Director: Instance-aware Gaussian Splatting for Dynamic Scene Modeling and Understanding

Yuheng Jiang[1*], Yiwen Cai[1*], Zihao Wang[2], Yize Wu[1], Sicheng Li[2], Zhuo Su[2], Shaohui Jiao[2], and Lan Xu[1]

[1] ShanghaiTech University, Shanghai, China
[2] ByteDance, Shanghai, China
nowheretrix123@gmail.com

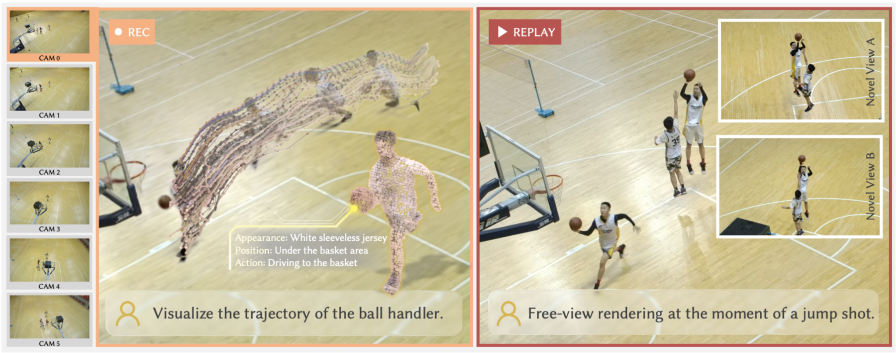

Fig. 1: We introduce **Director**, an instance-aware spatio-temporal Gaussian representation that enables robust human performance tracking, high-fidelity rendering, and instance-level understanding for open-vocabulary queries.

Abstract. Volumetric video seeks to model dynamic scenes as temporally coherent 4D representations. While recent Gaussian-based approaches achieve impressive rendering fidelity, they primarily emphasize appearance but are largely agnostic to instance-level structure, limiting stable tracking and semantic reasoning in highly dynamic scenarios. In this paper, we present Director, a unified spatio-temporal Gaussian representation that jointly models human performance, high-fidelity rendering, and instance-level semantics. Our key insight is that embedding instance-consistent semantics naturally complements 4D modeling, enabling more accurate scene decomposition while supporting robust dynamic scene understanding. To this end, we leverage temporally aligned instance masks and sentence embeddings derived from Multimodal Large Language Models to supervise the learnable semantic features of each Gaussian via two MLP decoders, enabling language-aligned 4D representations and enforcing identity consistency over time.
To enhance temporal stability, we bridge 2D optical flow with 4D Gaussians and finetune their motions, yielding reliable initialization and reducing drift. For the training, we further introduce a geometry-aware

[*] Equal Contribution.

SDF constraints, along with regularization terms that enforces surface continuity, enhancing temporal coherence in dynamic foreground modeling. Experiments demonstrate that Director achieves temporally coherent 4D reconstructions while simultaneously enabling instance segmentation and open-vocabulary querying.

Keywords: Gaussian Splatting · 4D Segmentation · Understanding

1 Introduction

We see what moves. We shape what persists. We understand what matters. Volumetric video aspires to capture this very progression, transforming dynamic visual observations into structured 4D representations, laying the foundation for scene understanding.

Yet, existing solutions [4,11,16,60] largely optimize dynamic scenes for photometric consistency, evolving from mesh-based models to NeRF [42] and Gaussian Splatting [28]. Despite achieving remarkable rendering fidelity, these approaches remain largely agnostic to the underlying scene structure, treating scenes as collection of radiance primitives without explicit instance-consistent constraints. Consequently, in highly dynamic scenarios, the optimization process often suffers from identity drift, foreground–background entanglement, unstable semantic association over time. A few recent works [19, 36] begin to move toward deeper structural understanding by introducing instance-level segmentation. Another line of research [29,34] explores natural language as an interface for scene querying and interaction. However, these approaches still adhere to a "reconstruction-then-understanding" paradigm, treating geometric modeling and semantics as loosely coupled components rather than mutually reinforcing processes. Moreover, they typically handle only carefully curated slow-motion scenarios and remain fragile when confronted with complex or fast dynamic motions.

In this paper, we argue that dynamic reconstruction and understanding should not follow the reconstruction-then-understanding pipeline, but instead be jointly optimized through instance-consistent primitive-level modeling. Instance-level segmentation and semantic cues enable more precise dynamic–static or inter-instance decomposition, improving tracking and appearance modeling. Conversely, long-term tracking and coherent 4D modeling provide temporally consistent context that enhances segmentation quality and forms a reliable foundation for open-vocabulary queries over dynamic scenes.

To "shape what persists and understand what matters", we introduce a unified and instance-aware spatio-temporal Gaussian representation for highly dynamic scenes (see Fig. 1). Our core idea is to leverage a set of temporally consistent 4D Gaussian primitives equipped with learnable semantic feature attributes to represent the scene. Guided by SAM3-derived instance masks and MLLM-derived semantic embedding, the Gaussian primitives are optimized to achieve accurate tracking, high-fidelity rendering, and detailed instance-level segmentation.

Our method, Director, first establishes reliable instance correspondences across multiple camera views. Leveraging temporally consistent SAM3 [3] masks, we

resolve cross-view ambiguities to obtain spatially and temporally consistent instance masks. To "understand what matters" beyond geometric partitioning, we generate time-varying, instance-wise captions from Multimodal Large Language Models (MLLMs) and encode them into sentence embeddings, which serve as language- and instance-aligned supervision for the training of the dynamic semantic field.

With the instance masks and semantic embedding as guidance, we explicitly decompose the scene into a static background and a dynamic foreground to enhance the consistency. For the static background, we aggregate observations across the entire temporal sequence and optimize a temporally consistent Gaussian layer. This provides a global reference for the scene, and facilitates accurate separation of dynamic foreground elements. For the dynamic foreground, we equip each Gaussian with a learnable semantic feature vector and use two MLP decoders to predict instance identity and language-aligned semantics. To enforce spatial coherence, we apply a KL-divergence regularization that encourages neighboring Gaussians within the same instance to maintain consistent semantic features.

To handle fast and complex motions in highly dynamic scenes, we explicitly warp the Gaussians from the previous frame using optical flow, providing reliable position initialization. During training, we first fine-tune only the Gaussian positions and rotations with an as-rigid-as-possible regularizer, enforcing local geometric consistency and correcting misalignments. We then leverage geometry-aware SDF constraints and regularization terms to jointly optimize Gaussian motion, appearance, and semantic features, maintaining instance-level consistency and temporal coherence. After training, our method identifies the target instance via identity query, renders its instance-level feature map, and retrieves corresponding relevant frame segments with segment queries, effectively filtering out interference from other dynamic objects.

