Complet4R: Geometric Complete 4D Reconstruction

Weibang Wang[1,*] Kenan Li[1,*] Zhuoguang Chen[1,2,*] Yijun Yuan[1,†] Hang Zhao[1,2,3,†]

[1]IIIS, Tsinghua University [2]Shanghai Artificial Intelligence Laboratory [3]Shanghai Qi Zhi Institute

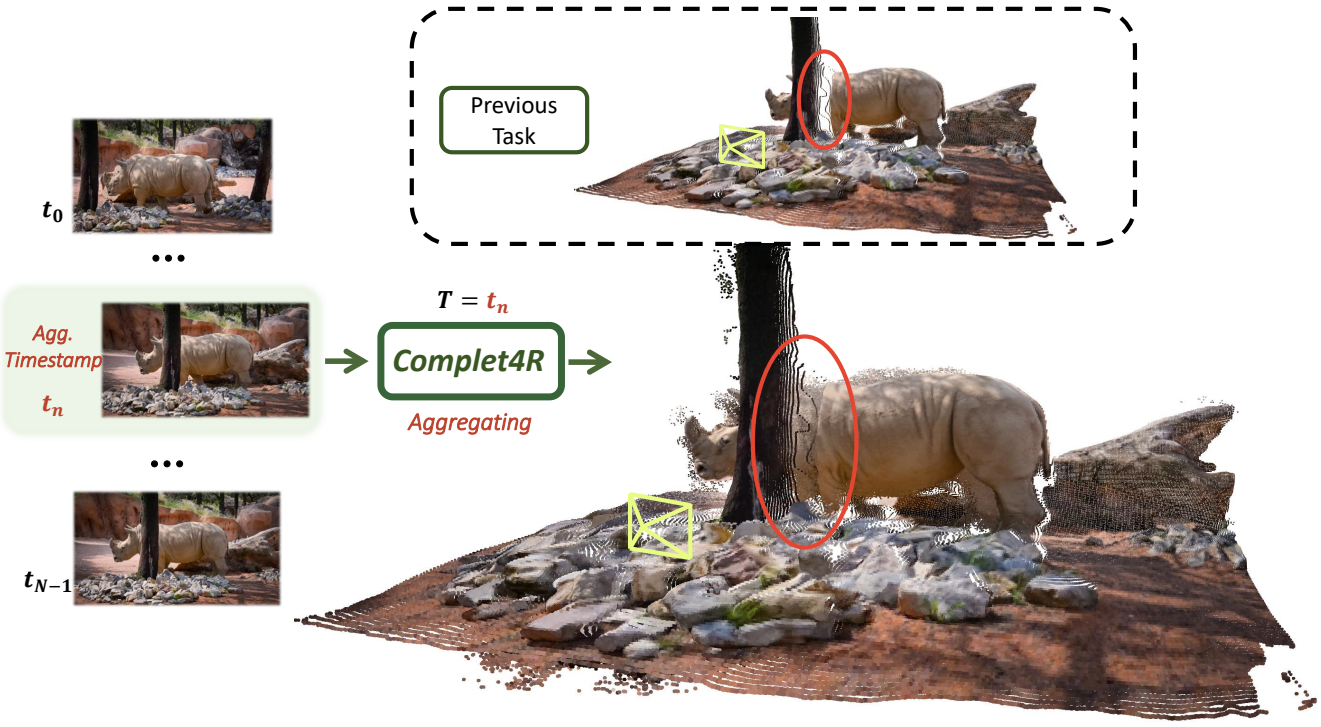

Figure 1. **Complete and Consistent 4D Reconstruction.** Our model, Complet4R, aggregates 3D point maps from all frames into a specific timestamp, forming a complete geometric representation that recovers occluded regions visible from other views. By alternating the aggregated timestamp, our method achieves complete and consistent 4D reconstruction, producing temporally coherent and geometrically complete representations directly from sequential video input.

Abstract

We introduce Complet4R, a novel end-to-end framework for Geometric Complete 4D Reconstruction, which aims to recover temporally coherent and geometrically complete reconstruction for dynamic scenes. Our method formalizes the task of Geometric Complete 4D Reconstruction as a unified framework of reconstruction and completion, by directly accumulating full contexts onto each frame. Unlike previous approaches that rely on pairwise reconstruction or local motion estimation, Complet4R utilizes a decoder-only transformer to operate all context globally directly from sequential video input, reconstructing a complete geometry for every single timestamp, including occluded regions visible in other frames. Our method demonstrates the state-of-the-art performance on our proposed benchmark for Geometric Complete 4D Reconstruction and the 3D Point Tracking task. Code will be released to support future research.

1. Introduction

Modeling the 3D world has long been a central problem [30] in both computer vision and robotics. Classical structure-from-motion (SfM) and simultaneous localization and mapping (SLAM) approaches reconstruct static environments by combining point matching, triangulation, and bundle adjustment [1, 32, 41]. While highly effective for rigid scenes, these formulations break down in the presence

* Equal contribution. †Equal advising.

of dynamics: moving objects violate the rigidity assumption, often appearing distorted or being entirely removed from the reconstruction. Traditional methods tend to treat dynamic regions as noise to maintain spatial consistency, which inevitably degrades reconstruction quality when motion is prevalent. However, dynamics are far more than noise. They encode rich spatial and temporal correlations that are vital for understanding the real world. By accurately modeling the dynamics, 3D perception can be extended into the spatiotemporal (4D) domain, where consistency must hold not only across space but also over time. Such 4D representations provide a foundation for higher-level reasoning about the physical world, enabling the development of world models and deeper insights into causality.

