymbol{T}_i\|_2} - \frac{\hat{\boldsymbol{T}}_i}{\|\hat{\boldsymbol{T}}_i\|_2} \right\|_2 \\ &+ \sum_{i=1}^{n} |p_i - p_{\text{gauge}}| + |\hat{p}_{\text{gauge}} - p_{\text{gauge}}|,\end{aligned} \tag{2}$$

which supervises the model for accurate rotation $\boldsymbol{q}_i$, translation direction, consistent relative scaling. and the scale prediction header. Our full training objective then combines this metric-aware camera loss with multi-view photometric supervision. The photometric loss $\mathcal{L}_{\text{input}} = \sum_{i=0}^{n}(\|\boldsymbol{I}_i - \hat{\boldsymbol{I}}_i\|_2 + \lambda_{\text{SSIM}}\text{SSIM}(\boldsymbol{I}_i, \hat{\boldsymbol{I}}_i) + \lambda_{\text{LPIPS}}\text{LPIPS}(\boldsymbol{I}_i, \hat{\boldsymbol{I}}_i))$ is applied to the n input views, and a corresponding loss $\mathcal{L}_{\text{novel}}$ is applied to m held-out novel views, ensuring high-fidelity reconstruction across all perspectives. Thus, the total loss for our feed-forward 3D Gaussian reconstruction model is $\mathcal{L}_{\text{3D}} = \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{input}} + \mathcal{L}_{\text{novel}}$. Supervised by this combined objective, our model not only infers coherent 3DGS from uncalibrated RGB inputs, but also enables metric measuring, which we discuss thoroughly in Supplementary C.

3.3 Stage 2: Aligning Dynamic Human Streaming with State-token

In stage one, we obtain a feed-forward network for static 3D Gaussian reconstruction from sparse-view images. However, to apply for video scenarios, the output scale of this network is not aligned across timestamps, leading to temporal inconsistency. To address this, one vanilla way is to stack all sparse video tokens for global attention in VGGT, which results in OOM and low reconstruction speed issues. Instead, we decompose the spatial and temporal dimensions of sparse-view videos and introduce a state token to enforce the temporal consistency in an efficient and memory-friendly way. Specifically, we employ a learnable *state token* that iteratively encodes information from all previous frames and broadcasts it to the current frame. These tokens inject temporal information by serving as the Key and Value in a cross-attention layer applied to the current frame's features. Conversely, the token itself is updated by attending to the current frame's features, where it serves as the Query. To further enhance the temporal stability, we extend the metric gauge regularization in Sec. 3.2 to a temporal form. The global scale factor is now computed over all n cameras and k timestamps as $p_{\text{gauge}} = \frac{1}{(n-1)(k-1)}\sum_{t=1}^{k}\sum_{i=1}^{n} p_i^t$, which is also utilized to supervise the general gauge $\hat{p}_{\text{gauge}}^t$ prediction in the scale prediction header. Consequently, we generalize the camera loss to supervise the temporal cross-attention layers across the

entire sequence as:

$$\begin{aligned}\mathcal{L}_{\text{cam}} = &\sum_{t=0}^{k}\sum_{i=0}^{n}\|\boldsymbol{q}_i^t - \hat{\boldsymbol{q}}_i^t\|_2 + \sum_{t=0}^{k}\sum_{i=1}^{n}\left\|\frac{\boldsymbol{T}_i^t}{\|\boldsymbol{T}_i^t\|_2} - \frac{\hat{\boldsymbol{T}}_i^t}{\|\hat{\boldsymbol{T}}_i^t\|_2}\right\|_2 \\ &+ \sum_{t=0}^{k}\sum_{i=1}^{n}|p_i^t - p_{\text{gauge}}| + \sum_{t=0}^{k}|\hat{p}_{\text{gauge}}^t - p_{\text{gauge}}|.\end{aligned} \tag{3}$$

This ensures consistent camera rotation, translation direction, and global scale across time.

3.4 Stage 3: Novel-timestamp 3D Gaussian Synthesis with Dense Motion Prediction and Dynamic Gaussian Fusion

A general 4D representation should enable the synthesis of 2D images from arbitrary camera viewpoints at *any moment* in time. This requires the capability to interpolate the representation to intermediate timestamps beyond the input frames. In this work, we adopt dynamic 3DGS as our fundamental 4D representation and assume a linear motion model that propagates 3DGS between consecutive frames, following previous work on 4D reconstruction [48]. Mathematically, we formulate our 4D Gaussian representation as:

$$\mathcal{G}_{4D} = \{\{\mathcal{G}_{3D}^t\}_{t=0}^{k}, \{\boldsymbol{M}_{i,\{1,2\}}^t\}_{i=0,t=0}^{n,k}, \boldsymbol{F}_\theta\}, \tag{4}$$

where $\mathcal{G}_{3D}^t$ denotes the 3D Gaussian attribute maps at time t, $\boldsymbol{M}_{i,\{1,2\}}^t \in \mathbb{R}^{2\times 3\times H\times W}$ represents the associated 3D motion field for view i, and $\boldsymbol{F}_\theta$ is a learnable fusion function that adaptively combines Gaussian attributes at novel timestamps in account of occlusion relationships.