To summarize, our main contributions include:

- A primitive-level instance-consistent 4D Gaussian formulation that tightly couples dynamic reconstruction and semantic modeling within a unified representation.
- Language- and instance-aligned semantic features embedded into dynamic Gaussians, where instance-level mask supervision and language priors jointly enforce structural constraints for temporally coherent 4D modeling, enabling detailed instance segmentation and open-vocabulary querying.
- A two-layer Gaussian representation that explicitly separates static and dynamic components and incorporates optical-flow-guided optimization for robust human performance tracking and high-fidelity rendering.

2 Related Work

Non-Rigid Tracking. Recent research on human performance non-rigid tracking has been extensively explored for various applications [12–14, 21, 25, 31, 35,

51, 59, 61, 70, 74, 76, 84, 86]. DynamicFusion [43] pioneered this direction by mapping dynamic frames into a canonical space via a dense volumetric warp field. VolumeDeform [17] enhanced robustness via sparse color feature matching. Fusion4D [6] scaled to multi-view capture for large motions and topological changes. KillingFusion and SobolevFusion [54, 55] replaced explicit correspondence estimation with differential regularization on signed distance fields, and DoubleFusion [82] embedded a parametric body model into a dual-layer representation for better occlusion handling. Later, Motion2Fusion [5] and UnstructureFusion [75] accelerated capture through learned correspondences in monocular and unstructured multi-camera settings, while RobustFusion [57,58] addressed human-object interactions via semantics-aware scene decomposition. DDC [13] further shifted toward weakly supervised learning, achieving photorealistic real-time reconstruction with fine-grained clothing details. Dynamic 3D Gaussians [40] allowed Gaussians to move and rotate over time to represent scene motions. Notably, recent works [8, 26, 38, 67, 69] have made significant progress in point tracking from casual monocular videos. Building on this line of work, our method incorporates multi-view 2D semantic priors and optical flow to maintain robust tracking in highly dynamic scenes with complex interactions and challenging occlusions.

Dynamic Scene Representation. Dynamic scene representation has undergone a fundamental paradigm shift over the past decade. Early mesh-based methods [4–6, 81] reconstructed textured dynamic meshes from dense camera rigs, but struggled to capture view-dependent appearance. Neural Radiance Fields (NeRF) [42] substantially elevated rendering fidelity through implicit volumetric representations, with dynamic extensions following either canonical-space deformation [33, 41, 45, 46, 86] or spatio-temporal tensor factorization [2, 10, 16, 52, 56]. Despite their quality gains, NeRF-based methods remain bottlenecked by costly ray-marching, limiting practical efficiency. 3D Gaussian Splatting (3DGS) [28] has since enabled photorealistic quality with fast training and real-time rendering, spurring three main directions for dynamic scenes: (1) per-Gaussian motion tracking across frames [20, 23, 24, 40, 50, 53, 60, 68, 87], (2) canonical Gaussian representations with deformation fields [22, 32, 44, 48, 72, 78], and (3) native 4D spacetime Gaussian representations [7, 37, 77], which parameterize Gaussians directly in 4D for real-time rendering via anisotropic temporal slicing. Our work follows the per-Gaussian motion tracking paradigm, augmenting it with semantic knowledge distilled from imperfect 2D semantic masks to achieve robust, high-quality dynamic scene representation.

Instance-level Scene Understanding and Editing. Bridging scene representations with 2D vision–language foundation models [1, 15, 66] through language fields has emerged as a dominant paradigm for open-vocabulary 3D scene understanding. Early works [27, 30, 47, 62, 63, 80, 83] lift 2D features into static 3D scenes. LERF [29] distills CLIP features into NeRF to enable language-driven queries. With the advent of 3D Gaussian Splatting (3DGS) [28], LangSplat [49] integrates 3DGS with SAM-derived hierarchical semantic maps for more precise language field modeling. Recent works [19] extend this paradigm to dynamic settings. 4DLangSplat [34] supervises per-object trajectories using temporally

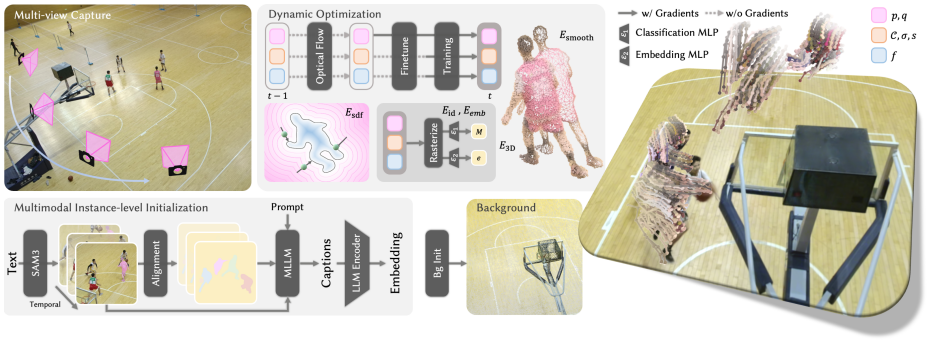

Fig. 2: Overview of Director. Using temporally consistent SAM3 masks and sentence embeddings, our method decomposes the scene into static background and dynamic foreground, learning language- and instance-aligned features for robust tracking, high-quality rendering, and accurate instance segmentation.

consistent captions and models state transitions via deformation networks. 4-LEGS [9] embeds spatio-temporal video features into 4D Gaussians to support text-based event localization across space and time. 4DLangVGGT [73] builds upon VGGT [64], a feed-forward visual geometry foundation model, and jointly learns spatio-temporal geometry and language alignment, enabling efficient open-vocabulary 4D understanding without per-scene optimization. Inspired by these advances, our work leverages long-term tracking and coherent 4D modeling to enhance segmentation quality and establish a reliable foundation for open-vocabulary queries over dynamic scenes.

3 Method

Our method is built upon a key design principle: instance-level consistency should be enforced directly at the Gaussian primitive level. To achieve this, we embed instance identities and language-aligned sentences embeddings into each Gaussian's semantic feature, and jointly optimize them with motion and appearance. The overall methodology is illustrated in Fig. 2.

Our approach first predicts temporally consistent masks and generate instance captions with MLLMs, using both as supervision during training to enforce identity consistency and semantic alignment (Sec. 3.1). We then explicitly disentangle foreground and background components. For the static background, we aggregate observations across the entire temporal sequence and jointly optimize a temporally consistent Gaussian representation $\mathcal{G}_{bg}$ to ensure stable spatio-temporal reconstruction. For the foreground, we equip each Gaussian with a learnable semantic feature and use two MLP heads to predict instance identity and language-aligned embedding (Sec. 3.2). Guided by optical flow, we then learn a unified dynamic representation $\mathcal{G}_{fg}$ that jointly models motion (position p, rotation q), appearance (third-order spherical harmonic $\mathcal{C}$, opacity σ, and scaling

s), and semantic feature f attributes in an end-to-end manner. This unified 4D semantic formulation, optimized with SDF constraints and regularization terms, ensures temporal coherence and photorealistic rendering, while laying a solid foundation for instance-aware editing and open-vocabulary querying (Sec. 3.3).

