

Gaussians on their Way: Wasserstein-Constrained 4D Gaussian Splatting with State-Space Modeling

Junli Deng^{1*}

Communication University of China
Beijing, China
dengjunliok@cuc.edu.cn

Yihao Luo^{2†*}

Imperial College London
London, UK
y.luo23@imperial.ac.uk

Abstract

Dynamic scene rendering has taken a leap forward with the rise of 4D Gaussian Splatting, but there's still one elusive challenge: how to make 3D Gaussians move through time as naturally as they would in the real world, all while keeping the motion smooth and consistent. In this paper, we unveil a fresh approach that blends state-space modeling with Wasserstein geometry, paving the way for a more fluid and coherent representation of dynamic scenes. We introduce a State Consistency Filter that merges prior predictions with the current observations, enabling Gaussians to stay true to their way over time. We also employ Wasserstein distance regularization to ensure smooth, consistent updates of Gaussian parameters, reducing motion artifacts. Lastly, we leverage Wasserstein geometry to capture both translational motion and shape deformations, creating a more physically plausible model for dynamic scenes. Our approach guides Gaussians along their natural way in the Wasserstein space, achieving smoother, more realistic motion and stronger temporal coherence. Experimental results show significant improvements in rendering quality and efficiency, outperforming current state-of-the-art techniques.

1. Introduction

Dynamic scene rendering is a fundamental problem in computer vision, with widespread applications in virtual reality, augmented reality, robotics, and film production. Accurately capturing and rendering dynamic scenes with complex motions and deformations remains a challenging task due to the high computational demands and the intricate nature of dynamic environments [1, 2].

Neural representations have advanced dynamic scene modeling, with Neural Radiance Fields [3] revolutionizing

novel view synthesis through neural network-parameterized continuous functions. Extensions to dynamic scenes [4–9] have been proposed, but they often suffer from high computational costs and limited real-time capabilities. 4D Gaussian Splatting [10–15] enables real-time dynamic scene rendering using dynamic 3D Gaussians and differentiable splatting [16]. However, accurately modeling scene dynamics remains challenging due to limitations in estimating precise Gaussian transformations [17, 18].

In this paper, we draw inspiration from control theory [19] and propose a novel approach that integrates a State Consistency Filter into the 4D Gaussian Splatting framework. By modeling the deformation of each Gaussian as a state in a dynamic system, we estimate Gaussian transformations by merging prior predictions and observed data, accounting for uncertainties in both.

To ensure smooth and consistent parameter updates, we incorporate Wasserstein distance [20, 21] as a key metric between Gaussian distributions. This metric effectively quantifies the optimal transformation cost between distributions, considering both positional and shape differences. By using Wasserstein distance regularization, we preserve the underlying Gaussian structure while enhancing temporal consistency and reducing rendering artifacts.

Additionally, we introduce Wasserstein geometry [21, 22] to model Gaussian dynamics, capturing both translational motion and shape deformations in a unified framework. This approach enables more physically plausible evolution of Gaussians, leading to improved motion trajectories and rendering quality. Our main contributions are:

- We propose a novel framework that integrates a State Consistency Filter into 4D Gaussian Splatting, enabling more accurate Gaussian motion estimation by optimally merging prior predictions and observed data.
- We introduce Wasserstein distance regularization, which smooths Gaussian parameter updates over time, ensuring temporal consistency and reducing artifacts.
- We leverage Wasserstein geometry to model both translational motion and shape deformations of Gaussians, en-

*Equal contribution.

†Corresponding author.

hancing the physical plausibility of Gaussian dynamics.

2. Related Work

2.1. Dynamic Novel View Synthesis

Synthesizing new views of dynamic scenes from multi-time 2D images remains challenging. Recent works have extended Neural Radiance Fields (NeRF) to handle dynamic scenes by learning spatio-temporal mappings [23–32]. While classical approaches using the plenoptic function [33], image-based rendering [34, 35], or explicit geometry [36, 37] face memory limitations, implicit representations [4, 26, 38–40] have shown promise through deformation fields [4, 7, 25] and specialized priors [39, 41–45].