Early efforts to handle scene dynamics were largely built on 2D cues, such as optical flow [3] and long-range pixel correspondences [37]. Although these techniques yield dense trajectories in the image plane, they cannot resolve the fundamental ambiguities of projection: a single 3D point may map to different pixels under changing viewpoints, while distinct 3D points may collapse into one pixel through occlusion or perspective. Consequently, recovering accurate geometry or enforcing temporal consistency across views remains formidable.

Recent advances in 3D tracking and dynamic reconstruction have lifted motion analysis from the image plane to 3D space. By estimating dense point maps and establishing frame-to-frame correspondences [26, 48], these methods represent dynamic scenes as temporally coherent 3D videos [55]. To further tame non-rigid deformation, some approaches decompose the task into a set of rigid-motion sub-problems [10], gaining local stability and interpretability without sacrificing global consistency.

Despite the progress, the dominant paradigm remains pairwise: each step reasons over only two frames. This shortsighted matching makes errors accumulate through the sequence and offers no lever for global spatiotemporal regularization. Moreover, their intermediate representations, point maps or 3D flow fields, are still frame-centric, limiting their capacity to represent the continuous evolution of real-world dynamics. Consequently, they can hardly distill the full sequence into a single, globally consistent 4D structure that fuses shape and motion across time directly.

Unlike prior approaches that either ignore temporal consistency [55] or only perform frame-level tracking [10], our framework offers a new concept to explore spatiotemporal coherence by completing 4D across the whole stream. Which is, for every single frame, the 3D geometry is completed from all observations, so spatiotemporal coherence is baked in from the outset rather than retrofitted.

Our method provides a foundational representation for reasoning about real-world dynamics and serves as a step toward constructing physically grounded models. Our main contributions are:

1. We introduce a new problem, complete consistent 4D scene reconstruction, which aims to recover temporally coherent and geometrically complete 4D representations from dynamic scenes.
2. To tackle this challenging problem, we propose Complet4R, a novel framework that effectively integrates motion information into a consistent, complete 4D geometric representation.
3. Extensive experiments demonstrate Complet4R achieves state-of-the-art or competitive performance on both the proposed Geometric Complete 4D Reconstruction and its sibling 3D Point Tracking task benchmarks.

2. Related Work

2.1. Camera Estimation and Scene Reconstruction

Joint prediction of camera poses and scene geometry has long been a classical problem in computer vision. Traditional Structure-from-Motion (SfM) methods like [40], VGGSfM [44] and Simultaneous Localization and Mapping (SLAM) methods like MonoSLAM [7], ORB-SLAM [32] require extensive parameter tuning and incur high computational costs, which limit practical deployment.

Recently, a variety of learning-based approaches for monocular and video depth estimation have brought new opportunities. For example, DUSt3R [48] predicts camera poses and scene geometry from image pairs, but its computationally intensive post-processing limits scalability to large scenes. Building on this trend, several notable works, such as VGGT [45], enable high-precision prediction of camera poses, depth maps, and 2D tracking through a single forward pass of multi-view inputs, representing a new paradigm of data-driven 3D scene understanding.

However, these methods target static scenes, while dynamic scene reconstruction remains challenging. Existing approaches, such as R-CVD [22], CasualSAM [56], and MegaSAM [28], jointly optimize camera parameters and dense depth maps with depth priors. A notable follow-up to DUSt3R [48], St4RTrack [10] uses a dual-branch architecture for dynamic reconstruction and tracking, but its pairwise input still limits long-range predictions.

Therefore, to introduce a globally consistent model that addresses above limitations is of great interest to the field.

2.2. 4D Reconstruction

The 4D reconstruction problem has been extensively studied in non-rigid reconstruction. Early approaches relied on RGB-D sensors [4, 9, 16, 33, 58] or strong hand-crafted priors [5, 11, 24, 36, 38], while later works leveraged monocular depth priors [29, 56] for better generalization. Recent advances in neural rendering [31] and Gaussian Splatting [21] have further improved dynamic scene reconstruc-

tion, yet most methods [27, 42, 53] focus on novel view synthesis and photorealistic rendering rather than recovering physically meaningful geometry, thus neglecting geometric 4D consistency, an increasingly critical property for spatial visual intelligence. Many approaches also depend on auxiliary priors or assumptions [46], including known camera parameters, motion cues [6], rigidity constraints, and even semantics from large language models [2, 51, 54]. More recent efforts jointly infer camera poses, persistent geometry, and 3D trajectories from monocular videos [25, 35, 47], but they still require per-scene optimization and off-the-shelf priors. Although [10, 12, 43] reveal the potential of complete dynamic reconstruction from multiple observations, their pairwise correspondence backbone limits performance to pairwise tracking and frame-wise reconstruction.

In contrast, we propose a feed-forward framework that achieves 4D complete reconstruction. It reconstructs the scene and objects not only from the current frame, but also completes them from all observed parts in monocular videos, enabling geometrically consistent 4D representations and moving toward a complete physically grounded understanding of dynamic scenes.

2.3. 3D Point Tracking

The tracking task was first introduced in Particle Video [39]. Early optical flow and scene flow methods [13, 15], which typically estimate motion only between adjacent frames, struggle to handle long video sequences. CoTracker [20] was the first to leverage a transformer architecture with joint attention to enable tracking through occlusions. Following this, BootsTAPIR [8] and CoTracker3 [19] explored the use of unlabeled data to further enhance performance. However, many existing long-sequence tracking methods [19, 20, 50] operate only in the camera coordinate system, making it difficult to achieve globally consistent tracks.