To predict the 3D motion map $\boldsymbol{M}_{i,\{1,2\}}^t$, we introduce a dense motion prediction block that operates on the static streaming reconstruction described in Sec. 3.2. This block predicts a dense, pixel-aligned dual motion field for each 3D Gaussian. Specifically, when processing a new frame at time t, the block infers: (1) a backward 3D motion $\boldsymbol{M}_{i,1}^t \in \mathbb{R}^{3\times H\times W}$ that warps the current 3DGS (time t) to the previous timestamp $t-1$, and (2) a forward 3D motion $\boldsymbol{M}_{i,2}^{t-1} \in \mathbb{R}^{3\times H\times W}$ that warps the 3DGS from the previous frame (time $t-1$) to the current frame t. This process is repeated symmetrically when the subsequent frames arrive. The proposed block comprises the same number of attention blocks as the backbone model. Each motion attention block takes as input the corresponding intermediate features from frames t and $t-1$ produced by the backbone. The output features from all motion attention blocks are aggregated and passed to a motion DPT head to produce the final dual motion prediction $\boldsymbol{M}_{i,\{1,2\}}^t$.

Given 3D Gaussian assets at two successive frames t and $t-1$, along with their corresponding motions, the 3DGS at the middle timestamp t' can be acquired by warping these two frames with a constant velocity assumption. This assumption is a first-order simplification of the polynomial 4D Gaussians [**?**], and

is empirically found to perform adequately for small temporal intervals in our cases. Specifically, for each 3D Gaussian in frame t, it is deformed to the middle timestamp by adding a displacement proportional to its temporal distance to time t. The 3DGS in frame $t-1$ are also deformed to this middle timestamp t' in the same way. This deforming process can also be represented as:

$$\begin{aligned} \boldsymbol{P}_i^{t\to t'} &= \boldsymbol{P}_i^t + |t'-t| \cdot \boldsymbol{M}_{i,1}^t, \boldsymbol{P}_i^{t-1\to t'} \\ &= \boldsymbol{P}_i^{t-1} + |t'-(t-1)| \cdot \boldsymbol{M}_{i,2}^{t-1}. \end{aligned} \tag{5}$$

However, there are no ground truth dense motions for supervision. To effectively train this block and the deformation process, we introduce a novel *retargeting loss* that optimizes the predicted 3D motion using 2D photometric constraints in a self-supervised manner. Specifically, for 3DGS at time t, we deform their positions to time $t-1$ with the prediction dense motion as $\boldsymbol{P}_i^{t\to t-1} = \boldsymbol{P}_i^t + \boldsymbol{M}_{i,1}^t$, while retaining other attributes. This forms an intermediate 3D Gaussian representation $\mathcal{G}_{3D}^{t\to t-1} = \{\boldsymbol{P}_i^{t\to t-1}, \boldsymbol{O}_i^t, \boldsymbol{C}_i^t, \boldsymbol{Q}_i^t, \boldsymbol{S}_i^t\}_{i=0}^n$. Since $\boldsymbol{M}_{i,1}^t$ aims to recover the true motion between frames, $\mathcal{G}_{3D}^{t\to t-1}$ should closely align with the ground-truth Gaussians $\mathcal{G}_{3D}^{t-1}$. which can be supervised with rendering consistency. The retargeting loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{retarget}} = \sum_{i=0}^{n} (&||\hat{\boldsymbol{I}}_i^{t\to t-1} - \hat{\boldsymbol{I}}_i^{t-1}||_2 + \lambda_{\text{S}}\text{SSIM}(\hat{\boldsymbol{I}}_i^{t\to t-1}, \hat{\boldsymbol{I}}_i^{t-1}) \\ &+ \lambda_{\text{L}}\text{LPIPS}(\hat{\boldsymbol{I}}_i^{t\to t-1}, \hat{\boldsymbol{I}}_i^{t-1})), \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{I}}_i^{t\to t-1} = \mathcal{R}(\mathcal{G}_{3D}^{t\to t-1}, \boldsymbol{E}_i, \boldsymbol{K}_i)$ and $\hat{\boldsymbol{I}}_i^{t-1} = \mathcal{R}(\mathcal{G}_{3D}^{t-1}, \boldsymbol{E}_i, \boldsymbol{K}_i)$ denote rendered images for view i, with $\mathcal{R}$ representing the rendering function for 3D Gaussian Splatting, and $\boldsymbol{E}_i$, $\boldsymbol{K}_i$ denoting camera extrinsics and intrinsics, respectively.

To ensure real-world plausibility and resolve ambiguities, we incorporate an *occlusion-aware optical flow loss* for a stronger regularization. We compute pseudo-ground-truth flow $\boldsymbol{\mu}_i^{t\to t-1}$ using SEA-RAFT [50] and project the predicted 3D motion $\boldsymbol{M}_{i,1}^t$ to 2D scene flow as $\hat{\boldsymbol{\mu}}_i^{t\to t-1}$. A cyclic consistency mask $\mathbf{1}_{\text{cyc}}$ penalizes inconsistencies between forward and backward flows and removes occluded regions. The flow loss is defined as:

$$\mathcal{L}_{\text{flow}} = \sum_{i=0}^{n} \mathbf{1}_{\text{cyc}}(\boldsymbol{\mu}_i^{t\to t-1}, \boldsymbol{\mu}_i^{t-1\to t}) \cdot ||\boldsymbol{\mu}_i^{t\to t-1} - \hat{\boldsymbol{\mu}}_i^{t\to t-1}||_2, \tag{7}$$

