FVVSPLAT: FEED-FORWARD STREAMABLE FREE-VIEWPOINT VIDEO RECONSTRUCTION FROM UNCALIBRATED INPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Streamable Free-Viewpoint Video (FVV) reconstruction aims to recover dynamic 3D scenes from multi-view observations, requiring high-quality, on-the-fly reconstruction under strict constraints. Existing methods rely on known camera parameters and slow per-scene optimization, which hinders practical application. In this work, we introduce **FVVSplat**, a Feed-Forward Streamable **F**ree-**V**iewpoint **V**ideo Reconstruction framework that instantly predicts dynamic 3D Gaussian **Splat**ting representations from uncalibrated multi-view video streams. There are three key innovations are proposed. 1) an adaptive radius sampling mechanism for accurate Gaussian prediction from uncalibrated inputs; 2) a bidirectional deformation that ensures reliable, persistent Gaussians across frames and mitigates long-term error accumulation; 3) an efficient attention is integrated into both static prediction and dynamic deformation with linear complexity, substantially reducing computational cost while maintaining reconstruction quality across video resolutions. Extensive experiments on multiply streamable FVV reconstruction benchmarks demonstrate that FVVSplat significantly improves both reconstruction quality and inference time, with a $15\times$ speedup and 2.16 dB PSNR gain over existing optimization-based methods on the N3DV dataset. Our source code and models will be available at `http://github.com`.

1 INTRODUCTION

Free-viewpoint video (FVV), which enables immersive experiences, has emerged as a promising application in virtual reality, telepresence, and remote education. In recent years, 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) has enabled efficient training and real-time rendering via differentiable rasterization, significantly advancing the field of novel view synthesis. To achieve streamable FVV reconstruction, some methods extend 3DGS into the temporal dimension via dynamic 3D Gaussian Splatting, enabling on-the-fly per-frame optimization. They typically follow two essential steps: (i) camera calibration Schonberger & Frahm (2016) and optimization of static 3D Gaussians on the initial frame Chen et al. (2024; 2025); and (ii) prediction of a deformation field across frames to capture Gaussian dynamics and maintain temporal coherence Yang et al. (2024c); Zhu et al. (2024). For example, 3DGStream Sun et al. (2024a) pioneers streamable FVV reconstruction via initialization of static 3D Gaussian from calibrated frames and learnable neural transformation cache, while subsequent works improve reconstruction performance through advanced techniques such as rate-aware training Hu et al. (2025a), vector-quantized residuals Girish et al. (2024), and Gaussian motion optimization Gao et al.; Fu et al. (2026); Lee et al. (2026).

Despite recent advances, these methods still suffer from suboptimal reconstruction quality and slow training speed (typically requiring over 2 seconds on the benchmarks). Recent feed-forward techniques have significantly advanced dynamic scene reconstruction owing to their strong representation ability and fast single-forward-pass inference. Although some recent feed-forward 4D scene reconstruction methods Wu et al. (2026); Chen et al. (2026b); Balice et al. (2026); Xu et al. (2025); Yang et al. (2026) enable real-time superior reconstruction, but they typically require access to all video frames and operate at small resolutions e.g., 256×256. This hinders streamable reconstruction and complicates on-the-fly high-resolution FVV reconstruction.

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

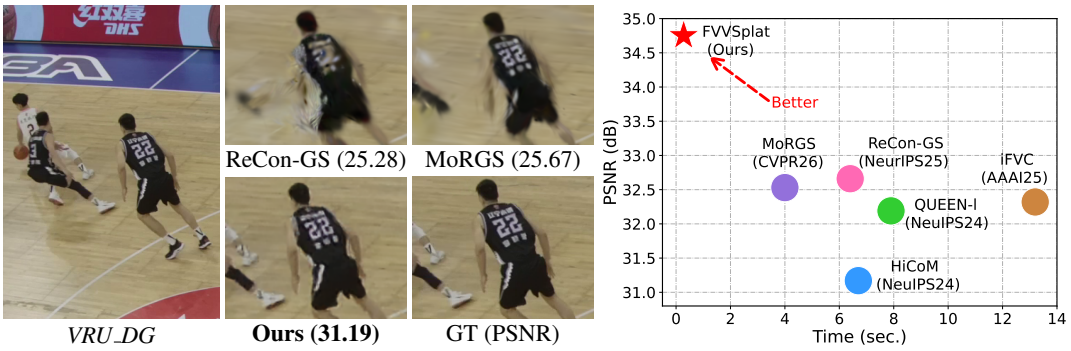

Figure 1: Comparisons of optimization-based streamable FVV reconstruction methods and our feed-forward method. **Left:** A qualitative comparison between the recent optimization-based methods for MoRGS Lee et al. (2026), ReCon-GS Fu et al. (2026), and the proposed feed-forward method. FVVSplat significantly outperforms other methods. **Right:** Comparison among streamable FVV reconstruction methods on the N3DV dataset. FVVSplat (Ours) achieves achieves the state-of-the-art performance in terms of reconstruction quality and speed on standard benchmark.

To address these issues, we introduce **FVVSplat**, an efficient feed-forward dynamic 3D Gaussian Splatting framework for streamable FVV reconstruction. Given uncalibrated multi-view video streams, FVVSplat directly and on-the-fly predicts dynamic 3DGS representations in frame-by-frame streaming manner. There are three technical innovations are proposed: (1) Adaptive anchored position sampling: A mechanism that accurate predicts 3D Gaussians from uncalibrated inputs. This approach captures geometric uncertainty and avoids local minima common in feed-forward models Charatan et al. (2024). (2) bidirectional deformation: A method that estimates both forward and backward deformation the previous canonical state and the current observation. This yields reliable Gaussians across frames and mitigates long-term error accumulation. (3) Efficient attention with liner complexity: An operation naturally introduced into the static prediction and dynamic deformation to reduce computational cost while maintain reconstruction quality. The primary contributions are summarized as:

- We introduce **StreamSplat**, first feed-forward framework for Streamable Free-Viewpoint Video reconstruction from uncalibrated inputs. Existing optimization-based method rely on the known camera parameters and per-scene optimization training.
- We propose three key designs: 1) an adaptive radius sampling mechanism for accurate 3DGS position prediction, 2) bidirectional deformation for robust dynamic Gaussian modeling, and 3) an efficient attention that substantial reduces complexity cost on desired high-resolution datasets for static prediction and dynamic deformation while maintaining the reconstruction quality.
- We evaluate our framework on the standard streamable FVV reconstruction benchmarks and achieve state-of-the-art performance with a more than **15× speedup** and **2.16 dB PSNR gain** over optimization-based baselines, as shown in Figure 1.