Temporally extended 3D Gaussian Splatting has also been explored for dynamic view synthesis. Luiten *et al.* [15] assign parameters to 3D Gaussians at each timestamp and use regularization to enforce rigidity. Yang *et al.* [46] model density changes over time using Gaussian probability to represent dynamic scenes. However, they require many primitives to capture complex temporal changes. Other works [47–52] leverage Multi-Layer Perceptrons (MLPs) to represent temporal changes. In 4D Gaussian Splatting, the motion of Gaussians should adhere to physical laws. By incorporating control theory, we can predict the motion of Gaussians more accurately.

2.2. Dynamic Scene State Estimation

Recent advances in dynamic scene reconstruction have explored various approaches for tracking and modeling temporal changes. In object tracking, methods like SORT [53] and SLAM systems [54] have provided robust frameworks for state estimation. The integration of learning-based approaches has enhanced these methods, particularly in handling complex scenarios with limited observations [55, 56].

Recent works have explored optimal transport and probabilistic approaches for dynamic scenes. Gaussian-Cube [57] models scenes with probabilistic distributions for robust deformation handling, while Shape of Motion [58] leverages geometric transformations for temporal coherence. KFD-NeRF [59] applies Kalman filtering to NeRF but is constrained by its discrete point representation, while optimal transport has shown promise in improving dynamic NeRF convergence [60]. Our work differs by leveraging Gaussian Splatting, which models dynamic elements as full probability distributions in the Wasserstein space. This approach captures the geometric nature of distribution transformations, which is especially beneficial for scenes with significant deformations or rapid motions.

3. Method

Our framework integrates three key components for dynamic scene rendering (Figure 1). First, a state-space updat-

ing mechanism with neural Gaussian deformation estimates motion patterns. Second, Wasserstein distance regularization ensures smooth parameter updates. Third, we model Gaussian dynamics under Wasserstein geometry for accurate motion prediction with intrinsic physical plausibility. We detail each component in the following sections.

3.1. Filter for State Consistency

3.1.1. Observer: Neural Gaussian Deformation Field

3D Gaussian Splatting represents static scenes as a collection of 3D Gaussians, each parameterized by its mean position μ and covariance matrix Σ . The covariance matrix is typically decomposed into rotation R and scaling S matrices [16]:

$$\Sigma = RSS^T R^T. \quad (1)$$

This decomposition allows for efficient modeling of oriented Gaussian distributions in 3D space. 4D Gaussian Splatting extends this representation to dynamic scenes by allowing these Gaussian parameters to vary over time $\gamma(t)$, enabling the modeling of moving and deforming objects.

Building upon this foundation, we introduce a more principled approach to modeling temporal variations through a neural deformation field. Given a canonical Gaussian distribution $\mathcal{N}^c = \mathcal{N}(\mu^c, \Sigma^c)$ and a time parameter t , our neural deformation field predicts the observed Gaussian distribution $\mathcal{N}_t^{Ob}(\mu_t^{Ob}, \Sigma_t^{Ob})$ at time t :

$$\mathcal{N}_t^{Ob} = f_\theta(\mathcal{N}^c, t), \quad (2)$$

where θ represents the learnable parameters of the neural network f_θ implemented as a Multi-Layer Perceptron (MLP). f_θ takes the concatenation of the canonical Gaussian parameters and the positional time encoding [3] as input and outputs the transformation parameters that map the canonical Gaussian to the observed state:

$$\begin{aligned} \mu_t^{Ob} &= \mu^c + \Delta\mu_t, \\ \Sigma_t^{Ob} &= \Sigma^c + \Delta\Sigma_t, \end{aligned} \quad (3)$$

where $\Delta\mu_t$ is the translation offset and $\Delta\Sigma_t$ is the deformation of the covariance matrix.

Using deformed Gaussian distributions \mathcal{N}_t^{Ob} predicted by the neural deformation field to represent dynamic scenes is a common practice in 4D Gaussian Splatting frameworks [10–15]. However, these methods often suffer from flickering artifacts due to abrupt changes in Gaussian parameters between frames. To address this issue, we use the above deformed Gaussian distributions as observations in a state consistency filter, which merges the predicted states with the observed data to obtain the final Gaussian parameters for rendering.