3.1 Multimodal Instance-level Initialization

Existing 4D representation methods [20,23,24] heavily rely on background matting [39], which is unreliable in sports scenarios due to dynamic lighting and reflections. To mitigate this issue, we employ SAM3 [3] and use text prompts such as "basketball player" and "basketball" to extract instances masks for basketball scenes. Leveraging the memory bank mechanism of SAM3, we obtain temporally consistent instance identities within each video. To further establish cross-view identity consistency, we concatenate the first frame from each view into a canonical video and reapply SAM3. Since adjacent cameras capture highly overlapping content from closely spaced viewpoints, the resulting canonical masks are naturally aligned and provide a unified identity reference. We compute the IoU between canonical and per-view first-frame masks to establish one-to-one correspondence, and propagate the aligned identities to all frames, yielding temporally and spatially consistent instance masks across all views. Formally, given V views, T frames and D instances, we denote the segmentation of view v at frame t as a multi-class mask $M^{v,t} \in \{0, 1, \ldots, D\}$. The background region is defined as $M_{bg}^{v,t} = M_0^{v,t}$, while $M_d^{v,t}, d \in \{1, \ldots, D\}$ corresponds to the d-th instance identity.

While instance-level geometric partitioning provides spatially grounded identity, it lacks high-level semantic abstractions aligned with natural language. To bridge this gap, we generate instance-wise captions and encode them into embeddings, providing language-aligned supervision for the dynamic semantic field.

Recent MLLMs, such as GPT-4o [15], Qwen3-VL [1], enable high-quality language generation from multimodal inputs. Leveraging the capabilities, we combine visual cues and textual prompts to guide the MLLM in producing temporally consistent and object-specific captions across video frames, capturing appearance, position, and motion patterns. Inspired by 4DLangsplat [34], we generate both time-invariant global captions and time-dependent per-frame captions for each instance, followed by language embedding extraction. For the time-invariant global caption C_i^{global}, we highlight the instance in the first view using the SAM3 mask rendered as a red contour and prompt Qwen3-VL [1] to describe its overall appearance and action characteristics. For the per-frame caption, we aggregate all views at frame t into a spatially consistent video clip and condition the MLLM on the highlighted instance observations $\bar{\mathcal{P}}_i^{v,t}$, the global caption, and a temporal prompt $\mathcal{T}$:

$$C_i^t = \text{MLLM}\big(\{\bar{\mathcal{P}}_i^{v,t}\}_{v=0}^{V-1},\ C_i^{\text{global}},\ \mathcal{T}\big). \tag{1}$$

The generated time-dependent captions are encoded into high-dimensional embeddings using an LLM [65] and further compressed into a compact 6-dimensional

embeddings via a lightweight autoencoder. The embeddings e, together with the instance masks, serve as 2D semantic supervision for learning the dynamic semantic field, enforcing temporally consistent representations across frames.

3.2 Gaussian Initialization

With the aligned instance masks and semantic embeddings as guidance, we explicitly decompose the scene into static background and dynamic foreground, and optimize each layer, $\mathcal{G}_{bg}$ and $\mathcal{G}_{fg}$, with tailored training strategies to enhance consistency and achieve detailed 4D segmentation.

Static Background Initialization. Our goal is to explicitly model the temporally consistent components of the scene, facilitating more accurate foreground modeling and separation. To this end, we uniformly sample ground-truth images across both temporal and spatial dimensions, and jointly optimize global background Gaussians $\mathcal{G}_{bg}$ using the following loss:

$$\begin{aligned} E^{bg}_{\text{color}} &= \|M_{bg}(\hat{\mathbf{C}} - \mathbf{C})\|_1, \\ E_{\text{bg}} &= \lambda_{\text{iso}} E_{\text{iso}} + \lambda_{\text{size}} E_{\text{size}} + E^{bg}_{\text{color}}, \end{aligned} \tag{2}$$

where the color loss E^{bg}_{color} is evaluated exclusively over the background mask M_{bg}. $\hat{\mathbf{C}}$ denotes the blended color after rasterization, and $\mathbf{C}$ is the ground truth. The isotropic loss E_{iso} and the size loss E_{size} regularize the scaling of the Gaussians, preventing them from becoming overly elongated or excessively large. We refer to DualGS [23] for the detailed formulation. The background layer provides a global reference for subsequent optimization, while the Gaussians in the background are also continuously trained in each frame.

First Frame Initialization. Our objective is to construct a temporally coherent 4D representation that jointly models dynamic motion and photometric appearance, while preserving instance-level structure and higher-level semantic embeddings. To achieve this, we equip the Gaussian representation with learnable semantic feature attributes f that explicitly encode instance-level semantic information. Each Gaussian primitive is associated with a learnable, view-independent semantic attribute vector f (8-dimensional in our experiments), initialized by randomly sampling from a uniform distribution.

During optimization, these semantic features are jointly optimized alongside other Gaussian parameters to encode the instance identities in the scene. Given the pre-trained background layer, we optimize $\mathcal{G}_{bg}$ and $\mathcal{G}_{fg}$ together by feeding both into the differentiable Gaussian rasterization. In addition to rendering the color map via alpha blending, we also rasterize the semantic attributes to produce a 2D semantic feature map $F \in \mathbb{R}^{H\times W\times 8}$. To supervise instance identity, F is projected through a classification MLP decoder into $(D{+}1)$ channels, followed by a softmax to produce per-pixel class probabilities P. These per-pixel probabilities are supervised with instance masks M via a cross-entropy loss:

$$E_{\text{id}} = -\sum_{d=0}^{D} y_d \log P_d, \tag{3}$$

where y_d is the one-hot encoding of the ground-truth instance label. We also utilize a separate MLP decoder to produce language-aligned embeddings $\hat{e}$, which are supervised with the ground-truth embeddings e via an L1 loss:

$$E_{\text{emb}} = \|\hat{e} - e\|_1, \tag{4}$$

This dual-decoder design ensures that the model simultaneously captures low-level instance identities and high-level semantic information.

To further enforce the spatial consistency of the semantic feature, we follow the Gaussian Grouping [79] to introduce the 3D regularization term:

$$E_{\text{3D}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} D_{\text{KL}} \left(P_i \| P_j \right), \tag{5}$$

where $\mathcal{S}$ denotes the sampled foreground Gaussians, $\mathcal{N}_k(i)$ is the set of k-nearest Gaussians of i in 3D space, and $D_{\text{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler divergence. This loss enforces spatial smoothness by encouraging neighboring Gaussians within the same instance to learn consistent identity representations, enabling occluded or interior regions to acquire coherent features that facilitate flexible scene manipulation. Overall, the objective function in the first frame training is formulated as follows:

$$E_{\text{init}} = \lambda_{\text{id}} E_{\text{id}} + \lambda_{\text{emb}} E_{\text{emb}} + \lambda_{\text{3D}} E_{\text{3D}} + E_{\text{color}}, \tag{6}$$

where the color loss E_{color} follows the original 3DGS [28].