2 Related Work

2.1 Dynamic Scene Reconstruction

Early dynamic 3D scene reconstruction approaches depend on the **Optimization-based** models that use implicit MLPs optimized per scene Ouyang et al. (2024); Chen et al. (2021); Sitzmann et al. (2020), or add time-conditioned deformation fields Park et al. (2021a;b); Li et al. (2022b), but still struggle with large, complex motion. Recent methods adopt explicit 3D primitives, particularly dynamic Gaussian Splattings Luiten et al. (2024); Li et al. (2024); Duan et al. (2024); Sun et al. (2024b); Lee et al. (2024), for efficient, real-time rendering. However, they still rely on calibrated camera poses and extensive per-scene optimization, making them unsuitable for fully uncalibrated and real-time scenarios. By directly predicting 3D scene representations via neural networks, **Feed-**

Forward-based methods have emerged as promising alternatives to optimization-based approaches. Recent scene-coordinate-based methods Wang et al. (2024); Leroy et al. (2024); Smart et al. (2024); Wang et al. (2025) can estimate camera parameters and 3D structure from uncalibrated inputs but target only static scenes. MonST3R Zhang et al. (2024) extends to dynamic scene estimations, but typically requires pose/geometry post-optimization, and coordinate-based point clouds remain hard to deform Luiten et al. (2024). Recent dynamic variants Liang et al. (2024); Yang et al. (2024a); Shen et al. (2025) are proposed. However, they still rely on camera calibration and offline manner with full-sequence access, thus failing to support uncalibrated online streaming.

2.2 Streamable Free-Viewpoint Video Reconstruction

Compared to offline methods, online reconstruction enables free-viewpoint video (FVV) to be trained and transmitted incrementally in a frame-by-frame manner, allowing users to preview or interact with the content in real-time. Some advanced structures have been proposed, i.e., neural radiance field (NeRF) Mildenhall et al. (2021) and 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) for streeamable FVV reconstruction. Recently, with the fast training, real-time rendering, and high-fidelity synthesis of 3DGS Kerbl et al. (2023), many 3DGS-based methods Sun et al. (2024a); Gao et al. (2024); Yan et al. (2025); Girish et al. (2024); Fu et al. (2026); Chen et al. (2026a); Lee et al. (2026) have been proposed. For example, 3DGStream Sun et al. (2024a) introduces a hash-based MLP to encode the position and rotation transformation of Gaussians at each frame, and adopts a Gaussian addition strategy for the adaptation of dynamic changes across frames. HiCoM Gao et al. (2024) designs a hierarchical coherent motion mechanism to effectively capture and represent scene motion for fast and accurate training. QUEEN Girish et al. (2024) proposes a Gaussian residual-based framework for model expressiveness reconstruction and a learned quantization-sparsity framework for residuals compression. ReCon-GS Fu et al. (2026) designs a dynamic hierarchy reconfiguration strategy that preserves localized motion expressiveness through anchor re-hierarchization, while ensuring temporal consistency through intra-hierarchical deformation. However, they still depend on calibrated camera poses and suffer from heavy per-frame optimization latency, making them inapplicable to fully uncalibrated and real-time scenarios. Recently, some feed-forward streamable dynamic scene reconstruction methods Wu et al. (2026); Chen et al. (2026b); Balice et al. (2026); Xu et al. (2025); Yang et al. (2026); Jiang et al. (2026); Choi et al. (2025); Hu et al. (2025b) proposed that uses the feed-forward fundation models to achieve the significant performance improvements. In contrast, FVVSplat operates strictly online on uncalibrated inputs, maintaining temporally consistent dynamic Gaussians via bidirectional deformation and adaptive Gaussian fusion.

3 Preliminary

3D Gaussian Splatting. 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) represents a static scene using a collection of 3D Gaussians. Each Gaussian $\mathcal{G}$ is defined by its mean (position) $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, opacity $\boldsymbol{\alpha}$, and color represented by spherical harmonics (SH) coefficients $\boldsymbol{c}$. The final opacity of a 3D Gaussian at a given point $\mathbf{x}$ is computed as:

$$\boldsymbol{\alpha}(\mathbf{x}) = \boldsymbol{\alpha} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{1}$$

Normally, the covariance matrix $\boldsymbol{\Sigma}$ is decomposed into a diagonal scaling matrix $\mathbf{S}$ and a rotation quaternion $\mathbf{R}$ for differentiable optimization: $\boldsymbol{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$. Given a set of N 3DGS $\mathcal{G} = \{G_i\}_{i=1}^N$, the rendering process involves splatting them onto the 2D image plane and then blending their colors based on their opacity and depth. The 3D Gaussians are projected using the approximate transformation Zwicker et al. (2001): $\boldsymbol{\Sigma}' = \mathbf{JW\Sigma W}^T\mathbf{J}^T$, where $\mathbf{J}$ is the Jacobian of the orthographic projection function defined as Wu et al. (2026):

$$\begin{aligned}(u, v) &= \Big(f_x \cdot x + c_x, f_y \cdot y + c_y\Big), \\ \mathbf{J} &= \frac{\partial(u,v)}{\partial(x,y,z)} = \begin{pmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \end{pmatrix}.\end{aligned} \tag{2}$$