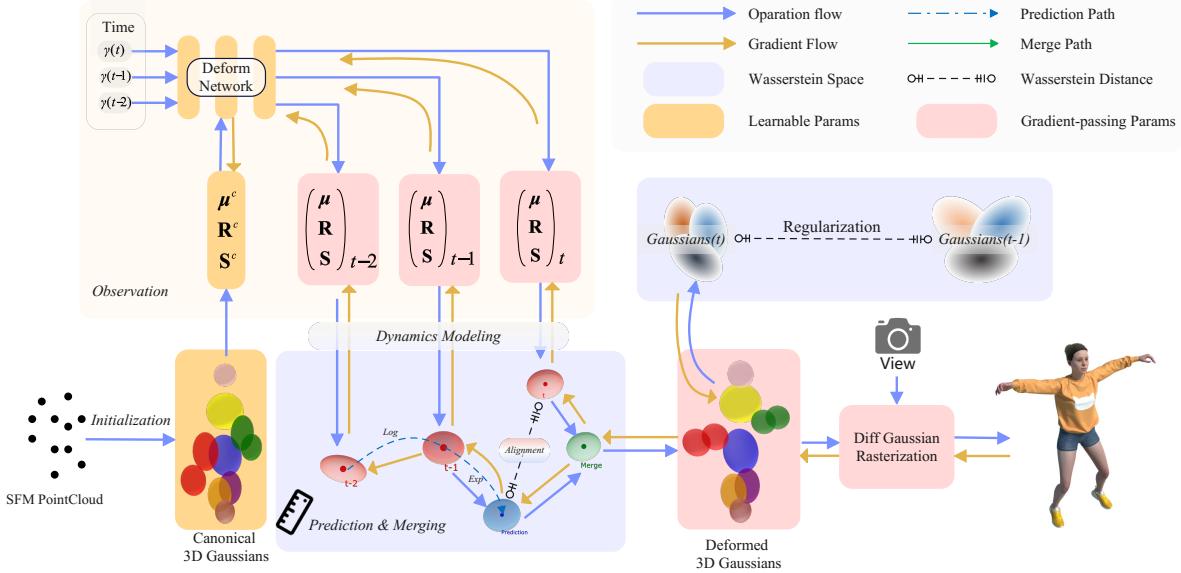


Figure 1. Overview of our proposed method. Starting from a Structure-from-Motion (SFM) point cloud, we initialize canonical 3D Gaussians including position μ^c , rotation R^c , and scale S^c parameters. The deform network predicts these parameters (μ, R, S) at different timestamps $\gamma(t)$. In the Wasserstein space, our state-updating mechanism merges predictions with observations, while ensuring temporal coherence between frames by regularization. The merged Gaussians are then rendered via differentiable rasterization.

3.1.2. Predictor: Time-Independent Linear Dynamics

Traditional Kalman Filters [61] model the state evolution as a linear dynamical system, where the state at time t is a linear transformation of the state at time $t - 1$ combined the control input at time t . The distribution of the state is updated based on the observed data and the predicted state. In our case, we directly model the Gaussian distributions as states (no distribution of states is considered) with the mean and covariance as the state variables. The Euclidean state transition is given by

$$\begin{aligned} \mathcal{N}_{t+1}^P &= \mathcal{N}_t + \mathbf{v}_t \Delta t, \\ \mathbf{v}_t^P &= \mathcal{N}_t - \mathcal{N}_{t-1}, \\ \mathbf{v}_t^P &= \mathbf{v}_{t-1}, \end{aligned} \quad (4)$$

where \mathcal{N}_{t+1}^P is the predicted Gaussian distribution at time $t + 1$, \mathcal{N}_t is the Gaussian distribution at time t , \mathbf{v}_t is the velocity of the Gaussian at time t , and $\Delta t = 1$ is the time step. In Euclidean metric, the velocity \mathbf{v}_t can be decomposed into Euclidean difference of means and covariances, i.e.,

$$\mathcal{N}_{t+1} - \mathcal{N}_t = \mathbf{v}_t = (\mu_t - \mu_{t-1}, \Sigma_t - \Sigma_{t-1}).$$

Conventionally, The predicted velocity \mathbf{v}_t^P is computed as the Euclidean difference between the Gaussian distributions at time t and $t - 1$. Similarly, the first equation in (4) only considers the first-order linear dynamic in Euclidean space. In Section 3.3, we will introduce the Wasserstein

dynamic of Gaussian distributions to replace the Euclidean one for a better depiction of 4D Gaussian splitting. Abstractly, Wasserstein difference is defined as

$$\begin{aligned} \mathbf{v}_t &= -\log_{\mathcal{N}_t}(\mathcal{N}_{t-1}), \\ \mathcal{N}_{t+1} &= \exp_{\mathcal{N}_t}(\mathbf{v}_t), \end{aligned} \quad (5)$$

where the Exponential \exp maps a tangent (velocity) vector to an endpoint Gaussian, and Logarithm \log does the inverse side, assigning the endpoint Gaussian to a tangent vector. Exponential and Logarithm will be determined by the Riemannian metric endowed on the manifold of all Gaussian distributions.