Recently, increasing attention has been given to 3D tracking. For example, SpatialTrackerV2 [50] achieves pixel-level tracking by jointly estimating camera poses and monocular depth. Another notable method is MV-Tracker [52], which extracts multi-level scene features to achieve high-precision 3D tracking across multiple views. However, its reliance on pre-calibrated camera parameters and ground-truth depth maps significantly restricts its applicability in real-world scenarios. While our model can also support the 3D tracking task without those limitations.

3. Geometric Complete 4D Reconstruction

In this section, we introduce a novel task named **Geometric Complete 4D Reconstruction**, which aims to reconstruct the complete geometry, including regions that are occluded in the current frame but visible in other frames. In essence, the objective is to *aggregate information across time* so that all observed points, from both dynamic objects and static backgrounds, contribute to the reconstruction of a specific target timestamp. The key challenge of this task lies in accurately estimating the motion of dynamic points between timestamps, even when they are occluded in the target frame. To address this challenge, we propose **Complet4R**, a decoder-only transformer framework that globally reasons over sequential video inputs to achieve geometrically consistent and complete 4D reconstruction.

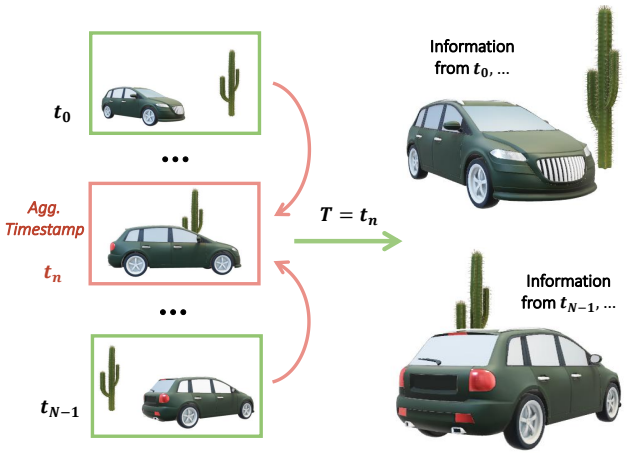

Figure 2. Geometric 4D complete reconstruction from observations. Given input frames, Complet4R aggregates contextual information across all timestamps. Consequently, at each timestamp T, the reconstructed scene incorporates the geometry from frame T along with complementary information from all other frames.

In the following subsections, we first define the task formulation and clarify its differences from prior related tasks. We then describe the architecture of Complet4R and its training strategy in detail.

3.1. Task Formulation

Revisiting Prior Task Definitions. Prior 4D reconstruction works, such as MonST3R [55], have made notable progress in modeling dynamic scenes from monocular videos. These methods typically regress the *visible* pointmap corresponding to each timestamp, focusing on reconstructing the geometry observable in the current frame. However, they overlook that different portions of dynamic objects or static regions may only become visible at other timestamps due to occlusions or motion. Consequently, these approaches provide *spatiotemporally incomplete* geometry representations.

Our Definition: Geometric Complete 4D Reconstruction. We propose a novel task named Geometric Complete 4D Reconstruction, as depicted in Fig. 2, which aims to reconstruct a *complete geometry* for every timestamp, covering both visible and occluded regions, by leveraging information across the entire temporal sequence. In essence, this task enables temporal aggregation of 3D geometry such that

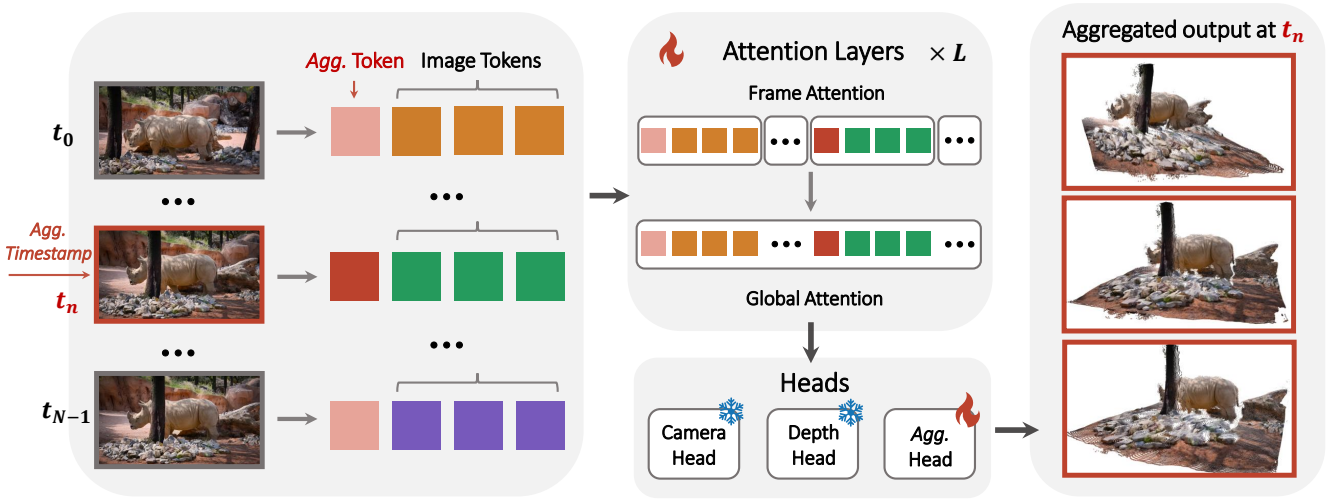

Figure 3. **Architecture Overview.** By concatenating special aggregation tokens, Complet4R identifies the specific timestamp for aggregation. The Aggregation head then outputs the positions of 3D points from other views at this timestamp, aggregating 3D point maps across frames to form a complete geometric representation.

all points observed at different timestamps are aligned and aggregated into a chosen target timestamp.