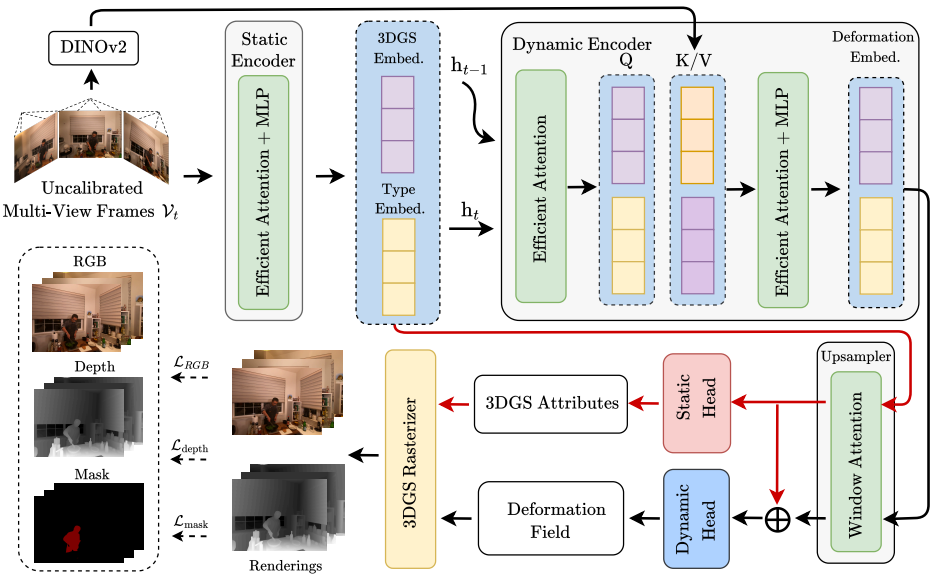

Figure 2: Overview of **FVVSplat**. Given the neighboring multi-view frames $\{I_t^v\}_{v=1}^V$, we first encode them using the Static Encoder to produce canonical 3D Gaussians, and then pass the 3DGS Embeddings to the Dynamic Decoder to predict the deformation field. The predicted dynamic 3D Gaussians are rendered to obtain the reconstructed video and corresponding depths.

The pixel color $\boldsymbol{C}$ is obtained by alpha-blending the projected 2D Gaussians sorted by depth:

$$\boldsymbol{C} = \sum_{i=1}^{N} \boldsymbol{c}_i \boldsymbol{\alpha}_i \prod_{j=1}^{i-1} (1 - \boldsymbol{\alpha}_j) \tag{3}$$

where $\boldsymbol{\alpha}_i$ is the opacity of the i-th projected Gaussian obtained by Eq.(1).

4 Method

Given a time-synchronized uncalibrated, multi-view video stream $\mathcal{S} = \{\mathcal{V}_t\}_{t=1}^T$ with V views, where $\mathcal{V}_t = \{I_t^v\}_{v=1}^V$ denotes the multi-view frames at timestep t. We present FVVSplat, a framework designed to instantly transform uncalibrated video streams into dynamic 3D Gaussian Splatting (3DGS) representations capable of capturing scene dynamics. Figure 2 provides an overview. We first encode the current frame into static 3D Gaussians in a canonical space (Section 4.1), then predict a bidirectional deformation field between Gaussians encoded from current frame and propagated from the previous frame, and finally fuse them into a unified dynamic representation (Section 4.2). An efficient attention (Section 4.3) is integrated into the Static Encoder and Dynamic Decoder to achieve the on-the-fly and high-quality dynamic scene reconstruction (Section 4.4).

4.1 Static 3D Gaussian Encoding

Canonical 3D Space. To support a unified model for in-the-wild videos with unknown and varying intrinsics, we adopt a shared orthographic canonical space Sun et al. (2024b); Shen et al. (2025). This bypasses per-scene calibration: camera motion and perspective effects are absorbed into the Gaussian dynamics and handled by the Dynamic Decoder (Section 4.2). It provides a simplified yet effective foundation for predicting dynamic 3D Gaussians directly from uncalibrated videos.

Structured Static 3D Gaussian Encoding. Given an input frame I, we form an RGB-D image and partition it into 8×8 patches. An Efficient Attention-based Static Encoder produces 3DGS embeddings $\mathbf{h}$ and a lightweight upsampler Liu et al. (2021); Xu et al. (2024) then yields per-2×2-patch Gaussian tokens $\hat{\mathbf{E}}$. Each token $\hat{\mathbf{E}}_i$ is decoded by linear heads into 3DGS parameters: position

offset $\boldsymbol{o}_i \in [\mathbf{a}_i,\ \mathbf{b}_i]$, rotation $\mathbf{R}_i$, scale $\mathbf{S}_i$, opacity $\boldsymbol{\alpha}_i$, and color $\boldsymbol{c}_i$, where $\mathbf{a}_i$ and $\mathbf{b}_i$ are dynamic truncation bounds estimated by the adaptive radius sampling. The 3D position is computed via pixel-aligned prediction: $\boldsymbol{\mu}_i = (u_i + \boldsymbol{o}_{i,0}, v_i + \boldsymbol{o}_{i,1}, g(\boldsymbol{o}_{i,2}))$, where (u_i, v_i) is the pixel coordinate, $\boldsymbol{o}_{i,2}$ denotes the inverse depth Yang et al. (2024b) for better depth estimation near the camera, and $g(z) = 2/(1+z)$ is the depth mapping.