Notice that both the predicted Gaussian distribution and velocity contain components for position and covariance. In the above model, we assume that the acceleration of the Gaussian distribution vanishes and the velocity remains constant over time for smoothness. The dynamics of the Gaussian distributions are modeled as a naive linear system, which is an oversimplified model and far from the real-world dynamics, but provides higher robustness for 4D Gaussian Splatting. Subsequently, we introduce a Kalman-like state updating mechanism to refine the Gaussian distributions based on the observed data and the predicted state.

3.1.3. Merging: Kalman-like State Updating

The Kalman Filter [61] is a recursive algorithm that estimates the state of a linear dynamical system from a series of noisy observations. It combines prior predictions with new

measurements to produce optimal state estimates, accounting for uncertainties in both the process and the observations. In our context, our prediction and observation are the Gaussian distributions themselves. The counterbalancing of prior predictions and new observations allows for robust tracking of the Gaussian states over time, enabling accurate rendering of dynamic scenes. We directly apply the updated equations of the Kalman Filter to merge the predicted Gaussian \mathcal{N}_t^{Ob} distributions with the observed data \mathcal{N}_t^P to obtain the updated Gaussian distributions $\hat{\mathcal{N}}_t$:

$$K = \Sigma_t^{Ob} (\Sigma_t^{Ob} + \Sigma_t^P)^{-1}, \quad (6)$$

$$\hat{\mathcal{N}}_t = \mathcal{N}_t^{Ob} + K(\mathcal{N}_t^P - \mathcal{N}_t^{Ob}),$$

where K is the Kalman Gain. The Kalman Gain determines the weight given to the new observation relative to the prior prediction. A higher gain gives more weight to the observation, while a lower gain relies more on the prior prediction. The updated Gaussian distributions $\hat{\mathcal{N}}_t$ determine the final 3D representation at time t and are used to render the result RGB images.

3.2. Wasserstein Regularization

4D Gaussian Splatting essentially updates the parameters of 3D Gaussian distributions based on different input timestamps. Ensuring consistent and smooth updates of these parameters is crucial for high-quality dynamic scene rendering. We hypothesize that flickering artifacts arise when Gaussian distributions undergo abrupt changes in shape or position between consecutive frames.

Previous methods have attempted to constrain these frame-to-frame changes using simple Euclidean metrics. Some works [48, 49] apply Euclidean distance regularization on Gaussian means, while others either ignore covariance updates or use the Frobenius norm for regularization [62]. However, these approaches treat position and shape parameters independently, failing to capture the intrinsic geometric relationship between Gaussian distributions, leading to suboptimal results. Intuitively, instead of updating the 9D parameters (3D mean and 6D covariance) in a Euclidean manner, it is more reasonable to consider the 3D Gaussian distribution as a whole and update it accordingly.

As a solution, we leverage the Wasserstein distance [21] from optimal transport theory [63]. This metric is particularly suitable as it naturally captures both position and shape changes of Gaussian distributions by measuring the optimal mass transportation cost between them. Unlike Euclidean metrics that treat parameters independently, the Wasserstein distance provides a geometrically meaningful way to track the evolution of 3D Gaussians in dynamic scenes.

Specifically, the squared 2-Wasserstein distance between two Gaussian distributions $\mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2)$ is

given by [20]:

$$W_2^2 = \|\mu_1 - \mu_2\|^2 + \text{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1 \Sigma_2)^{\frac{1}{2}}), \quad (7)$$

where $\text{Tr}((\Sigma_1 \Sigma_2)^{\frac{1}{2}}) = \text{Tr}((\Sigma_2^{\frac{1}{2}} \Sigma_1 \Sigma_2^{\frac{1}{2}})^{\frac{1}{2}})$ provides a symmetric version for stable computation. The first term quantifies the squared Euclidean distance between means, and the trace term measures covariance differences in 3-dimensional symmetric positive definite manifold $\text{SPD}(3)$. This formulation captures the geometric and statistical ‘distance’ between the distributions, providing a comprehensive measure of their disparity.