Formally, given a sequence of N temporally continuous RGB images observing a dynamic 3D scene,

$$\mathbf{I}_i \in \mathbb{R}^{3\times H\times W}, \quad i = 0, \ldots, N-1,$$

and a target (aggregate) timestamp $a \in \{0, \ldots, N-1\}$, our goal is to reconstruct the *complete* 3D representation of the scene at time step a.

For a given target timestamp a, the model aggregates geometric cues from all other frames $\{\mathbf{I}_i\}_{i\neq a}$, including both static backgrounds and moving objects, to infer a *complete* pointmap $\mathbf{P}^a$. Unlike previous methods that only recover visible geometry, this formulation explicitly reconstructs occluded parts of moving objects that are observed in other frames, thereby implicitly embedding motion causality reasoning across time.

Relation to 3D Tracking. Geometric Complete 4D Reconstruction inherently establishes dense temporal consistency across frames by aggregating geometry over time. Through this process, the model implicitly infers the motion of each 3D point. In other words, the temporal alignment used for geometric completion naturally yields consistent point trajectories, enabling 3D tracking as a byproduct of reconstruction. In practice, by iteratively treating each frame $\mathbf{I}_i$ as the target timestamp $\mathbf{I}_a$, the model reconstructs temporally coherent geometry at each step, and the accumulated temporal consistency across these reconstructions directly produces continuous 3D motion trajectories for all points.

3.2. Model Architecture

Complet4R is built upon a unified transformer architecture that jointly reasons over temporal sequences to reconstruct complete geometry at a chosen target timestamp, while simultaneously estimating per-frame depth and camera parameters. Formally, the model is defined as a mapping:

$$f\left((\mathbf{I}_i)_{i=0}^{N-1}, a\right) = (\mathbf{P}_i^a, \mathbf{g}_i, \mathbf{D}_i)_{i=0}^{N-1}, \tag{1}$$

where $\mathbf{P}_i^a \in \mathbb{R}^{H\times W\times 3}$ represents the aggregated 3D points from frame i toward the target timestamp a, $\mathbf{g}_i \in \mathbb{R}^9$ denotes the camera parameters (including intrinsics and extrinsics) [44], and $\mathbf{D}_i \in \mathbb{R}^{H\times W}$ is the predicted depth map.

Each input image $\mathbf{I}$ is divided into K patches and embedded into a set of visual tokens $\mathbf{t}^I \in \mathbb{R}^{K\times C}$ using DINOv2. All subsequent computations are based on these tokens.

Aggregation-aware Token Design. To enable geometric aggregation and completion, we introduce a novel design of aggregation tokens $\mathbf{t}^D$. Specifically, two sets of aggregation tokens are initialized: one set $\mathbf{t}_a^D$ for the target timestamp and another set $\mathbf{t}_{/a}^D$ shared among all other timestamps. These tokens are concatenated with each frame's visual tokens, allowing the model to explicitly identify the aggregation target. By varying the target timestamp a, Complet4R learns to gather geometric cues from other frames toward the a, thereby reconstructing the complete scene geometry.

Following VGGT, we include camera tokens $\mathbf{t}^{\mathbf{g}}$ and registration tokens $\mathbf{t}^R$, which encode camera poses and align features to a shared coordinate system, respectively. In particular, we initialize two sets of registration tokens: one $\mathbf{t}_1^R$ for the first frame and another $\mathbf{t}_{2:N}^R$ shared by all remaining frames, so that the model learns a unified representation

under the coordinate system of the first frame. The aggregation tokens $\mathbf{t}^D$ further capture frame-specific semantics, facilitating temporal alignment across time. We also experimented with additive fusion instead of concatenation, and detailed comparisons are discussed later.

Attention Mechanism. During both the frame-level and global attention stages of the transformer, aggregation tokens from the target frame and the remaining frames interact through self-attention to exchange temporal information. This process progressively aligns the features of all frames toward the target frame, integrating observations from both past and future frames. Consequently, the model reconstructs a holistic 3D representation of the scene and synthesizes complementary viewpoints that are otherwise unobservable from a single frame.

Prediction Heads. We adopt the original camera and depth heads from VGGT and introduce an additional *aggregation head* to enable temporally consistent 4D reconstruction. Both the depth and aggregation heads are implemented using DPT-style decoders. The camera head predicts camera parameters directly from the camera tokens, while the depth head estimates depth maps from patch tokens.

The aggregation head takes the features of each input frame and predicts the corresponding 3D scene representation aligned with the target frame's timestamp. By combining the outputs of the aggregation head across all frames, we obtain a complete and temporally coherent 3D point map of the scene from the viewpoint of the chosen target frame.

3.3. Training

Training Losses. We train our Complet4R model f using a multi-task loss:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{point}} + \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}}. \tag{2}$$

To address the difficulty of supervising dynamic and misaligned regions in 4D completion, we introduce a novel *Focal-Weighted Point Loss*. Given predicted points $\hat{\mathbf{P}}_i^a$ warped to a target timestamp a and ground truth $\mathbf{P}_i^a$, we follow VGGT and use a predicted uncertainty map $\hat{\Sigma}_{i,a}^P$ for aleatoric weighting.

To further emphasize hard samples, we adopt a focal-style point weight:

$$\mathbf{w}_i^a = |\beta \mathbf{e}_i^a|^\gamma, \qquad \mathbf{e}_i^a = \hat{\mathbf{P}}_i^a - \mathbf{P}_i^a,$$

which increases supervision on points with larger alignment errors. This improves robustness in highly dynamic regions.