Adaptive Radius Position Sampling. 3D Gaussian Splatting is sensitive to position initialization Kerbl et al. (2023) and prone to local minima Charatan et al. (2024), especially in feed-forward models Charatan et al. (2024); Yi et al. (2024) that predict in a single forward pass without iterative refinement. To mitigate this, existing methods resort to dense or hard sampling (Charatan et al., 2024), sacrificing scene-independent adaptivity and degrading Gaussian prediction accuracy. In this work, we predict a normal distribution with a learnable adaptive radius for each 3D offset $\boldsymbol{o}$. For each Gaussian token $\hat{\mathbf{E}}_i$, let $\boldsymbol{o}_i \in \mathbb{R}^3$ denote its position offset and $\boldsymbol{\mu}_i \in \mathbb{R}^3$ its mean. The offset is modeled as a truncated normal distribution:

$$\boldsymbol{o}_i \sim \mathcal{N}_{[\mathbf{a}_i,\ \mathbf{b}_i]}(\boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i)\,, \quad \boldsymbol{\Sigma}_i = \mathrm{diag}(\exp(\mathbf{s}_i)) \tag{4}$$

where $\mathbf{s}_i \in \mathbb{R}^3$ is the predicted log-variance for normal distribution, and $\exp(s_{i,j})$ gives the variance along dimension j. We set the truncation radius to be proportional to the predicted scale. For token $\hat{\mathbf{E}}_i$, we first aggregate its per-dimension standard deviation $\exp(s_{i,j}/2)$ into a scalar scale:

$$\xi_i = \frac{1}{d}\sum_{j=1}^{d} \exp(s_{i,j}/2), \tag{5}$$

where $s_{i,j}$ is the j-th component of $\mathbf{s}_i$ and $d = 3$. We then map ξ_i to a per-token exploration radius:

$$r_i = \mathrm{clamp}(k \cdot \xi_i,\ r_{\min},\ r_{\max})\,, \tag{6}$$

where k controls the ratio between the exploration range and the model's uncertainty, $\mathrm{clamp}(\cdot)$ clips its input to $[r_{\min}, r_{\max}]$, $r_{\min}$ prevents the radius from collapsing to zero, and $r_{\max}$ prevents it from exploding. The dynamic truncation bounds are then constructed as $\mathbf{a}_i = -r_i \cdot \mathbf{1}$ and $\mathbf{b}_i = r_i \cdot \mathbf{1}$. Adaptive radius is a truncation radius for truncated Gaussian position sampling that is determined per token and dynamically from the token's predicted log-variance, making the exploration range proportional to the model's positional uncertainty.

4.2 Dynamic Deformation Prediction

Dynamic scene reconstruction involves large non-rigid motions and topological changes. We address this with bidirectional deformation design that provides reliable cross-frame associations and naturally handles newly appearing and disappearing content, maintaining long-term temporal coherence. Per-Gaussian deformation is parameterized by a 3D velocity $\mathbf{v}_i \in [\mathbf{a}_i, \mathbf{b}_i]$ and an opacity coefficient $\boldsymbol{\gamma}$ that controls visibility over time. Likewise, $\mathbf{v}$ is sampled by the same probabilistic position sampling mechanism.

Dynamic Deformation Decoding. Given two consecutive frames I_{t-1} and I_t, we obtain 3D Gaussian Splatting (3DGS) embeddings $\mathbf{h}_{t-1}$ and $\mathbf{h}_t$ from a frozen static encoder. An efficient attention-based dynamic decoder, conditioned on DINOv2 features Oquab et al. (2023), produces deformation embeddings. The same upsampler then yields per-2×2-patch deformation tokens $\mathbf{d}$, which are concatenated with the paired static tokens $\mathbf{h}$ and decoded by an MLP head into per-Gaussian velocity and opacity coefficients.

Bidirectional Deformation Field. For forward deformation, static tokens $\mathbf{h}_{t-1}$ and deformation tokens $\mathbf{d}_{t-1}$ are adopted to predict the corresponding velocity and opacity coefficients:

$$\begin{aligned} (\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\Sigma}, \boldsymbol{\gamma}_{t-1}) &= \textsc{Head}_{\mathrm{MLP}}(\mathbf{h}_{t-1}, \mathbf{d}_{t-1}) \\ \boldsymbol{\mu}(t) &= \boldsymbol{\mu}(t-1) + \mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}_{[\mathbf{a},\mathbf{b}]}(\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\Sigma}) \\ \boldsymbol{\alpha}(t) &= \boldsymbol{\alpha}(t-1) \cdot \sigma\big(\boldsymbol{\gamma}_{t-1}\big), \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\boldsymbol{\gamma}_{t-1}$ controls transition rate. Therefore, the forwardly deformed Gaussians are denoted as: $\mathcal{G}^+(t) = \{(\boldsymbol{\mu}(t), \boldsymbol{\alpha}(t), \mathbf{S}, \mathbf{R}, \boldsymbol{c})\}$. Similarly, the backwardly deformed Gaussians $\mathcal{G}^-(t)$ are predicted from static tokens $\mathbf{h}_t$ and deformation tokens $\mathbf{d}_t$. Then, we update static Gaussians $\mathcal{G}(t)$ with $\mathcal{G}^+(t)$, setting the active deformation of

matched 3DGS primitives to the new forward field, and fuse $\mathcal{G}_t^-(t)$ to obtain the final Gaussians: $\tilde{\mathcal{G}}(t) = \mathcal{C}\big(\mathcal{G}(t), \mathcal{G}^+(t)\big) \cup \mathcal{G}_t^-(t)$, where $\mathcal{C}(\cdot)$ is concatenation operation. This bidirectional deformation fields handle the move and fade Gaussians over time, recovering cross-time scene evolution. The time-dependent opacity implicitly fuses the forward and backward Gaussians, propagating persistent Gaussians while handling emerging and vanishing ones, thereby maintaining temporal coherence without hard assignments or iterative fusion.