Notably, the trace term in Eq. (7) is isometric under similarity transformations [64]. For 3D Gaussian Splatting with covariance matrices decomposed into rotation ($\mathbf{R}_1, \mathbf{R}_2$) and scale ($\mathbf{S}_1, \mathbf{S}_2$) matrices as Eq. (1), the trace term becomes:

$$D_\Sigma = \text{Tr} \left(\mathbf{S}_1 + \mathbf{S}_2 - 2 \left(\mathbf{S}_1^{1/2} \mathbf{E}_{12} \mathbf{S}_1^{1/2} \right)^{1/2} \right), \quad (8)$$

$$\mathbf{E}_{12} = \mathbf{R}_1^\top \mathbf{R}_2 \mathbf{S}_2 \mathbf{R}_2^\top \mathbf{R}_1,$$

where $\mathbf{S}_1^{1/2}$ is the square root of the diagonal scale matrix \mathbf{S}_1 , and \mathbf{E}_{12} is the covariance matrix of the transformed distribution \mathcal{N}_2 under the rotation \mathbf{R}_1 . This decomposition allows for computationally efficient and stable computation of the matrix square root and eigenvalue decomposition required in the Wasserstein distance calculation. The detailed implementation is provided in Algorithm 1.

Algorithm 1 Wasserstein Distance for 3D Gaussians

Input: Two Gaussians $\mathcal{N}_i = (\mu_i, \mathbf{S}_i, \mathbf{R}_i)$, $i = 1, 2$,
Output: Wasserstein distance $W_2(\mathcal{N}_1, \mathcal{N}_2)$,

- 1: Euclidean difference of means $D_\mu^2 = \|\mu_1 - \mu_2\|^2$
 - 2: Compute $\text{Tr}((\Sigma_1 \Sigma_2)^{1/2})$ by
 $\mathbf{E}_{12} = \mathbf{R}_1^\top \mathbf{R}_2 \mathbf{S}_2 \mathbf{R}_2^\top \mathbf{R}_1$,
 $\mathbf{C}_{12} = \mathbf{S}_1^{1/2} \mathbf{E}_{12} \mathbf{S}_1^{1/2}$,
 Re-symmetrize (Optional): $\mathbf{C}_{12} \leftarrow \frac{1}{2}(\mathbf{C}_{12} + \mathbf{C}_{12}^\top)$,
 Eigenvalue decomposition $\mathbf{C}_{12} \mathbf{e}_k = \lambda_k \mathbf{e}_k$,
 $\text{Tr}((\Sigma_1 \Sigma_2)^{1/2}) = \text{Tr}(\mathbf{C}_{12}^{1/2}) = \sum_k \sqrt{\lambda_k}$.
 - 3: Get the distance $W_2^2 = D_\mu^2 + \sum(\mathbf{S}_1 + \mathbf{S}_2) - 2 \sum_k \sqrt{\lambda_k}$,
 Clamp (Optional) $W_2^2 \geq 0$
 - 4: **return** $W_2(\mathcal{N}_1, \mathcal{N}_2) = \sqrt{W_2^2}$
-

We incorporate the Wasserstein distance into our optimization framework through two complementary losses. The first, our State-Observation Alignment Loss (SOA Loss), enforces physical motion consistency:

$$\mathcal{L}_{\text{SOA}} = W_2^2(\mathcal{N}_t^P, \mathcal{N}_t^{\text{Ob}}), \quad (9)$$

which encourages the predicted Gaussians to align with observations while maintaining physical plausibility. While

observations are inherently error-prone due to discrete temporal sampling, our predictions incorporate prior knowledge of kinematic models. By measuring the Wasserstein distance between predictions and observations, we ensure that our predicted states remain physically coherent while staying close to the observed data.

Secondly, we introduce a Wasserstein regularization term to ensure temporal consistency and mitigate artifacts between consecutive frames for all Gaussians:

$$\mathcal{L}_{\text{WR}} = \sum_t^T \sum_i^N W_2^2(\hat{\mathcal{N}}_t^{(i)}, \hat{\mathcal{N}}_{t-1}^{(i)}), \quad (10)$$

which specifically targets flickering artifacts by penalizing abrupt changes in Gaussian parameters between adjacent frames, promoting smooth motion and deformation over time.