The final loss for target frame a is:

$$\begin{aligned}\mathcal{L}_{\text{point}} = \sum_{i=1}^{N} \Big(&\|\hat{\Sigma}_{i,a}^P \odot \mathbf{w}_i^a \odot (\hat{\mathbf{P}}_i^a - \mathbf{P}_i^a)\| \\ &+ \|\hat{\Sigma}_{i,a}^P \odot (\nabla \hat{\mathbf{P}}_i^a - \nabla \mathbf{P}_i^a)\| - \alpha \log \hat{\Sigma}_{i,a}^P \Big),\end{aligned} \tag{3}$$

where $\odot$ denotes channel-broadcast multiplication.

This Focal-Weighted Point Loss is a key component of our framework, yielding more accurate 4D reconstruction in challenging regions.

Following VGGT, we employ the same loss functions to supervise both the camera parameters and depth estimation.

Training Setup. The model is initialized from VGGT, with both the camera and depth heads frozen during training. The Aggregation head is introduced in place of the original point head, inheriting its parameters.

Training is performed by minimizing the training loss (2) using the AdamW optimizer for 10 epochs, together with a cosine learning rate schedule featuring a peak learning rate of $1e-5$ and an 0.5-iteration warmup. Input frames, depth maps, and point maps are resized such that their longer side is at most 518 pixels, with width-to-height ratios randomly sampled between 0.5 and 3.4. We further apply data augmentations including random color jitter, gaussian blur, and grayscale conversion. Training is conducted on 8 A100 GPUs over 23 hours.

Training Data. We use three datasets for training: Point Odyssey [57], Dynamic Replica [18], and SAIL-VOS 3D [14]. They consist of dynamic scenes and provide depth maps, camera parameters, and 3D point trajectories. Their rich motion patterns, including people with storylines [14], make them suitable for learning pixel-level correspondences across space and time. Among them, SAIL-VOS 3D provides the most complete 3D labels, offering meshes at all timestamps, which we use to generate 3D trajectories. We represent each pixel using barycentric coordinates on a mesh tile, and enforce temporal consistency via vertex correspondences across timestamps. In addition, we split the original long sequences into shorter, temporally consistent clips by identifying large depth-shift changes, which correspond to abrupt camera cuts in the storytelling videos. The final processed SAIL-VOS 3D training dataset contains about 704 unique sequences, combined with 109 sequences in Point Odyssey and 483 sequences in Dynamic Replica, to serve as our whole training dataset.

4. Experiment

To assess the performance of Complet4R, we conduct experiments on three tasks: 4D Complete Reconstruction (Sec. 4.1), 3D Point Tracking (Sec. 4.2), followed by an ablation study (Sec. 4.3).

4.1. 4D Complete Reconstruction

We evaluate the 4D complete reconstruction performance of Complet4R by measuring Accuracy, Completion, and Normal Consistency on the Complet4R-benchmark, to provide a comprehensive assessment.

Methods	Acc.↓		Complet.↓		N.C.↑	
	Mean	Med.	Mean	Med.	Mean	Med.
St4RTrack-seq	0.92	0.71	3.10	0.17	0.48	0.47
St4RTrack-pairs	0.94	0.77	2.67	0.14	0.46	0.44
Complet4R (Ours)	**0.50**	**0.37**	**0.26**	**0.11**	**0.49**	**0.49**

Table 1. **4D Complete Reconstruction on SAIL-VOS 3D-test.** We report Accuracy (Acc.), Completion (Complet.), and Normal Consistency (N.C.) on all points; each metric is shown as Mean and Median. Best results are **bold**.

Datasets. To evaluate and compare the 4D complete reconstruction capability of models for the task, we select a subset of the SAIL-VOS 3D validation split as *SAIL-VOS 3D-test* for the benchmark, which contains 44 sequences with the same distribution as the original dataset, with the same ratio of indoor/outdoor conditions and similar spatial–temporal distribution of motion amplitude and patterns.

Baselines. To the best of our knowledge, this task represents a novel problem, and there are no existing methods that can be directly used for comparison. The most similar method is St4RTrack [10], which leverages pairwise correspondences for joint tracking and reconstruction. Due to the limitation of pairwise tracking, achieving full 4D reconstruction requires anchoring every frame and subsequently aggregating the tracking results for each timestamp. We adopt this method as our baseline and evaluate the two checkpoints provided for different training modes [10].

The 3D point tracking series [49, 50] focus on sparse-point tracking and are not designed for reconstruction. On our task, they cannot perform dense per-pixel tracking on the same hardware and lack supervision for currently occluded regions. Like St4RTrack, they also require post-aggregation for complete reconstruction. Due to these limitations, which are not fair for these tracking methods in the task, we do not use them as our baselines.

Evaluation Metrics. To evaluate 4D complete reconstruction performance, we follow the 3D reconstruction evaluation protocol of CUT3R [47]. Concretely, we evaluate the 3D reconstruction performance of each selected target frame, which aggregates points from the current and all other observations. For computational efficiency, the points are randomly downsampled.

Results. The evaluation results are summarized in Tab. 1. Our model consistently surpasses the baselines across all metrics, establishing a new state-of-the-art on the 4D complete reconstruction task. Specifically, we achieve nearly 50% relative improvements in Accuracy, and an order-of-magnitude improvement in the mean Completion metric. These results highlight the model's superior performance in recovering complete 3D scenes with consistent geometry.

4.2. 3D Point Tracking

As previously mentioned, Complet4R implicitly supports 3D point tracking. We therefore evaluate its performance on 3D point tracking benchmarks using Average Percent of Points within Delta (APD) and Endpoint Error (EPE), following the protocol of St4RTrack [10]. Since Complet4R is not specifically designed for 3D point tracking, we perform 4D complete reconstruction for each frame and aggregate corresponding points along the temporal axis to form trajectories for evaluation.