4.3 Efficient Attention

Existing feed-forward models typically rely on standard self-attention for Gaussian encoding and dynamic prediction. While this yields promising reconstruction quality, it incurs substantial computational cost, especially for high-resolution videos, due to the quadratic complexity in the sequence length. Let $Q, K, V \in \mathbb{R}^{B\times H\times L\times D}$ denote the query, key, and value, where B is the batch size, H the number of heads, L the sequence length, and D the per-head dimension. The standard self-attention is defined as

$$\text{Self-Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{D}}\right)V. \tag{8}$$

To maintain high efficiency across various video resolutions, we introduce an efficient attention that replaces the softmax with a non-negative kernel $\phi(x) = \text{ELU}(x) + 1$:

$$\text{E-Attn}(Q, K, V) = \frac{\phi(Q)\big(\phi(K)^\top V\big)}{\phi(Q)\big(\phi(K)^\top \mathbf{1}\big)}. \tag{9}$$

Here L denotes the input sequence length, i.e., the number of tokens. Since $L \gg D$ and D is fixed, this formulation avoids the explicit $L \times L$ attention matrix and reduces the complexity from $O(L^2D)$ to $O(LD^2)$.

4.4 Training Strategy

To achieve the jointly optimization of static 3D Gaussian encoding and dynamic deformation prediction, we adopt a two-stage training protocol.

Stage 1: Static 3DGS Encoder Training. The static encoder aims to reconstruct static 3DGS from a single input frame I_t and pseudo-depth D_t. It produces 3DGS primitives that are used to render RGB $\hat{I}_t$ and depth $\hat{D}_t$. The training loss combines photometric and depth supervision:

$$\mathcal{L}_{\text{static}} = \mathcal{L}_{\text{recon}}(\hat{I}_t, I_t) + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}(\hat{D}_t, D_t), \tag{10}$$

where $\mathcal{L}_{\text{recon}}$ is the sum of L_2 loss (RGB space) and LPIPS loss Zhang et al. (2018), and $\mathcal{L}_{\text{depth}}$ is a scale- and shift-invariant depth loss Ranftl et al. (2020):

$$\mathcal{L}_{\text{depth}} = \mathbb{E}\|\tau(\hat{D}_t) - \tau(D_t)\|, \quad \text{where} \quad \tau(\mathbf{x}) = \frac{\mathbf{x} - \text{median}(\mathbf{x})}{\mathbb{E}\|\mathbf{x} - \text{median}(\mathbf{x})\|}. \tag{11}$$

Stage 2: Dynamic Deformation Decoder Training. To improve the dynamic decoder performance and maintain the temporal consistency for reconstructed videos, we introduce the bidirectional deformation at neighboring timesteps to supervise their renderings for training Dynamic Deformation Decoder. Given two neighboring frames I_{t-1} and I_t, we first encode them to obtain static 3DGS $\mathcal{G}_{t-1}$ and $\mathcal{G}_t$. Then we use the dynamic decoder to predict the their bidirectional deformation fields and final Gaussians. The renderings, depths, and masks of these two neighboring frames are adopted into training objective is to reconstruct :

$$\mathcal{L}_{\text{dynamic}} = \mathbb{E}_t\left[\mathcal{L}_{\text{recon}}(\hat{I}_t, I_t) + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}(\hat{D}_t, D_t) + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}(\hat{I}_t \odot M_t, I_t \odot M_t)\right], \tag{12}$$

where $\mathcal{L}_{\text{mask}}$ is an auxiliary reconstruction loss that encourages the model to focus on moving foreground regions using a binary segmentation mask M_j from SAM2 Ravi et al. (2025).

5 Experiments

5.1 Experimental Settings

Table 1: Quantitative results of Optimization-based (Optim.) and Feed-Forward (FF) FVV reconstruction methods on the N3DV (1352 × 1014), and the MeetRoom (1280 × 720) datasets. **Bold** and underlined values denote the best and second-best results, respectively.

	Method	N3DV				MeetRoom			
		PSNR(dB)↑	SSIM↑	LPIPS↓	Time(sec.)↓	PSNR(dB)↑	SSIM↑	LPIPS↓	Time(sec.)↓
Optim.	StreamRF	30.61	0.930	-	15.0	26.71	0.913	-	10.2
	3DGStream	31.67	0.948	0.130	8.5	30.79	0.921	0.188	7.2
	HiCoM	31.17	0.953	0.130	6.7	29.61	-	0.182	4.1
	4DGC	31.58	0.943	-	50	28.08	0.922	-	24.5
	GIFStream	30.76	0.937	0.147	-	24.37	0.881	0.239	-
	iFVC	32.32	0.949	0.132	13.2	<u>32.38</u>	<u>0.960</u>	0.168	8.4
	DLGStream	31.93	0.945	0.141	-	27.28	0.917	0.206	-
	StreamSTGS	32.30	0.943	0.147	67	27.41	0.920	0.210	29
	QUEEN-s	31.89	0.945	0.139	4.7	31.95	<u>0.960</u>	<u>0.159</u>	2.9
	QUEEN-l	32.19	0.946	0.136	7.9	32.21	**0.962**	**0.152**	3.5
	ReCon-GS	<u>32.66</u>	<u>0.957</u>	0.123	6.4	30.84	0.955	<u>0.163</u>	3.9
	MoRGS	32.53	0.950	<u>0.118</u>	<u>4.0</u>	31.79	0.957	**0.152**	<u>2.3</u>
FF	**Ours**	**34.82**	**0.965**	**0.112**	**0.26**	**35.06**	<u>0.960</u>	**0.152**	**0.17**

Training Datasets. To train our feed-forward model, we collect several multi-view dynamic scene datasets: WideRange4D Yang et al. (2025), N3DV Li et al. (2022b), Meetroom Li et al. (2022a), VRU Wu et al. (2025), and MultiEgo Li et al. (2025). Following prior works Sun et al. (2024a); Fu et al. (2026), we downsample N3DV by 2× for both training and testing, keeping the original resolution for other datasets. We adopt SAM2 Ravi et al. (2025) to obtain segmentation masks for all datasets, use the official train/test splits, and train FVVSplat on the training sets.