3.3 Dynamic Optimization

After initialization, we adopt a frame-by-frame optimization strategy that unifies tracking, modeling, and semantic learning within a single representation and trains them jointly. The optimization process consists of two stages: optical-flow-aided explicit warping, and the training phase.

Explicit Warping. Assuming that the foreground Gaussians $\mathcal{G}_{fg}^{t-1}$ from the preceding frame have been well-optimized, we aim to explicitly warp each primitive with position p_{t-1} to $\bar{p}_t$, providing a reliable initialization for the current frame t. Since optical flow operates in the 2D image space, we establish geometric constraints through projection and multi-view least squares. Specifically, we project p_{t-1} onto each camera plane to obtain pixel coordinates $u = (x_v, y_v)$, for $v = 0, ..., V-1$. These 2D coordinates are then used to query the optical flow predicted by SEA-RAFT [71] between frame $t-1$ and t, yielding the warped pixel locations $u'_v = (x'_v, y'_v)$. These pixel positions indicate the potential projections of the Gaussian position in the current frame. We therefore initialize the updated 3D position $\bar{p}_t$ by minimizing the multi-view reprojection error:

$$\bar{p}_t = \arg\min_{p \in \mathbb{R}^3} \sum_{v=0}^{V-1} \|\pi_v(p) - u'_v\|_2^2, \tag{7}$$

where $\pi_v(\cdot)$ denotes the perspective projection under view v. This least-squares problem can be further converted into solving a linear equation.

Training. Notably, the least-squares triangulation is inaccurate due to noise and occlusions, which can introduce incorrect correspondences and lead to implausible initialization. Inspired by HiFi4G [24], we introduce a refinement stage that employs a locally as-rigid-as-possible (ARAP) regularizer to enforce local geometric consistency:

$$E_{\text{smooth}} = \sum_i \sum_{k \in \mathcal{N}(i)} w_{i,k}^{t-1} \| R\left(q_{i,t} * q_{i,t-1}^{-1}\right)\left(p_{k,t-1} - p_{i,t-1}\right) - \left(p_{k,t} - p_{i,t}\right) \|_2^2, \tag{8}$$

where i indexes a Gaussian primitive and $R(\cdot)$ converts a quaternion to a rotation matrix and w corresponds to the blending weights $w_{i,k}^t = \exp\left(-\|p_{i,t} - p_{k,t}\|_2^2 / l^2\right)$. l is the influence radius. In this phase, we keep the appearance and semantic attributes fixed and optimize only the Gaussian positions and rotations to correct physically implausible deformations caused by explicit warping. The objective is formulated as:

$$E_{\text{ft}} = \lambda_{\text{smooth}} E_{\text{smooth}} + E_{\text{color}}, \tag{9}$$

After refinement, we jointly optimize the Gaussians' motion, appearance, and semantic feature attributes. Foreground Gaussian primitives dynamically split and prune to accommodate new observations, while cloned Gaussians inherit all parents attributes, ensuring continuous trajectory tracking. To maintain semantic feature consistency, we constrain each Gaussian's 2D projections to remain within its assigned instance masks, promoting effective learning. For each instance d, we first pre-compute a signed distance field (SDF) $\Phi_d^{v,t}$ from the its instance mask $M_d^{v,t}$, where each pixel u is assigned the Euclidean distance to the nearest mask boundary. Recall that the classification MLP produces per-pixel semantic logits, which can also be applied to each Gaussian's semantic feature. We then compute a softmax to predict Gaussian i's instance label c_i, and define the SDF loss only for Gaussians whose predicted label matches instance d, defined as:

$$E_{\text{sdf}} = \sum_i y_d \cdot \left[\max(0, \Phi_d^{v,t}(\pi_v(p_{i,t})))\right]^2, \tag{10}$$

This formulation penalizes Gaussians that lie outside the 2D silhouette of their assigned instance, reinforcing spatial consistency. Besides, we also apply several regularization terms to maintain temporal consistency. For the background layer, we continuously optimize the appearance attributes $a_{i,t}$ to capture potential lighting and shadow changes, while constraining them to remain close to the global background appearance attributes a_i from $\mathcal{G}_{bg}$ to prevent jitter:

$$E_{\text{temp}}^{bg} = \sum_i \|a_{i,t} - a_i\|_2^2, \tag{11}$$

We further extend this regularization across consecutive frames, denoted as E_{temp}, to prevent abrupt changes in all attributes, avoiding unstable tracking or