where $\mathbf{1}_{\text{cyc}}(\boldsymbol{\mu}_i^{t\to t-1}, \boldsymbol{\mu}_i^{t-1\to t}) = \exp(-\boldsymbol{r}_i^t \cdot ||\boldsymbol{\mu}_i^{t\to t-1} + \boldsymbol{\mu}_i^{t-1\to t}[\boldsymbol{p}_i^t + \boldsymbol{\mu}_i^{t-1\to t}]||_2)$ acts as an occlusion-aware weighting term, $\boldsymbol{r}_i^t$ is a hyperparameter related to the length of each flow, and $\boldsymbol{\mu}_i^{t-1\to t}[\boldsymbol{p}_i^t + \boldsymbol{\mu}_i^{t-1\to t}]$ represents a pixel-wise indexing process. The *retargeting loss* and *occlusion-aware optical flow loss* are combined together as a matching supervision: $\mathcal{L}_{\text{matching}} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{retarget}}$.

Table 1. Quantitative results of novel-view synthesis of static reconstruction.

Dataset				DNA-Rendering			Genebody		
Method	Type		w. Cam. pose	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DualGS	Optimization	4D*	Yes	18.7408	0.8932	0.1515	16.1721	0.7663	0.1891
D-3DGS	Optimization	4D	Yes	20.7767	0.8944	0.1162	15.7067	0.7981	0.1790
Queen	Streaming	3D	Yes	15.4966	0.8941	0.1213	16.4514	0.9484	0.0710
GPS-Gaussian	Feed-Forward	3D	Yes	24.2963	0.9247	0.0867	25.1734	0.9346	0.0756
NoPoSplat	Feed-Forward	3D	**No**	11.7632	0.8092	0.2846	13.9554	0.8721	0.1664
AnySplat	Feed-Forward	3D	**No**	26.1157	0.9430	0.1513	25.8010	0.9287	0.1355
Ours	Feed-Forward	4D	**No**	**29.8167**	**0.9606**	**0.0542**	**28.0819**	**0.9523**	**0.0548**

Table 2. Quantitative results of novel-time and novel-view synthesis.

Dataset			DNA-Rendering			Genebody		
Method	Type	w. Cam. pose	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
D-3DGS	Optimization	Yes	20.9158	0.8994	0.1163	15.5823	0.8002	0.1502
SpaceTimeGS	Optimization	Yes	17.2189	0.8879	0.1247	14.6119	0.8488	0.1805
L4GM	Feed-Forward	Yes	18.0325	0.9152	0.1367	14.8572	0.9144	0.1727
Ours	Feed-Forward	**No**	**29.0378**	**0.9566**	**0.0535**	**27.4247**	**0.9459**	**0.0601**

Additionally, to naturally fuse the two sets of deformed 3DGS $\mathcal{G}_{3D}^{t\to t'}$, $\mathcal{G}_{3D}^{t-1\to t'}$ while preserving 3DGS from occluded regions, we further deform the 3D Gaussian with a *dual consistency factor* $\boldsymbol{D}_i^t$ (or $\boldsymbol{D}_i^{t-1}$ for frame $t-1$). This factor is calculated by measuring the distance of the deformed concurrent 3D Gaussian point to the retrieved 3D Gaussian point at the retargeted frame via the projected 2D flow, as shown in Fig. 3. This factor serves as the guidance of areas masked in the next frame. 3DGS with a factor larger than a threshold τ will be kept as occluded 3DGS $\{\bar{\mathcal{G}}_{3D}^t, \bar{\mathcal{G}}_{3D}^{t-1}\}$, while the other 3DGS $\{\hat{\mathcal{G}}_{3D}^t, \hat{\mathcal{G}}_{3D}^{t-1}\}$ from two nearby timestamps will be merged into one by a two-layer MLP $\boldsymbol{F}_\theta$ to eliminate temporal redundancy. The final 3D Gaussian assets $\mathcal{G}_{3D}^{t'}$ are merged with the remaining occluded 3DGS and the fused 3DGS as:

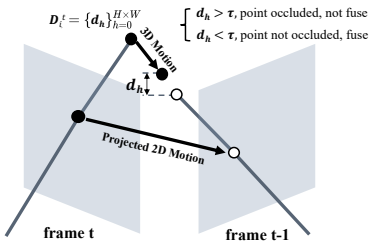

Fig. 3. Illustration of dual consistency factor. We identify the occluded points with a predicted 2D/3D motion consistency.

$$\mathcal{G}_{3D}^{t'} = \{\bar{\mathcal{G}}_{3D}^{t\to t'}, \bar{\mathcal{G}}_{3D}^{t-1\to t'}, \boldsymbol{F}_\theta(\hat{\mathcal{G}}_{3D}^{t\to t'}, \hat{\mathcal{G}}_{3D}^{t-1\to t'})\}. \tag{8}$$

This fusion process is supervised by the photometric loss at novel time t', defined as $\mathcal{L}_{\text{fusion}} = \sum_{i=0}^{n}(\|\boldsymbol{I}_i^{t'} - \hat{\boldsymbol{I}}_i^{t'}\|_2 + \lambda_{\text{S}}\text{SSIM}(\boldsymbol{I}_i^{t'}, \hat{\boldsymbol{I}}_i^{t'}) + \lambda_{\text{L}}\text{LPIPS}(\boldsymbol{I}_i^{t'}, \hat{\boldsymbol{I}}_i^{t'}))$. We supervise Stage 3 with the loss function $\mathcal{L}_{4D} = \mathcal{L}_{\text{matching}} + \mathcal{L}_{\text{fusion}}$.