Table 2: Quantitative results on the VRU (1920 × 1080) dataset. * denotes results reproduced by released code. **Bold** and underlined values denote the best and second-best results, respectively.

	Method	VRU			
		PSNR(dB)↑	SSIM↑	LPIPS↓	Time(sec.)↓
Optim.	3DGStream	26.20	0.876	0.247	32.1
	4DGC	26.61	0.889	0.322	125.1
	QUEEN-l*	26.87	0.886	0.220	22.4
	ReCon-GS*	26.02	0.877	0.216	24.8
	MoRGS*	<u>26.92</u>	<u>0.890</u>	<u>0.214</u>	<u>20.4</u>
FF	**FVVSplat**	**31.87**	**0.926**	**0.193**	**0.28**

Implementation Details. FVVSplat is trained on 4 NVIDIA A800 GPUs with FlashAttention-2 Dao (2023), gradient checkpointing Chen et al. (2016), and BF16 mixed-precision training for efficiency. In adaptive radius sampling, we set $k = 1$, $r_{\min} = 0.6$, $r_{\max} = 1.2$. We apply image-level augmentations following EDM Karras et al. (2022) and optimize using AdamW Loshchilov & Hutter (2017) with gradient clipping set to 1.0. For Stage 1, we use a batch size of 4, a learning rate of 5×10^{-4} with 20K linear warm-up iterations, and weight decay of 0.05. For Stage 2, we use a batch size of 3, a learning rate of 1×10^{-4} with 100K linear warm-up iterations, and weight decay of 0.05. Additional training details are provided in the supplementary.

Evaluation Settings. Following prior optimization-based works Sun et al. (2024a); Girish et al. (2024); Fu et al. (2026), we adopt averaged per-frame peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) Wang et al. (2004), and LPIPS Zhang et al. (2018) across all videos at the standard benchmark resolutions to evaluate reconstruction quality, ensuring fair comparison Charatan et al. (2024); Jain et al. (2024). We report averaged per-frame training time for optimization-based methods and averaged per-frame inference time for feed-forward method.

5.2 Comparisons of State-of-the-art Methods

Quantitative comparisons. To evaluate our method on standard benchmarks, we collect some streamable FVV reconstruction methods Wang et al. (2023); Sun et al. (2024a); Wu et al. (2024); Yan et al. (2025); Hu et al. (2025a); Gao et al. (2024); Fu et al. (2026); Girish et al. (2024). Notably, we reproduced QUEEN-l, ReCon-GS Fu et al. (2026), and MoRGS Lee et al. (2026) on the VRU dataset, using their official implementations to ensure a fair comparison.

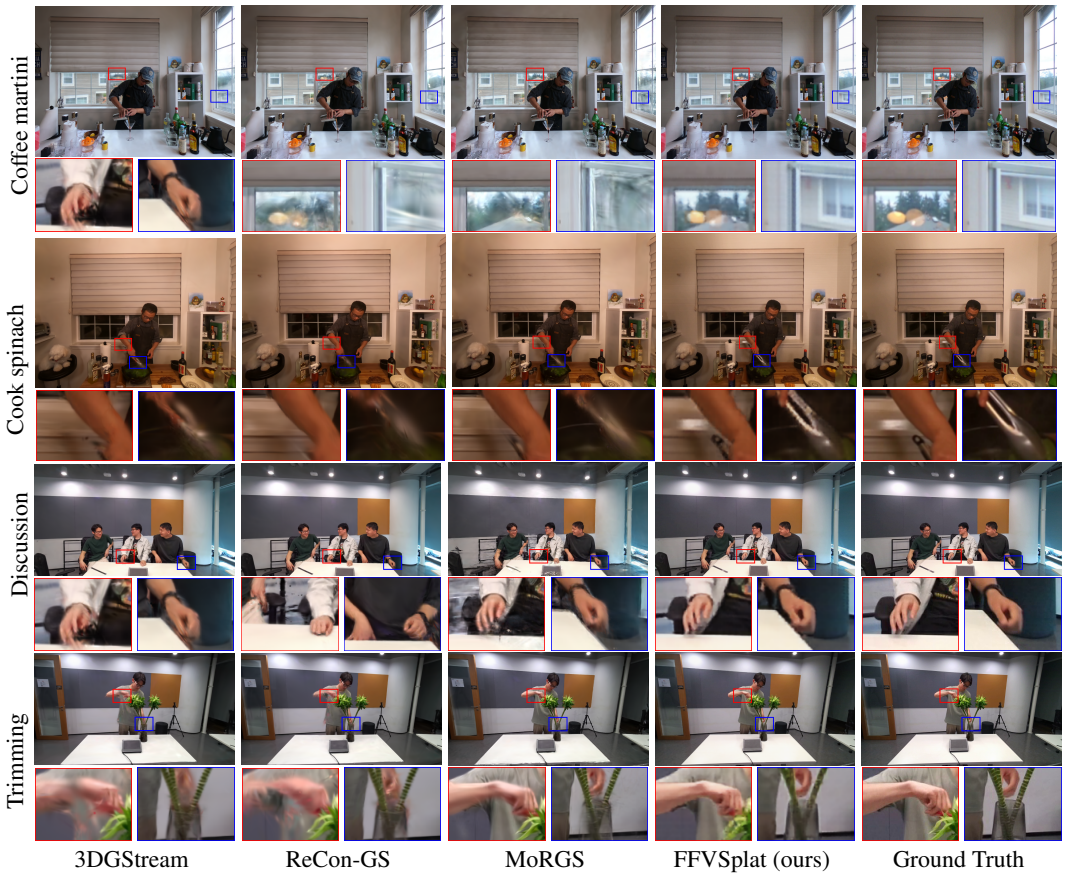

Figure 3: Visual comparisons on the N3DV (*coffee_martini*, *cook_spinach*), and Meetroom (*discussion*, *trimming*) datasets. FVVSplat achieves better visual results compared with other optimization-based methods.