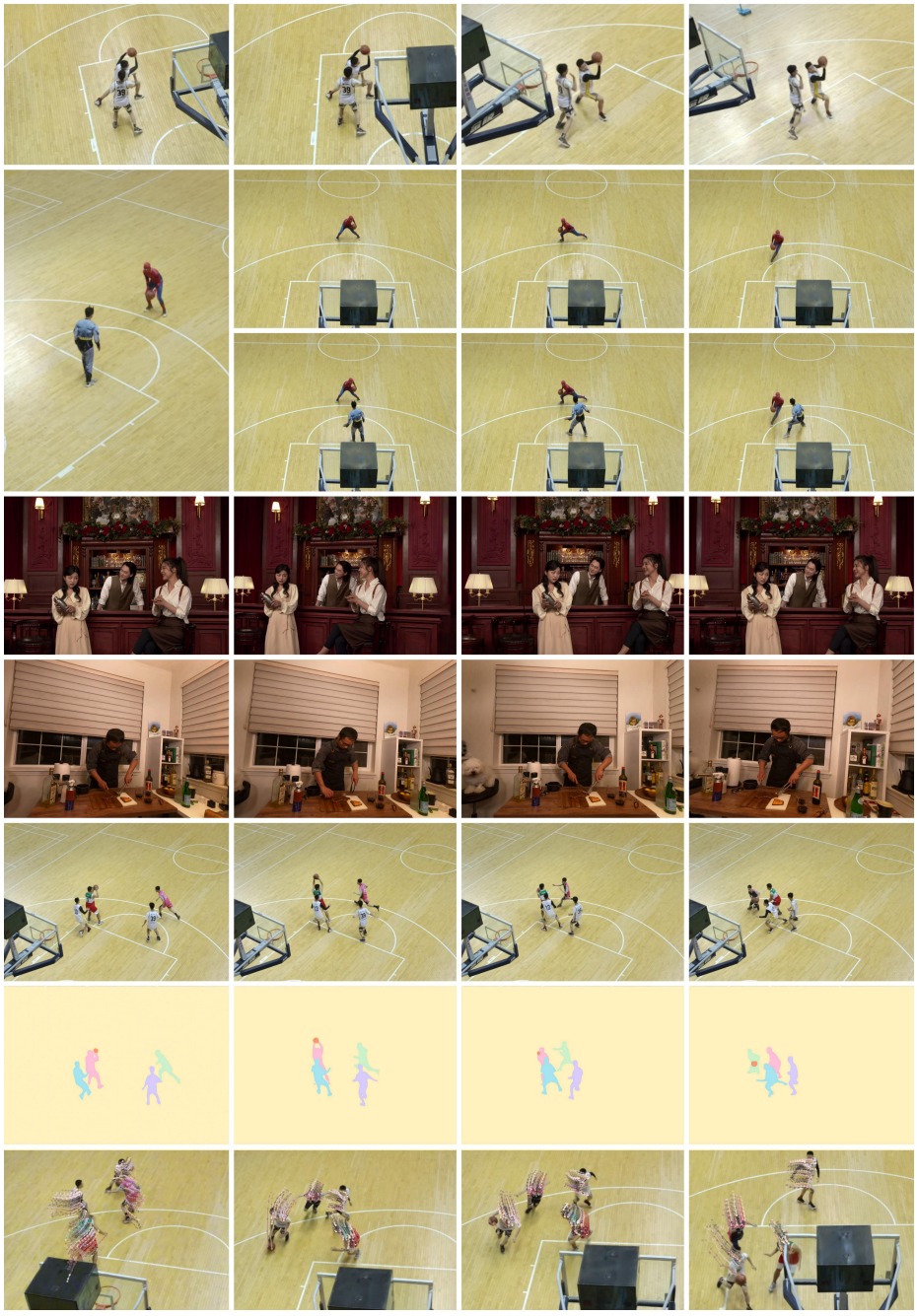

Fig. 3: Gallery of our results. All images are rendered from novel views. The first five rows show high-fidelity rendering results under challenging scenarios, including fast motions and severe occlusions. The sixth row presents the corresponding instance-level 4D segmentation of the fifth row. The last row visualizes trajectory renderings

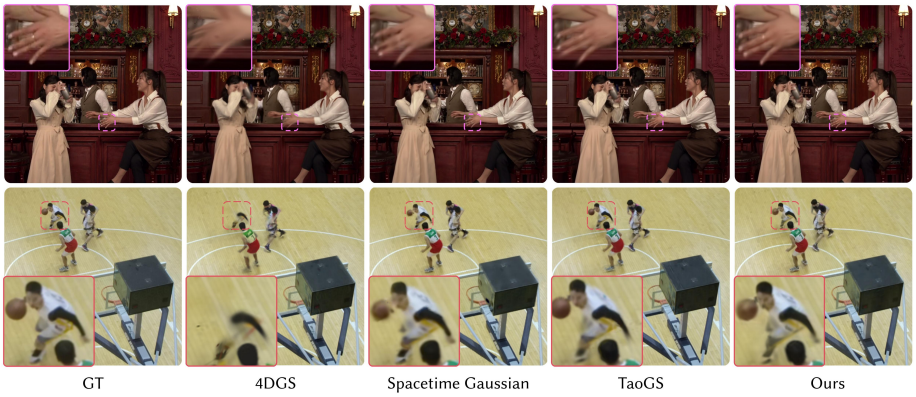

Fig. 4: Qualitative comparison with 4DGS [72], Spacetime Gaussian [37], and TaoGS [20]. Ours shows the best quality. Note that the zoomed-in regions are extremely small in the original image, where even the ground truth is not very sharp.

Table 1: Rendering quality comparison on the corresponding dataset. Green and yellow cells indicate the best and second-best results, respectively.

Method	ST-NeRF			MPEG GSC		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
4DGS	34.722	0.956	0.0625	33.542	0.936	0.0912
Spacetime Gaussian	37.949	0.965	0.0524	37.807	0.954	0.0526
TaoGS	37.719	0.963	0.0564	36.137	0.952	0.0671
Ours	38.912	0.967	0.0463	38.241	0.959	0.0537

sudden semantic shifts. Overall, the training-phase loss is defined as:

$$E_{\text{train}} = \lambda_{\text{sdf}} E_{\text{sdf}} + \lambda_{\text{temp}}^{bg} E_{\text{temp}}^{bg} + \lambda_{\text{temp}} E_{\text{temp}} + \lambda_{\text{smooth}} E_{\text{smooth}} + \lambda_{\text{id}} E_{\text{id}} + \lambda_{\text{emb}} E_{\text{emb}} + \lambda_{\text{3D}} E_{\text{3D}} + E_{\text{color}}. \tag{12}$$

Querying. After training, we decode the rendered feature map into high-dimensional space and compute relevance scores via cosine similarity with the LLM-processed query. Specifically, we first retrieve the instance ID in the first frame using the identity query, then render the RGB and feature maps for that instance only, and finally retrieve its most relevant frame segments via the segment query. This approach effectively filters out interference from other dynamic instances, enabling precise instance-level retrieval and analysis.

Implementation Details. For static background initialization, we set $\lambda_{\text{iso}} = 0.0005$ and $\lambda_{\text{size}} = 0.02$, optimizing for 20,000 iterations. For first frame initialization, we jointly optimize the Gaussian and two MLPs for 30,000 iterations, using $\lambda_{\text{id}} = 2$, $\lambda_{\text{emb}} = 10$, and $\lambda_{\text{3D}} = 2$. For the E_{3D}, we sample 6000 foreground

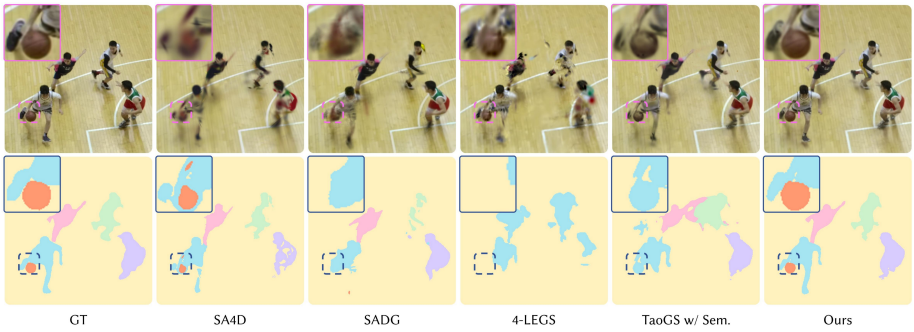

Fig. 5: Qualitative comparison with 4D segmentation methods, including SA4D [19], SADG [36], and 4-LEGS [9], as well as TaoGS [20] w/ Sem.. Ours achieves more accurate instance segmentation.