In this way, our model achieves the task of generalized 4D human-object reconstruction by decomposing it into static 3D Gaussian streaming and dense motion prediction. High-quality novel-view images at any novel time can be acquired by interpolating the predicted dynamic 3DGS and then rendering onto the corresponding image planes.

4 Experiment

4.1 Experimental Settings

Datasets. *Forge4D* is trained on the DNA-Rendering [7] training set, which comprises 2,078 human-object interaction video sequences that exhibit diverse subject ages, appearances, and motion patterns. 4D synthesis is evaluated on three benchmarks: 1) an in-domain held-out test set that contains all sequences from 10 distinct identities in DNA-Rendering, 2) the out-of-domain complete Genebody [8] dataset, and 3) an in-the-wild benchmark composed of two casually captured datasets: SelfCap [48] and VVC [44]. For motion and metric prediction, due to a lack of ground truth annotations in real-world datasets, we construct a synthetic dataset, MetaHuman4D, with ground truth annotations for evaluation. The details of MetaHuman4D are in the Supplementary B.

Evaluation Metrics. The synthesized image quality is measured using standard metrics: PSNR, SSIM, and LPIPS at a resolution of 518×518, unless specific ones are required by architectural constraints, such as 512×512 for GPS-Gaussian and L4GM. All models are trained and evaluated using 4 input views with a camera angle of around 45°, which ensures sufficient coverage of the frontal human appearance. The dense motion prediction task is benchmarked using the L2 distance and the retargeted point distance. The metric scale prediction is evaluated by computing the L2 distance between predicted 3D points and their nearest corresponding points on the GT scale mesh.

Baselines. We establish comparisons under two experimental settings. For novel-view synthesis at input timestamps, we compare against optimization-based methods (DualGS [17], Queen [10], D-3DGS [59]) and feed-forward models (GPS-Gaussian [67]), along with pose-free feed-forward methods (NoPosplat [60], AnySplat [15]). For novel-time interpolation quality, we compare exclusively with methods capable of generating 3D representations at non-input timestamps: SpaceTimeGS [24], D-3DGS [59], and L4DM [37]. We provide comparisons with template-based methods in the Supplementary, noting that the methodological differences prevent a direct comparison.

Implementation Details. We initialize the backbone using pre-trained weights from VGGT. The scale head, 3D Gaussian position offset head, color offset head, scene attention blocks, and motion blocks are zero-initialized. See more details in the Supplementary C.

4.2 Evaluation

Evaluation on 3D Reconstruction and Novel-View Synthesis. As demonstrated in Tab. 1, our model outperforms all baseline methods across all metrics. Specifically, *Forge4D* surpasses previous pose-free feed-forward 3D reconstruction models, i.e., NoPoSplat and AnySplat, by up to +2.28 dB in PSNR. This significant performance gap stems not only from artifacts caused by multi-view mismatches and geometric inaccuracies but also from the inability of these methods to generate 3D models and camera configurations that align with the ground-truth scale and camera parameters. Moreover, the extensive white background

Fig. 4. Qualitative results of 3D reconstruction on test sets. Our *Forge4D* exhibits more stable synthesized novel-view images against artifacts, including blur, ghosting, and shape distortion.

Table 3. In-the-wild accessments.

Method	Novel-View			Novel-Time		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
LHM	13.3687	0.7867	0.3942	-	-	-
GPS-Gaussian	25.6316	0.9065	0.1019	-	-	-
AnySplat	23.9645	0.8930	0.1502	-	-	-
L4GM	15.4771	0.8494	0.2111	14.8831	0.8170	0.1951
Forge4D	26.2735	0.9160	0.0917	25.4726	0.9009	0.1025

Table 4. Metric Scale prediction evaluation.

Method	Point Distance
MoGe-2	0.3309m
Ours	**0.0264m**

Table 5. Motion prediction evaluation.

Method	Motion Error	Point Distance
POMATO	0.01274	0.7555
Ours	**0.00953**	**0.0215**

in human images further complicates the synthesis of plausible renderings, even with the camera optimization procedure described in Supplementary C. Compared to posed feed-forward 3D reconstruction models, *Forge4D* also exceeds the previous state-of-the-art method, GPS-Gaussian, by up to +2.90 dB in PSNR, which originates from *Forge4D*'s ability to produce more geometrically faithful reconstructions of challenging regions such as hands, heads, and accessories. Qualitative results in Fig. 4 further confirm that *Forge4D* delivers more photorealistic novel views without artifacts such as blur, ghosting, or shape distortion.