Figure 4: Visual comparisons on the VRU dataset. FVVSplat achieves better visual advantages compared with other optimization-based methods on challenging large motion scenarios.

We provide the quantitative results of existing optimzation-based methods and our FVVSplat for these three datasets in Tabs. 1–2. From these results can be obversed that FVVSplat consistently surpasses all optimization-based methods by a large margin. Compared with the state-of-the-art MoRGS, it achieves PSNR gains of 2.16dB, 3.27dB, and 4.95dB, together with speedups of 15×, 13.5×, and 72× on the N3DV, Meetroom, and VRU datasets, respectively. These results demonstrate the significant advantage of FVVSplat for streamable FVV reconstruction due to the introduce of feed-forward model and effective designs.

Table 3: Ablation study results of our proposed designs of FVVSplat on the N3DV dataset.

Models	PSNR(dB)↑	SSIM↑	LPIPS↓	Time(sec.)↓
Baseline	28.72	0.821	0.436	0.19
w/o ARS	33.49	0.947	0.264	0.24
w/o Bi-Def	33.10	0.938	0.271	0.21
w/o EA	34.05	0.960	0.116	0.43
Ours (full)	34.82	0.965	0.112	0.26

Table 4: Comparisons of two attentions with various resolutions on *flame_salmon*.

Resolution	Self-Attn.		Efficient Attn.	
	PSNR↑	Time↓	PSNR↑	Time↓
480P	33.62	0.020	33.85	0.076
720P	33.43	0.101	33.56	0.170
960P	33.26	0.179	33.39	0.226
1080P	32.94	0.510	32.91	0.382

Qualitative comparisons. We present the qualitative comparisons with three representative streamable FVV reconstruction methods, i.e., 3DGStream Sun et al. (2024a), Recon-GS Fu et al. (2026), and MoRGS Lee et al. (2026) on the scenes of the N3DV, MeetRoom, and VRU datasets in Fig. 3 and Fig. 4, respectively. Qualitative comparisons in these two figures show that FVVSplat consistently outperforms 3DGStream, Recon-GS, and MoRGS in preserving fine-grained structures, including the window and lighting in the *coffee_martini* scene, the background texture in the *discussion* scene, and the numerals and basketball net in the *DG* and *GZ* scenes. These results demonstrate that FVVSplat can accurately reconstruct dynamic scenes with high fidelity to intricate details, while simultaneously achieving a highly compact representation.

5.3 Ablation Studies

We conduct ablation studies to evaluate the contribution of each key component in FVVSplat. Quantitative and qualitative results on the N3DV dataset are presented in Table 3.

Effectiveness of adaptive radius sampling. We first construct a baseline that removes our three designs—ARS, Bi-Def, and EA. Although it achieves reasonable quality, its performance is limited by the absence of explicit dynamic modeling and efficient attention. Without ARS, the model still reaches 33.49 dB PSNR, but the rendering time increases from 0.19 s to 0.24 s, indicating that ARS mainly contributes to the quality–efficiency balance.

Effectiveness of bidirectional deformation. Removing the Bidirectional Deformation leads to a further drop to 33.10 dB PSNR, while LPIPS rises to 0.271. This indicates that Bidirectional Deformation is essential for capturing complex motion and maintaining perceptual fidelity: without cross-time deformation, features from different frames are not properly warped into a shared space, so the model fails to represent temporally varying geometry.

Effectiveness of efficient attention. The most revealing variant removes Efficient Attention. It matches the full model in quality but its rendering time jumps to 0.43 s, about 1.65× slower. Removing EA reverts to standard self-attention, whose quadratic cost grows rapidly with sequence length: at 1080P, self-attention takes 0.510 s versus 0.382 s for efficient attention, with the gap widening at higher resolutions. This confirms that EA is not a quality-improving module but a scalability mechanism, trading a small amount of accuracy for a substantial gain in efficiency.

Our full model achieves the best reconstruction performance, confirming that the proposed designs are complementary components that jointly enable high-quality reconstruction at a practical speed.

6 Conclusion

In this work, we introduce **FVVSplat**, a Feed-Forward Streamable **F**ree-**V**iewpoint **V**ideo Reconstruction framework that instantly predicts dynamic 3D Gaussian **Splat**ting representations from uncalibrated multi-view video streams. An adaptive radius sampling mechanism is designed for accurate Gaussian prediction from uncalibrated inputs, and a bidirectional deformation is proposed that ensures reliable, persistent Gaussians across frames and mitigates long-term error accumulation. Furthermore, an efficient attention is integrated into both static prediction and dynamic deformation with linear complexity, substantially reducing computational cost while maintaining reconstruction quality across video resolutions. Extensive experiments on multiply streamable FVV reconstruction benchmarks demonstrate that FVVSplat significantly improves both reconstruction quality and inference time over existing optimization-based methods.

REFERENCES

Matteo Balice, Yanik Kunzi, Chenyangguang Zhang, Matteo Matteucci, Marc Pollefeys, and Sungwhan Hong. No pose, no problem in 4d: Feed-forward dynamic gaussians from unposed multi-view videos. *arXiv preprint arXiv:2605.22190*, 2026.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Jiacong Chen, Qingyu Mao, Youneng Bao, Xiandong Meng, Fanyang Meng, Ronggang Wang, and Yongsheng Liang. Motion matters: Compact gaussian streaming for free-viewpoint video reconstruction. *Advances in Neural Information Processing Systems*, 38:120385–120409, 2026a.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, et al. Dggt: Feedforward 4d reconstruction of dynamic driving scenes using unposed images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1265–1276, 2026b.

Yihang Chen, Qianyi Wu, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Hac: Hash-grid assisted context for 3d gaussian splatting compression. In *European Conference on Computer Vision*, pp. 422–438. Springer, 2024.