Datasets and Baselines. The evaluation is conducted on the *WorldTrack* dataset released by St4RTrack, constructed from Aerial Digital Twin (ADT) [34] and Panoptic Studio [17] via TAPVid-3D [23], and further augmented with PointOdyssey and DynamicReplica. The dataset provides ground-truth 3D trajectories in the world coordinate system with corresponding 2D projections, covering diverse scenarios such as minimal motion, dynamic objects, and large camera motion.

To benchmark the effectiveness of our method, we compare it against the baselines provided by *St4RTrack* [10], including *SpaTracker* [49], *MonST3R* [55], and [10]. For fair comparison, we align the camera-coordinate outputs of *SpaTracker* using Procrustes and RANSAC, and transform the predictions of *MonST3R* into world coordinates using its estimated camera poses, enabling 3D tracking evaluation in a unified world coordinate system, which is not natively supported by these methods.

Metric and Results. If query points are selected from any frame, Complet4R can predict their 3D coordinates in any other frame. For evaluation consistency, we choose the dynamic visible points in the first frame of each sequence that have complete tracking information as query points.

Next, the predicted 3D points are aligned with the ground truth using the global median. We compute the norms of both predicted points and ground truth points, and derive a scaling factor from their medians for alignment.

Finally, the aligned predictions are evaluated using APD and EPE. APD is computed as the average percentage of points within four 3D thresholds, $\delta_{3D} \in \{0.1\text{m}, 0.3\text{m}, 0.5\text{m}, 1.0\text{m}\}$, across all sequences, with the final score averaged over these thresholds. EPE is calculated as the L2 distance between predicted and ground-truth points, averaged over all frames in the sequences.

The evaluation results are shown in Tab. 2. Although our model was not specifically trained for point tracking, it significantly outperforms the state-of-the-art across all metrics on the point tracking benchmark. This demonstrates that our approach can not only reconstruct complete scenes at each timestamp, but also recover temporally consistent and continuous dynamic motions, as further illustrated in Fig. 5.

Category	Methods	PO		DR		ADT		PStudio	
		APD↑	EPE↓	APD↑	EPE↓	APD↑	EPE↓	APD↑	EPE↓
Combinational	SpaTracker+RANSAC-Procrustes	53.77	43.58	58.58	104.44	66.49	16.00	52.05	42.66
	SpaTracker+MonST3R	58.61	40.85	59.21	91.36	69.94	15.11	50.16	48.37
Feed-forward	MonST3R	39.36	64.52	51.86	53.13	67.92	15.78	51.32	45.68
	SpaTracker	51.20	46.95	58.65	108.28	67.65	16.28	62.59	30.94
	St4RTrack	68.72	29.70	68.13	29.61	75.34	12.12	69.67	26.37
	Complet4R (Ours)	**80.17**	**16.07**	**80.65**	**15.99**	**77.34**	**9.72**	**76.88**	**18.52**

Table 2. **World Coordinate 3D Point Tracking on Dynamic Points.** We report both APD and EPE for four datasets: PO (Point Odyssey), DR (Dynamic Replica), ADT (Aria Digital Twin), and PStudio (Panoptic Studio). Best results are **bold**.

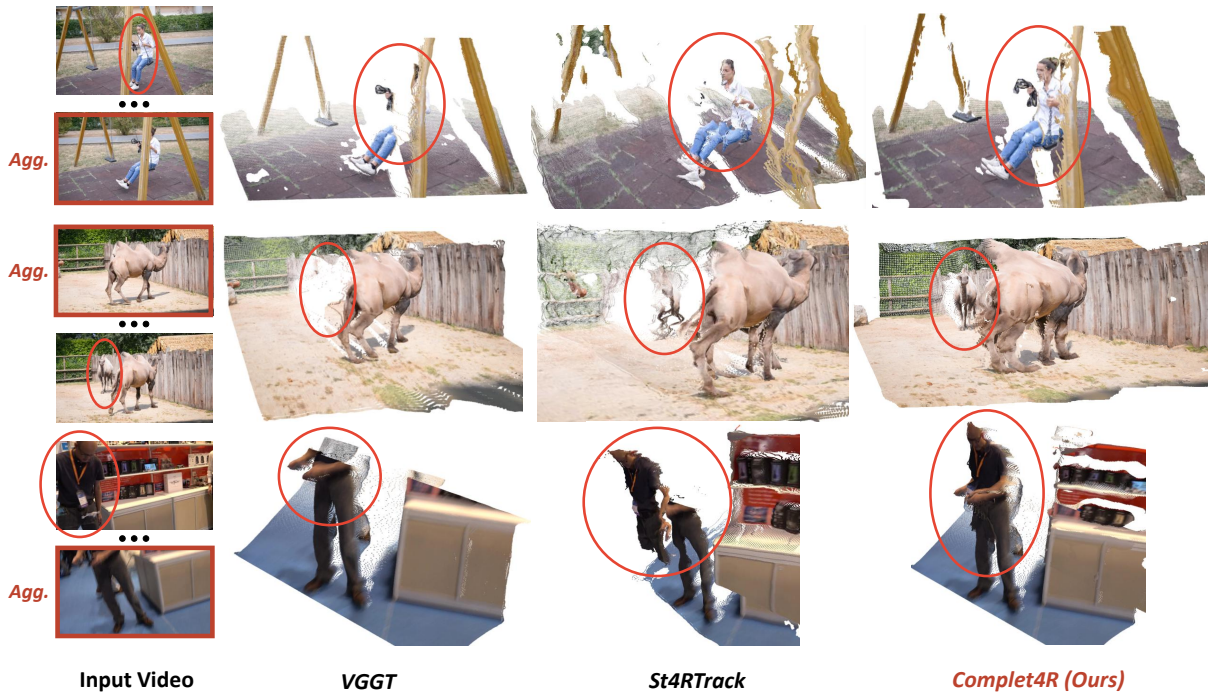

Figure 4. **Qualitative Results for 4D Complete Reconstruction.** The first column shows the video inputs, with red boxes indicating the target aggregation timestamp for each sequence (*Agg.*: aggregation). The subsequent columns present the outputs of different models. Our method successfully reconstructs the complete geometry at the target timestamp highlighted by the red ellipses, whereas other methods produce incomplete or geometrically inconsistent reconstructions.