Evaluation on 4D Reconstruction and Novel-Time Synthesis. Quantitative and qualitative results in Tab. 2 and Fig. 5 demonstrate *Forge4D*'s effectiveness in novel-time synthesis. The model exhibits strong capabilities in both generating 3D Gaussian assets and interpolating 3DGS for arbitrary intermediate timestamps between key frames. Our approach outperforms the previous state-of-the-art feed-forward 4D reconstruction model L4GM on both datasets with performance gains of up to +12.57 dB in PSNR. In comparison with optimization-based methods, our method surpasses all previous approaches, as these methods fail to generate reasonable novel views under such sparse-view conditions. The optimization-based techniques struggle with the limited input views, while our feed-forward approach maintains robust performance even with sparse camera arrangements.

Evaluation on In-the-wild Dataset. We evaluate the in-the-wild applicability of *Forge4D* on two casually captured datasets: SelfCap [48] and VVC [44].

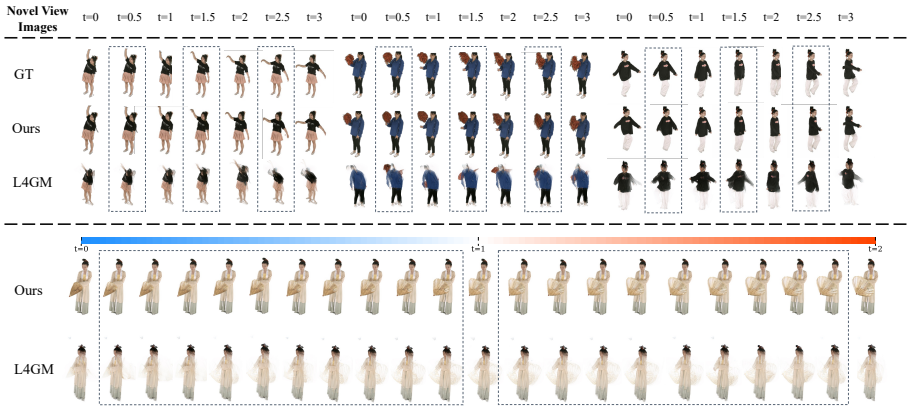

Fig. 5. Qualitative results of 4D reconstruction on novel-view and novel-time synthesis. *Forge4D* accurately reconstructs 3DGS at input frames while generating plausible intermediate 3DGS *at any time* with high-fidelity rendering quality (images in dashes).

From SelfCap, we select yoga and corgi sequences that effectively capture full-body human motion with objects, while using the complete validation and test sets from VVC. This comprehensive evaluation covers 9 diverse human-object interaction scenarios, representing a wide spectrum of real-world situations. We select 4 cameras with diverse viewing angles ranging from 20 to 70 degrees as model inputs, corresponding to their spatial arrangement in each test sequence. We provide quantitative comparisons with GPS-Gaussian [67] in Tab. 3 and qualitative comparisons in Fig. 6.

Evaluation on Dense Motion Prediction. We evaluate our motion prediction module using our MetaHuman4D dataset containing multi-view images and dense ground-truth motions derived from mesh correspondences for each frame. The predicted 3D motions are compared against the GT 3D motions using L2 distance, and benchmarked against the state-of-the-art dense motion prediction model POMATO [65]. Additionally, we report the retargeted point distance to the GT target-time mesh, where *Forge4D* consistently generates plausible outputs while POMATO fails to produce reasonable human geometry. Quantitative results in Tab. 5 demonstrate the effectiveness of our motion prediction framework. Further details are provided in Supplementary B.

Evaluation on Metric Scale Prediction. *Forge4D* is able to recover real-world metric scale points as a byproduct of the scale prediction header supervised with *metric gauge*. We evaluate the performance of *Forge4D* in metric scale recovering by measuring the mean distance of the predicted points to the GT human mesh on MetaHuman4D, which results in a 0.02 m error on average. A comparison with MoGe-2 is made in the same metric, with the results presented in Tab. 4. *Forge4D* outperforms MoGe-2 in mean distance to ground-truth mesh, primarily due to MoGe-2's inability to effectively align multi-view point correspondences.

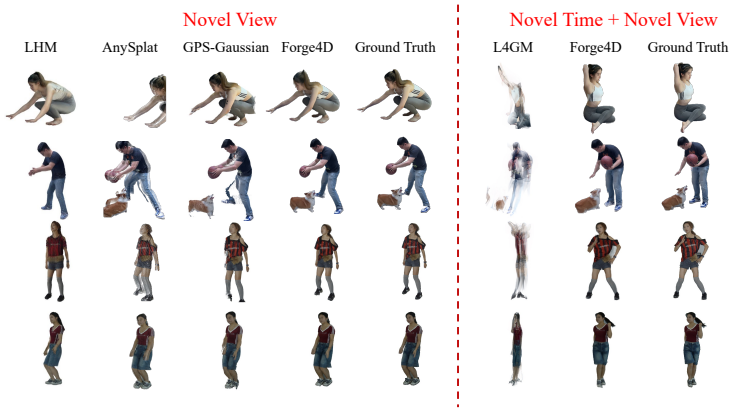

Fig. 6. Qualitative results on in-the-wild datasets of the human-centric methods.

Table 6. Ablation on model speed.

Module	Delay	Frame Rate
Key-frame Reconstruction	176.50 ms	-
Motion Prediction	47.77 ms	-
Interpolation (10 Steps)	1.46 ms	-
Full Model	224.27 ms	4.45 FPS
+Interpolate 10 Steps	225.10 ms	44.42 FPS
+Interpolate 20 Steps	226.01 ms	88.48 FPS

Table 7. Ablation on different components.