Table 2: Rendering quality and segmentation comparison on the basketball dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	Recall ↑	F_1 ↑
SA4D	33.951	0.942	0.0917	0.5235	0.5667	0.6464
SADG	33.421	0.942	0.1259	0.4995	0.5330	0.6052
4-LEGS	31.655	0.939	0.0784	–	–	–
TaoGS w/ Sem.	37.364	0.963	0.0530	0.5780	0.6799	0.6757
Ours	38.298	0.965	0.0470	0.8305	0.8764	0.8878

Gaussians and use $k = 4$ nearest neighbors for KL divergence. During the refinement stage, we apply $\lambda_{\text{smooth}} = 0.01$ and finetune for 3000 iterations. In the dynamic training stage, we adopt the following empirical weights: $\lambda_{\text{sdf}} = 0.01$, $\lambda_{\text{temp}}^{bg} = 0.001$, $\lambda_{\text{temp}} = 0.01$, $\lambda_{\text{smooth}} = 0.0001$, $\lambda_{\text{id}} = 1$, $\lambda_{\text{emb}} = 10$, and $\lambda_{\text{3D}} = 2$. We train each frame for 8000 iterations. During the training, we freeze the classification MLP, continue optimizing only the semantic MLP. Every 300 iterations, we clone and prune foreground Gaussians, ensuring that the total number of modified Gaussians per frame remains below 5%.

4 Experimental Results

To demonstrate the capabilities of Director, we evaluate our method on two challenging datasets: the ST-NeRF basketball dataset [85] and the MPEG GSC Dataset [18]. The basketball dataset is captured using 32 pre-synchronized and calibrated Z-Cam cameras, recording 4K videos at 30 fps. It features multi-player basketball games with fast and complex motions, as well as challenging multi-person interactions such as screens, defensive switches, fast-break layups, and jump shots. The MPEG GSC Dataset [18] contains two indoor multi-view sequences featuring fast motion scenarios, captured by 20 cameras at 1080p resolution. We conduct all experiments on a single NVIDIA RTX 3090 GPU.

Fig. 6: Visualization of querying results. The top row shows the query–frame similarity scores across frames. The bottom row presents several reference views.

Table 3: Quantitative ablation study on training stages.

	PSNR ↑	mIoU ↑	Recall ↑	F_1 ↑
w/o Semantic Feature	37.661	–	–	–
w/o Explicit Warping	37.091	0.8008	0.8420	0.8743
w/o Motion Fine-tuning E_{ft}	37.618	0.6987	0.7228	0.7881
w/o Training E_{train}	36.796	0.5105	0.5233	0.6065
Ours	**38.382**	**0.8328**	**0.8690**	**0.8940**

Table 4: Quantitative comparison of different loss term contributions.

	PSNR ↑	mIoU ↑	Recall ↑	F_1 ↑
w/o E_{sdf}	37.604	0.7404	0.7780	0.7952
w/o E_{temp}^{bg}	37.967	0.8095	0.8548	0.8693
w/o E_{temp}	37.529	0.7613	0.8284	0.8370
w/o E_{smooth}	37.216	0.7915	0.8653	0.8631
Ours	**38.284**	**0.8341**	**0.8794**	**0.8907**

Background and first-frame initialization take about 50 minutes, while dynamic optimization takes roughly 10 minutes per frame. As illustrated in Fig. 3, our method enables robust tracking and high-fidelity rendering, while also producing temporally consistent 4D instance segmentation for scene editing. This allows flexible manipulation of dynamic elements, such as changing an instance's position, size, or orientation, or removing it entirely.

4.1 Comparison

Qualitative Evaluation. We compare Director against state-of-the-art dynamic rendering methods, including 4DGS [72], Spacetime Gaussian [37], and TaoGS [20]. As shown in Fig. 4, 4DGS struggles to model fast motions effectively, while Spacetime Gaussian produces blurry results. TaoGS suffers from over-smoothing, leading to noticeable artifacts. In contrast, Director explicitly decomposes the scene, leverages optical-flow guidance, and incorporates regularization terms to achieve high-fidelity rendering even under rapid and challenging motions. To further validate the 4D segmentation performance, we compare our method against with four baselines, SA4D [19], SADG [36], 4-LEGS [9], all provided with SAM3 masks as input. Additionally, we construct a strict reconstruction-then-understanding baseline by first reconstructing the scene with TaoGS [20] and then applying the same semantic module to the fixed recon-

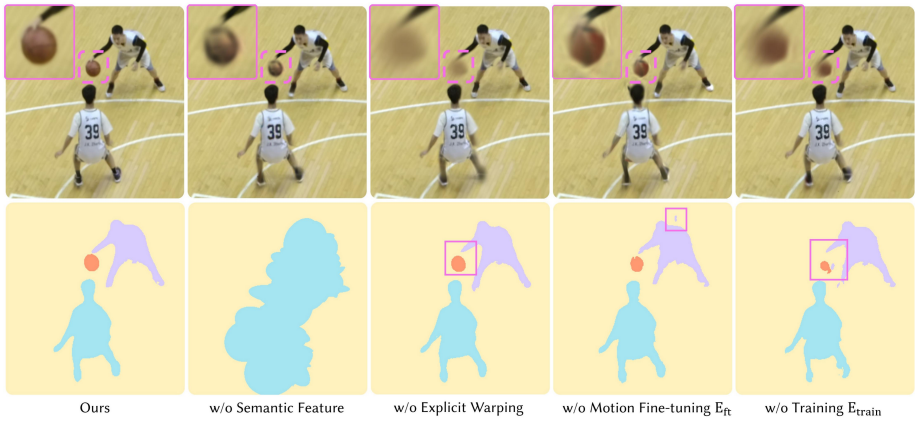

Fig. 7: Qualitative ablation study on key training components.

struction, denoted as TaoGS w/ Sem. As shown in Fig. 5, SA4D struggles with fast motions, resulting in blurred and incomplete masks. SADG also has difficulty with rapid movements and often merges closely interacting instances. 4-LEGS produces inferior rendering quality and supports only foreground-level mask queries, with instance-level queries fail entirely. Although TaoGS w/ Sem achieves reasonably good rendering quality, its decoupled two-stage design cannot fix inter-instance entanglement after photometric reconstruction. In contrast, our method delivers high-quality renderings while producing 4D segmentation results that are visually more accurate than the SAM3 inputs.

Quantitative Comparison. We select three 200-frame sequences from the ST-NeRF basketball dataset and two 100-frame sequences from the MPEG GSC dataset, evaluating rendering performance across three test views. For rendering, we evaluate PSNR, SSIM, LPIPS in Tab. 1. Our method achieves the highest rendering quality, surpassing existing approaches. To assess 4D segmentation, we evaluate mIoU, Recall, and F1 on instance masks, where mIoU measures mask overlap, Recall quantifies correctly predicted instance pixels, and F1 balances precision and recall. Instance-level masks for 4-LEGS are difficult to obtain, so we exclude it from this comparison. As shown in Tab. 2, our method achieves the highest 4D segmentation accuracy while maintaining superior rendering quality.

We also visualize the temporal evolution of query-frame similarity scores in Fig. 6, using the average across all frames as the threshold. For reference, we also provide the raw image with the target instance highlighted.