Variants	Loss	Agg. Repre.	Agg. Token	Acc.↓		Complet.↓		N.C.↑	
				Mean	Med.	Mean	Med.	Mean	Med.
(1)	Dynamic	-	-	**0.50**	0.47	**0.26**	0.20	0.48	0.47
(2)	-	Offset	-	0.63	0.42	0.50	0.11	0.44	0.40
(3)	-	-	Add	0.57	0.38	0.32	0.12	**0.50**	**0.49**
Complet4R (Ours)	**Focal**	**Endpoint**	**Concatenate**	**0.50**	**0.37**	**0.26**	**0.11**	0.49	**0.49**

Table 3. **Ablation Study on 4D Complete Reconstruction on SAIL-VOS 3D-test.** Accuracy (Acc.), Completion (Complet.), and Normal Consistency (N.C.) are reported over all points; each metric includes Mean and Median. Agg. Repre. means the aggregation representation, Agg. Token means aggregation tokens. Best results are **bold**. The symbol "-" indicates the default setting same as Complet4R.

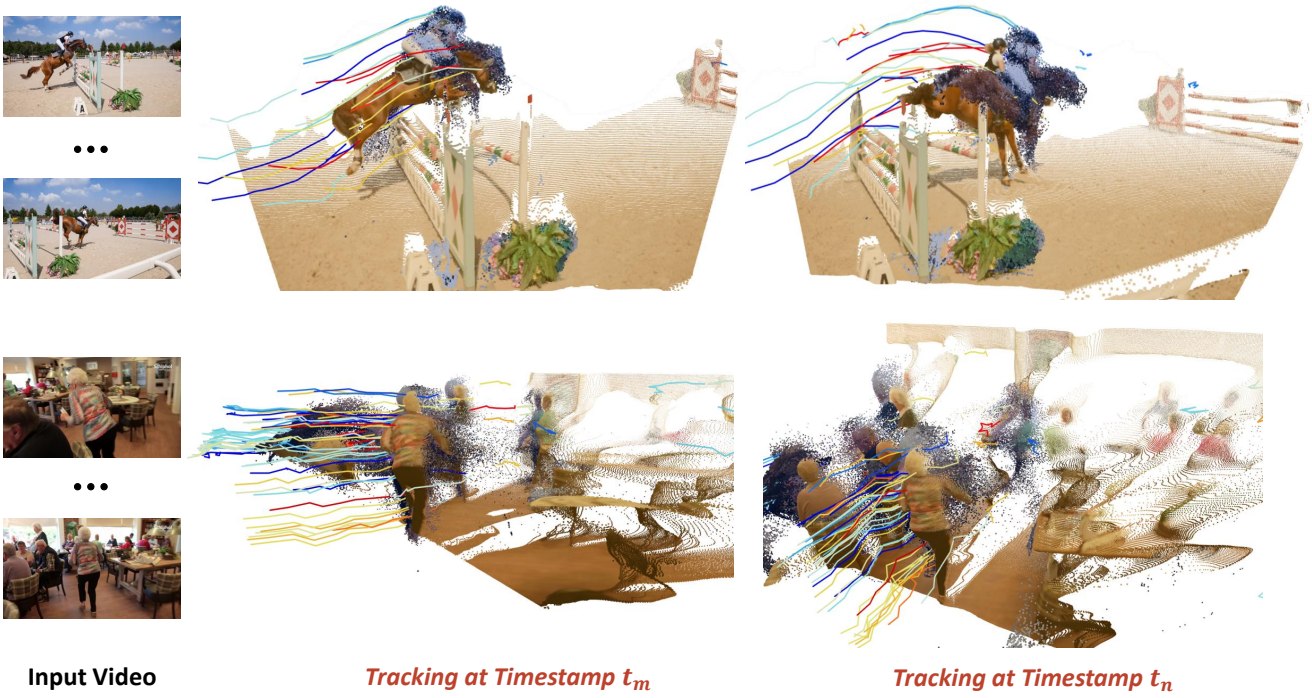

Figure 5. **Qualitative Results for 3D Dynamic Point Tracking.** The first column shows the input images; the second and third columns display the tracking trajectories produced by our method at successive time steps. The smooth trajectories demonstrate strong spatiotemporal geometric consistency.

4.3. Ablation Study

We conduct an extensive ablation on the SAIL-VOS 3D-test dataset to analyse three design axes that affect 4D complete reconstruction: the training loss, the aggregation representation, and the processing of aggregation tokens.

Complet4R employs a focal-weighted point loss to address the imbalance in point distribution, supervises absolute coordinates at the aggregation timestamp (Endpoint), and fuses aggregation tokens by concatenating them with image patch tokens.

For comparison, we evaluate variants that replace these components as follows; detailed configurations of each variant are provided in the Supplementary Material.

(1) Using a dynamic-weighted point loss instead of the focal-weighted point loss, which emphasizes dynamic points by multiplying their loss with a large factor.
(2) Supervising displacement relative to a reference frame (Offset) instead of absolute endpoint coordinates.
(3) Adding aggregation tokens directly to image patch tokens instead of concatenating them.