Evaluation Mode	Variants	PSNR↑	SSIM↑	LPIPS↓
Static Novel View	**Full Model**	29.8167	0.9606	0.0542
	w/o *gauge aligning*	13.2884	0.1194	0.2184
Dynamic Novel Time + Novel View	**Full Model**	29.0378	0.9566	0.0535
	w/o *state token*	28.5555	0.9513	0.0592
	w/o *retargeting loss*	28.4124	0.9530	0.0573
	w/o *optical flow loss*	28.8676	0.9551	0.0563
	w/o *Gaussian Fusion*	27.3459	0.9346	0.0683

4.3 Ablation Study

Ablations on Methodology. We show the effectiveness of the proposed *metric gauge*, the *retargeting loss*, the *occlusion-aware optical flow loss*, and the Gaussian fusion process in Tab. 7. A significant performance gap is observed when the *retargeting loss* and the occlusion-aware *optical flow loss* are replaced with direct supervision on the novel timestamps, and the training process will lead to a collapse when the *metric gauge* alignment is missing. Additionally, directly averaging the 3DGS from the nearby 2 frames will also lead to a suboptimal novel-view image quality. We provide novel-view and novel-time visualizations for the ablation study in Sec. 4.3 in Figure 7. The results demonstrate that without state token updates, temporal misalignment occurs at novel timestamps, while the absence of either retargeting loss or optical flow loss leads to unreliable human motion estimation, and vanilla averaging the Gaussian attributes leads to temporal redunancy and flickering issues. We also include a comparison of the Temporal Luminance Standard Deviation (TL-STD↓) to illustrate these flickering issues, where a higher TL-STD indicates more noticeable brightness jitters. Without Gaussian Fusion, the average TL-STD (0.9446) is 75.6% higher than that with Gaussian Fusion (0.5379).

Ablations on Model Speed. We evaluate the inference speed of *Forge4D* on an NVIDIA H200 Tensor Core GPU, with the results reported in Tab. 6. The key-frame reconstruction requires 176.50 ms per inference, the motion prediction

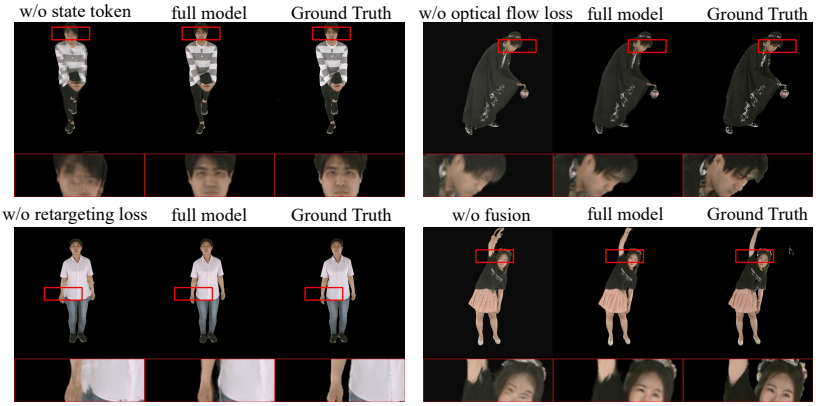

Fig. 7. Visualizations on the ablation study.

module takes 47.77 ms, and the intermediate-time interpolation for 10 frames adds 1.46 ms, resulting in a total latency of 224.27 ms per input frame pair. Given our model's capability for arbitrary-length interpolation, the effective output frame rate reaches 44 FPS when generating 10 interpolated frames per input interval.

5 Conclusion

We propose *Forge4D*, the first feed-forward model for 4D human-object reconstruction from uncalibrated sparse-view videos. Our approach simplifies the problem by decomposing it into streaming real-world metric-scale 3D Gaussian reconstruction and dense motion prediction for novel-time synthesis. *Forge4D* achieves state-of-the-art novel-view synthesis quality and enables interpolation to arbitrary timestamps while maintaining plausible intermediate representations. Nevertheless, *Forge4D* exhibits certain limitations. The performance degrades in the presence of large motions or longer inter-frame intervals, primarily due to reduced inter-frame correspondences and violations of the consistent motion assumption. We identify these limitations as directions for future work, focusing on improving motion modeling under extreme displacements and optimizing computational efficiency for better interactivity.

References

1. Baradel*, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., Lucas*, T.: Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In: ECCV (2024)
2. Carmigniani, J., Furht, B.: Augmented reality: an overview. Handbook of augmented reality pp. 3–46 (2011)

3. Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR (2024)
4. Chen, J., Li, C., Lee, G.H.: Weakly-supervised 3d pose transfer with keypoints. arXiv preprint arXiv:2307.13459 (2023)
5. Chen, J., Li, C., Zhang, J., Zhu, L., Huang, B., Chen, H., Lee, G.H.: Generalizable human gaussians from single-view image. arXiv preprint arXiv:2406.06050 (2024)
6. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. arXiv preprint arXiv:2403.14627 (2024)
7. Cheng, W., Chen, R., Yin, W., Fan, S., Chen, K., He, H., Luo, H., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., Ren, D., Yang, L., Liu, Z., Loy, C.C., Qian, C., Wu, W., Lin, D., Dai, B., Lin, K.Y.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. arXiv preprint **arXiv:2307.10173** (2023)
8. Cheng, W., Xu, S., Piao, J., Qian, C., Wu, W., Lin, K.Y., Li, H.: Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. arXiv preprint arXiv:2204.11798 (2022)
9. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d-rotor gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. In: Proc. SIGGRAPH (2024)
10. Girish, S., Li, T., Mazumdar, A., Shrivastava, A., Luebke, D., Mello, S.D.: QUEEN: QUantized efficient ENcoding for streaming free-viewpoint videos. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=7xhwE7VH4S
11. Guo, J., Liu, J., Chen, J., Mao, S., Hu, C., Jiang, P., Yu, J., Xu, J., Liu, Q., Xu, L., Chen, Z., Guo, C.: Auto-connect: Connectivity-preserving rigformer with direct preference optimization. arXiv preprint arXiv:2506.11430 (2025)
12. He, Y., Gu, X., Ye, X., Xu, C., Zhao, Z., Dong, Y., Yuan, W., Dong, Z., Bo, L.: Lam: Large avatar model for one-shot animatable gaussian head. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–13 (2025)
13. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
14. Hu, Y., Liu, Z., Shao, J., Lin, Z., Zhang, J.: Eva-gaussian: 3d gaussian-based real-time human novel view synthesis under diverse camera settings. arXiv preprint arXiv:2410.01425 (2024)
15. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. arXiv preprint arXiv:2505.23716 (2025)
16. Jiang, Y., Song, W., Li, S., Hao, A.: Hfhuman: High-fidelity human reconstruction from single image with multi-modality fusion. IEEE Transactions on Visualization and Computer Graphics **32**(2), 2152–2164 (2026). https://doi.org/10.1109/TVCG.2025.3623198
17. Jiang, Y., Shen, Z., Hong, Y., Guo, C., Wu, Y., Zhang, Y., Yu, J., Xu, L.: Robust dual gaussian splatting for immersive human-centric volumetric videos. ACM Transactions on Graphics (TOG) **43**(6), 1–15 (2024)
18. Jin, Y., Peng, S., Wang, X., Xie, T., Xu, Z., Yang, Y., Shen, Y., Bao, H., Zhou, X.: Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In: International Conference on Computer Vision (ICCV) (2025)

19. Lee, J., Won, C., Jung, H., Bae, I., Jeon, H.G.: Fully explicit dynamic guassian splatting. In: Proceedings of the Neural Information Processing Systems (2024)
20. Li, P., He, Y., Hu, Y., Dong, Y., Yuan, W., Liu, Y., Zhu, S., Cheng, G., Dong, Z., Guo, Y.: Panolam: Large avatar model for gaussian full-head synthesis from one-shot unposed image. arXiv preprint arXiv:2509.07552 (2025)
21. Li, P., He, Y., Hu, Y., Dong, Y., Yuan, W., Liu, Y., Zhu, S., Cheng, G., Dong, Z., Guo, Y.: Panolam: Large avatar model for gaussian full-head synthesis from one-shot unposed image. arXiv preprint arXiv:2509.07552 (2025)
22. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Li, M., Chi, X., Xia, S., Xue, W., et al.: Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. arXiv preprint arXiv:2409.10141 (2024)
23. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813
24. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8508–8520 (June 2024)
25. Lin, C., Lin, Y., Pan, P., Yu, Y., Yan, H., Fragkiadaki, K., Mu, Y.: Movies: Motion-aware 4d dynamic view synthesis in one second. arXiv preprint arXiv:2507.10065 (2025)
26. Lin, C.H., Lv, Z., Wu, S., Xu, Z., Nguyen-Phuoc, T., Tseng, H.Y., Straub, J., Khan, N., Xiao, L., Yang, M.H., Ren, Y., Newcombe, R., Dong, Z., Li, Z.: Dgs-lrm: Real-time deformable 3d gaussian reconstruction from monocular videos. arXiv preprint arXiv:2506.09997 (2025)
27. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21136–21145 (2024)
28. Liu, I., Xu, Z., Yifan, W., Tan, H., Xu, Z., Wang, X., Su, H., Shi, Z.: Riganything: Template-free autoregressive rigging for diverse 3d assets. arXiv preprint arXiv:2502.09615 (2025)
29. Liu, T., Wang, G., Hu, S., Shen, L., Ye, X., Zang, Y., Cao, Z., Li, W., Liu, Z.: Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In: European Conference on Computer Vision. pp. 37–53. Springer (2025)
30. Liu, Z., Hu, Y., Zhang, X., Song, R., Shao, J., Lin, Z., Zhang, J.: Dynamics-aware gaussian splatting streaming towards fast on-the-fly 4d reconstruction (2025), https://arxiv.org/abs/2411.14847
31. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 3DV (2024)
32. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
33. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024)
34. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., Bo, L.: Lhm: Large animatable human reconstruction model from a single image in seconds. In: arXiv preprint arXiv:2503.10625 (2025)