4.2 Ablation Study

Key Components. We perform an ablation study to evaluate the impact of key training components in our optimization. As shown in Fig. 7 and Tab. 3, disabling semantic features leads to under-constrained 4D Gaussian shapes, blurred

renderings, and failed instance-wise segmentation. Omitting Optical-Flow-Aided Explicit Warping causes failures in regions with fast motion. Skipping motion fine-tuning E_{ft} leads to artifacts that disrupt temporal tracking, while removing the training E_{train} leads to severe blurring. In contrast, our full method effectively leverages both RGB and semantic supervision, producing clear and temporally consistent reconstructions.

Regularizers. We evaluate the influence of regularizers during training. As shown in Tab. 4, the SDF term constrains Gaussians to lie within the instance silhouette, the temporal term enforces consistency across frames, and the smoothness term ensures physically plausible non-rigid deformations. Omitting any of these terms degrades performance. In contrast, our full method produces high-quality rendering results and accurate 4D segmentation.

5 Limitations

Despite Director enabling high-fidelity rendering, robust tracking, and accurate segmentation for dynamic scenes, it has several limitations. First, training dynamic Gaussians remains relatively slow, which limits practical applications. Accelerating this step is an important direction for future work. Second, our framework relies on multiple loss terms with carefully tuned hyperparameters. Adapting these to significantly different scenes may require additional tuning. Finally, due to current computational constraints, we encode semantic features into a compact, low-dimensional latent space to enable Gaussian training. This limits the richness of semantic information that can be represented.

6 Conclusion

In this work, we presented Director, a unified 4D Gaussian framework that integrates geometric modeling and semantic understanding for highly dynamic scenes. By leveraging spatially and temporally consistent instance masks alongside language-aligned embeddings from MLLMs, our method achieves robust human performance tracking, high-fidelity rendering, and precise instance-level segmentation. The two-layer Gaussian representation, combined with an optical-flow-aided optimization strategy, ensures high temporal consistency even under fast and complex motions. Experiments demonstrate that Director effectively couples reconstruction and semantic reasoning, enabling open-vocabulary queries and representing a significant step toward more immersive virtual experiences.

Acknowledgements. This work was supported by the Science and Technology Commission of Shanghai Municipality under the Science and Technology Program (25511106302), the National Natural Science Foundation of China under Grant W2431046, the Central Guided Local Science and Technology Foundation of China under Grant YDZX20253100001001, the MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence and HPC Platform of ShanghaiTech University.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
4. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics (TOG) **34**(4), 69 (2015)
5. Dou, M., Davidson, P., Fanello, S.R., Khamis, S., Kowdle, A., Rhemann, C., Tankovich, V., Izadi, S.: Motion2fusion: Real-time volumetric performance capture. ACM Trans. Graph. **36**(6), 246:1–246:16 (Nov 2017)
6. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., Izadi, S.: Fusion4d: real-time performance capture of challenging scenes. ACM Trans. Graph. **35**(4) (jul 2016). `https://doi.org/10.1145/2897824.2925969`, `https://doi.org/10.1145/2897824.2925969`
7. 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: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
8. Feng*, H., Zhang*, J., Wang, Q., Ye, Y., Yu, P., Black, M.J., Darrell, T., Kanazawa, A.: St4rtrack: Simultaneous 4d reconstruction and tracking in the world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
9. Fiebelman, G., Cohen, T., Morgenstern, A., Hedman, P., Averbuch-Elor, H.: 4-legs: 4d language embedded gaussian splatting. In: Computer Graphics Forum. vol. 44, p. e70085. Wiley Online Library (2025)
10. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
11. Gao, Q., Xu, Q., Cao, Z., Mildenhall, B., Ma, W., Chen, L., Tang, D., Neumann, U.: Gaussianflow: Splatting gaussian dynamics for 4d content creation. arXiv preprint arXiv:2403.12365 (2024)
12. Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., Liu, Y.: Real-time geometry, albedo and motion reconstruction using a single rgbd camera. ACM Transactions on Graphics (TOG) (2017)
13. Habermann, M., Liu, L., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM Transactions on Graphics **40**(4) (aug 2021)
14. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG) **38**(2), 14:1–14:17 (2019)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

16. Işık, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: Humanrf: High-fidelity neural radiance fields for humans in motion. ACM Transactions on Graphics (TOG) **42**(4), 1–12 (2023). https://doi.org/10.1145/3592415, https://doi.org/10.1145/3592415
17. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: VolumeDeform: Real-time Volumetric Non-rigid Reconstruction (October 2016)
18. Jeong, J.Y., Kim, J., Lee, B.H., Yun, K.J., Cheong, W., Yoo, S.H.: [INVR] Multi-view datasets Classroom and Bartender for 3D INVR activity. Tech. Rep. M69151, ISO/IEC JTC1/SC29/WG4 MPEG, Sapporo, Japan (July 2024)
19. Ji, S., Wu, G., Fang, J., Cen, J., Yi, T., Liu, W., Tian, Q., Wang, X.: Segment any 4d gaussians. arXiv preprint arXiv:2407.04504 (2024)
20. Jiang, Y., Guo, C., Wu, Y., Hong, Y., Zhu, S., Shen, Z., Zhang, Y., Jiao, S., Su, Z., Xu, L., et al.: Topology-aware optimization of gaussian primitives for human-centric volumetric videos. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
21. Jiang, Y., Jiang, S., Sun, G., Su, Z., Guo, K., Wu, M., Yu, J., Xu, L.: Neuralhofusion: Neural volumetric rendering under human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6155–6165 (2022)
22. Jiang, Y., Shen, Z., Guo, C., Hong, Y., Su, Z., Zhang, Y., Habermann, M., Xu, L.: Reperformer: Immersive human-centric volumetric videos from playback to photo-real reperformance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11349–11360 (2025)
23. 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)
24. Jiang, Y., Shen, Z., Wang, P., Su, Z., Hong, Y., Zhang, Y., Yu, J., Xu, L.: Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19734–19745 (2024)
25. Jiang, Y., Yao, K., Su, Z., Shen, Z., Luo, H., Xu, L.: Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 595–605 (2023)
26. Jin, L., Tucker, R., Li, Z., Fouhey, D., Snavely, N., Holynski, A.: Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
27. Kania, K., Yi, K.M., Kowalski, M., Trzciński, T., Tagliasacchi, A.: Conerf: Controllable neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18623–18632 (2022)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (ToG) **42**(4), 1–14 (2023)
29. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19729–19739 (2023)
30. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. Advances in neural information processing systems **35**, 23311–23330 (2022)