As shown in Table 3, our method consistently outperforms these variants across Accuracy, Completion, and Normal Consistency metrics, demonstrating the effectiveness of our design choices.

Specifically, the focal-weighted point loss dynamically adjusts supervision weights via an adaptive mechanism, encouraging the model to focus on challenging regions and thereby improving prediction performance. For the aggregation representation, endpoint outperforms offset, and for the aggregation token, concatenation outperforms addition, improving overall reconstruction quality.

5. Conclusion

In this paper, we introduce Complet4R, a novel end-to-end framework that constructs geometric complete 4D reconstruction for dynamic scenes. By jointly reasoning about geometry and motion across all frames, the network recovers geometrically complete and temporally coherent scene reconstruction, including areas occluded in any single view yet visible elsewhere. Unlike prior works that stitch pairwise reconstructions or tracks only local motion, Complet4R fuses long-range cues into one unified 4D representation, eliminating drift and ensuring cross-time consistency. Extensive experiments demonstrate that our method achieves state-of-the-art performance on both the newly proposed 4D complete reconstruction and the 3D point tracking benchmark. We believe that the proposed formulation and method will foster further research, paving the way toward more realistic and consistent world models.

6. Acknowledgements

This work is supported by the National Key R&D Program of China (2022ZD0161700) and Tsinghua University Initiative Scientific Research Program.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54 (10):105–112, 2011. 1

[2] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2024. 3

[3] Michael J Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *1993 (4th) International Conference on Computer Vision*, pages 231–236. IEEE, 1993. 2

[4] Aljaz Bozic, Michael Zollhofer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7002–7012, 2020. 2

[5] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, pages 690–696. IEEE, 2000. 2

[6] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *NeurIPS*, 2024. 3

[7] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007. 2

[8] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, Joao Carreira, et al. Bootstap: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision*, pages 3257–3274, 2024. 3

[9] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4): 1–13, 2016. 2

[10] Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J Black, Trevor Darrell, and Angjoo Kanazawa. St4rtrack: Simultaneous 4d reconstruction and tracking in the world. *arXiv preprint arXiv:2504.13152*, 2025. 2, 3, 6

[11] Katerina Fragkiadaki, Marta Salas, Pablo Arbelaez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. *advances in neural information processing systems*, 27, 2014. 2

[12] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. Dˆ 2ust3r: Enhancing 3d reconstruction with 4d pointmaps for dynamic scenes. *arXiv preprint arXiv:2504.06264*, 2025. 3

[13] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 3

[14] Yuan-Ting Hu, Jiahong Wang, Raymond A Yeh, and Alexander G Schwing. Sail-vos 3d: A synthetic dataset and baselines for object detection and 3d mesh reconstruction from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1418–1428, 2021. 5

[15] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7396–7405, 2020. 3

[16] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European conference on computer vision*, pages 362–379. Springer, 2016. 2

[17] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 6

[18] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13229–13239, 2023. 5

[19] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 3

[20] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pages 18–35. Springer, 2024. 3

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[22] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 2

[23] Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, Joao Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. Tapvid-3d: A benchmark for tracking any point in 3d. *Advances in Neural Information Processing Systems*, 37: 82149–82165, 2024. 6

[24] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *Proceedings of the IEEE international conference on computer vision*, pages 4649–4657, 2017. 2

[25] Jiahui Lei, Yijia Weng, Adam W Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6165–6177, 2025. 3

[26] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2

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

[28] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10486–10496, 2025. 2

[29] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. 2

[30] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 1

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[32] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1, 2

[33] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 2

[34] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 6

[35] LIU Qingming, Yuan Liu, Jiepeng Wang, Xianqiang Lyu, Peng Wang, Wenping Wang, and Junhui Hou. Modgs: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[36] Rene Ranftl, Vibhav Vineet, Qifeng Chen, and Vladlen Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4058–4066, 2016. 2

[37] Michael Rubinstein, Ce Liu, and William T Freeman. Towards longer long-range motion trajectories. 2012. 2

[38] Chris Russell, Rui Yu, and Lourdes Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *European conference on computer vision*, pages 583–598. Springer, 2014. 2

[39] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *International journal of computer vision*, 80(1):72–91, 2008. 3

[40] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[41] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 519–528. IEEE, 2006. 1

[42] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 3

[43] Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic point maps: A versatile representation for dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7295–7305, 2025. 3

[44] 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*, pages 21686–21697, 2024. 2, 4

[45] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2

[46] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 3

[47] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 3, 6

[48] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[49] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 6

[50] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy. *arXiv preprint arXiv:2507.12462*, 2025. 3, 6

[51] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024. 3

[52] Mengjie Xu, Yitao Zhu, Haotian Jiang, Jiaming Li, Zhenrong Shen, Sheng Wang, Haolin Huang, Xinyu Wang, Han Zhang, Qing Yang, et al. Mitracker: Multi-view integration for visual object tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27176–27185, 2025. 3

[53] Ling Yang, Kaixin Zhu, Juanxi Tian, Bohan Zeng, Mingbao Lin, Hongjuan Pei, Wentao Zhang, and Shuicheng Yan. Widerange4d: Enabling high-quality 4d reconstruction with wide-range movements and scenes. *arXiv preprint arXiv:2503.13435*, 2025. 3

[54] Bohan Zeng, Ling Yang, Siyu Li, Jiaming Liu, Zixiang Zhang, Juanxi Tian, Kaixin Zhu, Yongzhen Guo, Fu-Yun Wang, Minkai Xu, et al. Trans4d: Realistic geometry-aware transition for compositional text-to-4d synthesis. *arXiv preprint arXiv:2410.07155*, 2024. 3

[55] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 2, 3, 6

[56] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 2

[57] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 5

[58] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33(4): 1–12, 2014. 2