35. Qiu, L., Zhu, S., Zuo, Q., Gu, X., Dong, Y., Zhang, J., Xu, C., Li, Z., Yuan, W., Bo, L., et al.: Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In: CVPR (2025)
36. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
37. Ren, J., Xie, K., Mirzaei, A., Liang, H., Zeng, X., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H.: L4gm: Large 4d gaussian reconstruction model. In: Proceedings of Neural Information Processing Systems(NeurIPS) (Dec 2024)
38. Shen, Q., Wu, Z., Yi, X., Zhou, P., Zhang, H., Yan, S., Wang, X.: Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. arXiv preprint arXiv:2403.18795 (2024)
39. Sun, J., Jiao, H., Li, G., Zhang, Z., Zhao, L., Xing, W.: 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20675–20685 (June 2024)
40. Sun, Y.T., Huang, Y.H., Ma, L., Lyu, X., Cao, Y.P., Qi, X.: Splatter a video: Video gaussian representation for versatile processing. arXiv preprint arXiv:2406.13870 (2024)
41. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: LGM: large multi-view gaussian model for high-resolution 3d content creation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 15062, pp. 1–18. Springer (2024). https://doi.org/10.1007/978-3-031-73235-5_1, https://doi.org/10.1007/978-3-031-73235-5_1
42. Tu, H., Liao, Z., Zhou, B., Zheng, S., Zhou, X., Zhang, L., Wang, Q., Liu, Y.: Gbc-splat: Generalizable gaussian-based clothed human digitalization under sparse rgb cameras. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26377–26387 (2025)
43. Tu, H., Shao, R., Dong, X., Zheng, S., Zhang, H., Chen, L., Wang, M., Li, W., Ma, S., Zhang, S., Zhou, B., Liu, Y.: Tele-aloha: A telepresence system with low-budget and high-authenticity using sparse rgb cameras. In: ACM SIGGRAPH 2024 Conference Papers. SIGGRAPH '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3641519.3657491, https://doi.org/10.1145/3641519.3657491
44. Volumetric Video Challenge: (2025), https://www.4dv.ai/research/sig-asia2025-volumetric-video-challenges#dataset
45. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
46. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details (2025), https://arxiv.org/abs/2507.02546
47. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
48. Wang, Y., Yang, P., Xu, Z., Sun, J., Zhang, Z., Chen, Y., Bao, H., Peng, S., Zhou, X.: Freetimegs: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction. In: CVPR (2025), https://zju3dv.github.io/freetimegs

49. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Scalable permutation-equivariant visual geometry learning (2025), https://arxiv.org/abs/2507.13347
50. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. arXiv preprint arXiv:2405.14793 (2024)
51. Wang, Y., Huang, T., Chen, H., Lee, G.H.: Freesplat: Generalizable 3d gaussian splatting towards free-view synthesis of indoor scenes. arXiv preprint arXiv:2405.17958 (2024)
52. Wang, Z., Tan, J., Khurana, T., Peri, N., Ramanan, D.: Monofusion: Sparse-view 4d reconstruction via monocular fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8252–8263@ (October 2025)
53. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20310–20320 (June 2024)
54. Wu, Z., Yan, Q., Yi, X., Wang, L., Liao, R.: Streamsplat: Towards online dynamic 3d reconstruction from uncalibrated video streams. arXiv preprint arXiv:2506.08862 (2025)
55. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: CVPR (2025)
56. Xu, X., Wu, J.H., Li, Q.: What drives consumer shopping behavior in live streaming commerce? Journal of electronic commerce research **21**(3), 144–167 (2020)
57. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., Lv, Z.: 4dgt: Learning a 4d gaussian transformer using real-world monocular videos (2025)
58. Yan, J., Peng, R., Wang, Z., Tang, L., Yang, J., Liang, J., Wu, J., Wang, R.: Instant gaussian stream: Fast and generalizable streaming of dynamic scene reconstruction via gaussian splatting (2025), https://arxiv.org/abs/2503.16979
59. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. arXiv preprint arXiv:2309.13101 (2023)
60. Ye, B., Liu, S., Xu, H., Xueting, L., Pollefeys, M., Yang, M.H., Songyou, P.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. arXiv preprint arXiv:2410.24207 (2024)
61. Yi, X., Wu, Z., Shen, Q., Xu, Q., Zhou, P., Lim, J.H., Yan, S., Wang, X., Zhang, H.: Mvgamba: Unify 3d content generation as state space sequence modeling. arXiv preprint arXiv:2406.06367 (2024)
62. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: Guava: Generalizable upper body 3d gaussian avatar. arXiv preprint arXiv:2505.03351 (2025)
63. Zhang, H., Xu, H., Feng, C., Jampani, V., Ahuja, N.: Physrig: Differentiable physics-based skinning and rigging framework for realistic articulated object modeling. arXiv preprint arXiv:2506.20936 (2025)
64. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views (2025), https://arxiv.org/abs/2502.12138
65. Zhang, S., Ge, Y., Tian, J., Xu, G., Chen, H., Lv, C., Shen, C.: Pomato: Marrying pointmap matching with temporal motion for dynamic 3d reconstruction. arXiv preprint arXiv:2504.05692 (2025)
66. Zhang, X., Liu, Z., Zhang, Y., Ge, X., He, D., Xu, T., Wang, Y., Lin, Z., Yan, S., Zhang, J.: Mega: Memory-efficient 4d gaussian splatting for dynamic scenes. arXiv preprint arXiv:2410.13613 (2024)

67. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)