31. Kwon, Y., Liu, L., Fuchs, H., Habermann, M., Theobalt, C.: Deliffas: Deformable light fields for fast avatar synthesis. Advances in Neural Information Processing Systems **36** (2024)
32. Li, H., Li, S., Gao, X., Batuer, A., Yu, L., Liao, Y.: Gifstream: 4d gaussian-based immersive video with feature stream. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21761–21770 (2025)
33. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Lv, Z.: Neural 3d video synthesis (2021)
34. Li, W., Zhou, R., Zhou, J., Song, Y., Herter, J., Qin, M., Huang, G., Pfister, H.: 4d langsplat: 4d language gaussian splatting via multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22001–22011 (2025)
35. Li, Y., Habermann, M., Thomaszewski, B., Coros, S., Beeler, T., Theobalt, C.: Deep physics-aware inference of cloth deformation for monocular human performance capture. In: 2021 International Conference on 3D Vision (3DV). pp. 373–384. IEEE (2021)
36. Li, Y.J., Gladkova, M., Xia, Y., Cremers, D.: Sadg: Segment any dynamic gaussian without object trackers. arXiv preprint arXiv: 2411.19290 (2024)
37. 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. pp. 8508–8520 (2024)
38. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
39. Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8762–8771 (2021)
40. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 3DV (2024)
41. Luo, H., Xu, T., Jiang, Y., Zhou, C., Qiu, Q., Zhang, Y., Yang, W., Xu, L., Yu, J.: Artemis: Articulated neural pets with appearance and motion synthesis. ACM Trans. Graph. **41**(4) (jul 2022). https://doi.org/10.1145/3528223.3530086, https://doi.org/10.1145/3528223.3530086
42. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 405–421. Springer International Publishing, Cham (2020)
43. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015)
44. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1165–1175 (2024)
45. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)

46. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (dec 2021)
47. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 815–824 (2023)
48. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5020–5030 (2024)
49. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)
50. Rai, A., Xing, A., Agarwal, A., Cong, X., Li, Z., Lu, T., Prakash, A., Sridhar, S.: Packuv: Packed gaussian uv maps for 4d volumetric video. arXiv preprint arXiv:2602.23040 (2026)
51. Shao, R., Chen, L., Zheng, Z., Zhang, H., Zhang, Y., Huang, H., Guo, Y., Liu, Y.: Floren: Real-time high-quality human performance rendering via appearance flow using sparse rgb cameras. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–10 (2022)
52. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16632–16642 (2023)
53. Shen, Z., Cai, Y., Lu, Y., Hong, Y., Wu, Y., Zheng, M., Zhang, Y., Xu, L.: Dynamic gaussian streams for volumetric video via codebook-based quantization. 2025 IEEE International Workshop on Multimedia Signal Processing (MMSP) pp. 340–345 (2025), `https://api.semanticscholar.org/CorpusID:284725053`
54. Slavcheva, M., Baust, M., Cremers, D., Ilic, S.: Killingfusion: Non-rigid 3d reconstruction without correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1386–1395 (2017)
55. Slavcheva, M., Baust, M., Ilic, S.: Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2646–2655 (2018)
56. Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., Geiger, A.: Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. IEEE Transactions on Visualization and Computer Graphics **29**(5), 2732–2742 (2023)
57. Su, Z., Xu, L., Zheng, Z., Yu, T., Liu, Y., Fang, L.: Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 246–264. Springer International Publishing, Cham (2020)
58. Su, Z., Xu, L., Zhong, D., Li, Z., Deng, F., Quan, S., Fang, L.: Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
59. Sun, G., Chen, X., Chen, Y., Pang, A., Lin, P., Jiang, Y., Xu, L., Wang, J., Yu, J.: Neural free-viewpoint performance rendering under complex human-object interactions. In: Proceedings of the 29th ACM International Conference on Multimedia (2021)

60. 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. pp. 20675–20685 (2024)
61. Suo, X., Jiang, Y., Lin, P., Zhang, Y., Wu, M., Guo, K., Xu, L.: Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6226–6237 (2021)
62. Vora, S., Radwan, N., Greff, K., Meyer, H., Genova, K., Sajjadi, M.S., Pot, E., Tagliasacchi, A., Duckworth, D.: Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. arXiv preprint arXiv:2111.13260 (2021)
63. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3835–3844 (2022)
64. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
65. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Improving text embeddings with large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11897–11916 (2024)
66. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
67. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: International Conference on Computer Vision (2023)
68. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9660–9672 (2025)
69. Wang*, Q., Zhang*, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: CVPR (2025)
70. Wang, S., Geiger, A., Tang, S.: Locally aware piecewise transformation fields for 3d human mesh registration. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
71. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: European Conference on Computer Vision. pp. 36–54. Springer (2024)
72. 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)
73. Wu, X., Bai, Y., Li, M., Wu, X., Zhao, X., Lai, Z., Liu, W., Wang, X.: 4dlangvggt: 4d language-visual geometry grounded transformer. arXiv preprint arXiv:2512.05060 (2025)
74. Xiang, D., Prada, F., Wu, C., Hodgins, J.: Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In: 2020 International Conference on 3D Vision (3DV). pp. 322–332. IEEE (2020)
75. Xu, L., Su, Z., Han, L., Yu, T., Liu, Y., Fang, L.: Unstructuredfusion: realtime 4d geometry and texture reconstruction using commercial rgbd cameras. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2508–2522 (2019)

76. Xu, Z., Bi, S., Sunkavalli, K., Hadap, S., Su, H., Ramamoorthi, R.: Deep view synthesis from sparse photometric images. ACM Transactions on Graphics (TOG) **38**(4), 1–13 (2019)
77. Yang, Z., Yang, H., Pan, Z., Zhu, X., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023)
78. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20331–20341 (2024)
79. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: European conference on computer vision. pp. 162–179. Springer (2024)
80. Yu, H.X., Guibas, L.J., Wu, J.: Unsupervised discovery of object radiance fields. arXiv preprint arXiv:2107.07905 (2021)
81. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5746–5756 (2021)
82. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
83. Yuan, Y.J., Sun, Y.T., Lai, Y.K., Ma, Y., Jia, R., Gao, L.: Nerf-editing: geometry editing of neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18353–18364 (2022)
84. Zhang, H., Lin, S., Shao, R., Zhang, Y., Zheng, Z., Huang, H., Guo, Y., Liu, Y.: Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023)
85. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics (TOG) **40**(4), 1–18 (2021)
86. Zhao, F., Jiang, Y., Yao, K., Zhang, J., Wang, L., Dai, H., Zhong, Y., Zhang, Y., Wu, M., Xu, L., Yu, J.: Human performance modeling and rendering via neural animated mesh. ACM Trans. Graph. **41**(6) (nov 2022). https://doi.org/10.1145/3550454.3555451, https://doi.org/10.1145/3550454.3555451
87. Zhu, S., Guo, C., Lu, Y., Shen, Z., Wu, Y., Hong, Y., Cai, Y., Zheng, M., Zhang, Y., Xu, L., Yu, J.: A general framework for gaussian splatting-based human-centric volumetric videos. Visual Intelligence **4**, 8 (2026). https://doi.org/10.1007/s44267-026-00111-7, https://www.sciopen.com/article/10.1007/s44267-026-00111-7