Yihang Chen, Qianyi Wu, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Hac++: Towards 100x compression of 3d gaussian splatting. *arXiv preprint arXiv:2501.12255*, 2025.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8628–8638, 2021.

Changwoon Choi, Jeongjun Kim, Geonho Cha, Minkwan Kim, Dongyoon Wee, and Young Min Kim. Humans as a calibration pattern: Dynamic 3d scene reconstruction from unsynchronized and uncalibrated videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6598–6608. IEEE, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Jiaye Fu, Qiankun Gao, Chengxiang Wen, Yanmin Wu, Siwei Ma, Jiaqi Zhang, and Jian Zhang. Recon-gs: Continuum-preserved gaussian streaming for fast and compact reconstruction of dynamic scenes. *Advances in Neural Information Processing Systems*, 38:134751–134778, 2026.

Qiankun Gao, Jiarui Meng, Chengxiang Wen, Jie Chen, and Jian Zhang. Hicom: Hierarchical coherent motion for dynamic streamable scenes with 3d gaussian splatting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Qiankun Gao, Jiarui Meng, Chengxiang Wen, Jie Chen, and Jian Zhang. Hicom: Hierarchical coherent motion for dynamic streamable scenes with 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:80609–80633, 2024.

Sharath Girish, Tianye Li, Amrita Mazumdar, Abhinav Shrivastava, Shalini De Mello, et al. Queen: Quantized efficient encoding of dynamic gaussians for streaming free-viewpoint videos. *Advances in Neural Information Processing Systems*, 37:43435–43467, 2024.

Qiang Hu, Zihan Zheng, Houqiang Zhong, Sihua Fu, Li Song, Xiaoyun Zhang, Guangtao Zhai, and Yanfeng Wang. 4dgc: Rate-aware 4d gaussian compression for efficient streamable free-viewpoint video. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 875–885. IEEE, 2025a.

Yingdong Hu, Yisheng He, Jinnan Chen, Weihao Yuan, Kejie Qiu, Zehong Lin, Siyu Zhu, Zilong Dong, Steven Hoi, and Jun Zhang. Forge4d: Feed-forward 4d human reconstruction and interpolation from uncalibrated sparse-view videos. *arXiv preprint arXiv:2509.24209*, 2025b.

Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7341–7351, 2024.

Yiming Jiang, Hanzhang Tu, Wenfeng Song, Siyou Lin, Liang An, Shuai Li, Aimin Hao, and Yebin Liu. Hireff: High-resolution feedforward human reconstruction from uncalibrated sparse-view video. *arXiv preprint arXiv:2606.29333*, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

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.

Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic gaussian splatting. *Advances in Neural Information Processing Systems*, 37:5384–5409, 2024.

Wonjoon Lee, Sungmin Woo, Donghyeong Kim, Jungho Lee, Sangheon Park, and Sangyoun Lee. Morgs: Efficient per-gaussian motion reasoning for streamable dynamic 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41044–41053, 2026.

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

Bate Li, Houqiang Zhong, Zhengxue Cheng, Qiang Hu, Qiang Wang, Li Song, and Wenjun Zhang. Multiego: A multi-view egocentric video dataset for 4d scene reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12882–12889, 2025.

Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming radiance fields for 3d video synthesis. *Advances in Neural Information Processing Systems*, 35:13485–13498, 2022a.

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*, pp. 5521–5531, 2022b.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 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.

Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, et al. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. *arXiv preprint arXiv:2412.03526*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8089–8099, 2024.

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021a.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021b.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

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

Qiuhong Shen, Xuanyu Yi, Mingbao Lin, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. Seeing world dynamics in a nutshell. *arXiv preprint arXiv:2502.03465*, 2025.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 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, 2024a.

Yang-Tian Sun, Yihua Huang, Lin Ma, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Splatter a video: Video gaussian representation for versatile processing. *Advances in Neural Information Processing Systems*, 37:50401–50425, 2024b.

Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19706–19716, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025.

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*, pp. 20697–20709, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jiahao Wu, Rui Peng, Zhiyan Wang, Lu Xiao, Luyang Tang, Jinbo Yan, Kaiqiang Xiong, and Ronggang Wang. Swift4d: Adaptive divide-and-conquer gaussian splatting for compact and efficient reconstruction of dynamic scene. *arXiv preprint arXiv:2503.12307*, 2025.

Minye Wu, Zehao Wang, Georgios Kouros, and Tinne Tuytelaars. Tetrirf: Temporal tri-plane radiance fields for efficient free-viewpoint video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6487–6496, 2024.

Zike Wu, Qi Yan, Xuanyu Yi, Lele Wang, and Renjie Liao. Streamsplat: Towards online dynamic 3d reconstruction from uncalibrated video streams. In *International Conference on Learning Representations*, volume 2026, pp. 123114–123137, 2026.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024.

Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv. 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. *arXiv preprint arXiv:2506.08015*, 2025.

Jinbo Yan, Rui Peng, Zhiyan Wang, Luyang Tang, Jiayu Yang, Jie Liang, Jiahao Wu, and Ronggang Wang. Instant gaussian stream: Fast and generalizable streaming of dynamic scene reconstruction via gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16520–16531, 2025.

Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Apoorva Sharma, Maximilian Igl, Peter Karkus, Danfei Xu, et al. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. *arXiv preprint arXiv:2501.00602*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

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.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20331–20341, 2024c.

Xuanyu Yi, Zike Wu, Qiuhong Shen, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, Shuicheng Yan, Xinchao Wang, and Hanwang Zhang. Mvgamba: Unify 3d content generation as state space sequence modeling. *arXiv preprint arXiv:2406.06367*, 2024.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Ruijie Zhu, Yanzhe Liang, Hanzhi Chang, Jiacheng Deng, Jiahao Lu, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:101790–101817, 2024.

Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS'01.*, pp. 29–538. IEEE, 2001.