3.3. Modeling Gaussian Dynamics with Wasserstein Geometry

Building upon Wasserstein distance, we model Gaussian dynamics using Wasserstein geometry (Figure 2). The evolution of Gaussian distributions is captured through logarithmic map $-v_t = \log_{\mathcal{N}_t} \mathcal{N}_{t-1}$ for velocity computation and exponential map $\mathcal{N}_{t+1}^P = \exp_{\mathcal{N}_t} v_t$ for prediction.

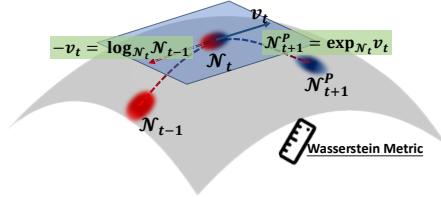


Figure 2. Gaussian dynamics modeling in Wasserstein space. The velocity v_t is computed via logarithmic map between \mathcal{N}_{t-1} and \mathcal{N}_t , then used to predict \mathcal{N}_{t+1}^P through exponential map. Gray regions show log/exp map operations in the wasserstein space; the blue region represents current state.

3.3.1. Logarithmic Map for Velocity Computation

As shown in the gray region of Figure 2, we compute the velocity v_t through the logarithmic map $-v_t = \log_{\mathcal{N}_t} \mathcal{N}_{t-1}$. For two Gaussian distributions characterized by their means μ_{t-1}, μ_t and covariances Σ_{t-1}, Σ_t , the Wasserstein logarithmic map for the mean is directly computed as the Euclidean difference:

$$\Delta\mu_t = \mu_t - \mu_{t-1}, \quad (11)$$

For the covariance, the Wasserstein logarithmic map attributes to the commutator of the matrix square root of the

covariance matrices, which is given from [64] as:

$$\begin{aligned} \log_{\Sigma_t} \Sigma_{t-1} &= (\Sigma_t \Sigma_{t-1})^{\frac{1}{2}} + (\Sigma_{t-1} \Sigma_t)^{\frac{1}{2}} - 2\Sigma_{t-1} \\ &= P + P^\top - 2\Sigma_t, \\ P &= \Sigma_{t-1}^{\frac{1}{2}} \left(\Sigma_{t-1}^{\frac{1}{2}} \Sigma_t \Sigma_{t-1}^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_t^{-\frac{1}{2}}. \end{aligned} \quad (12)$$

where $\Sigma_{t-1}^{1/2}$ is the matrix square root of Σ_{t-1} , $\Sigma_{t-1}^{-1/2}$ is its inverse, and $\log(\cdot)$ denotes the matrix logarithm.

3.3.2. Exponential Map for State Prediction

Following the velocity computation, we predict the future state \mathcal{N}_{t+1}^P using the exponential map $\mathcal{N}_{t+1}^P = \exp_{\mathcal{N}_t} v_t$, as illustrated in Figure 2. This operation maps the velocity vector back to the manifold of Gaussian distributions. The mean prediction conforms to the Euclidean update:

$$\mu_{t+1}^P = \mu_t + \Delta\mu_t \quad (13)$$

For the covariance, the Wasserstein exponential map is computed by solving the Sylvester equation [65]:

$$\begin{aligned} \Sigma_{t+1}^P &= \exp_{\Sigma_t} (\Delta\Sigma_t) \\ &= \Sigma_t + \Delta\Sigma_t + \Gamma_{\Sigma_t}(\Delta\Sigma_t) \Sigma_t \Gamma_{\Sigma_t}(\Delta\Sigma_t)^\top, \end{aligned} \quad (14)$$

where $\Gamma_{\Sigma_t}(\Delta\Sigma_t)$ symbolizes the root of Sylvester equation,

$$\Gamma_{\Sigma_t}(\Delta\Sigma_t) \Sigma_t + \Sigma_t \Gamma_{\Sigma_t}(\Delta\Sigma_t) = \Delta\Sigma_t. \quad (15)$$

This mapping ensures that the predicted covariance Σ_{t+1}^P remains a valid SPD matrix, preserving the geometric properties essential for accurate rendering. Details of its explicit solution are given in [64]. By operating in the tangent space through logarithmic and exponential maps, our approach naturally handles the non-linear nature of Gaussian transformations while maintaining their statistical properties. The complete implementation is summarized in Algorithm 2.

Algorithm 2 Wasserstein Gaussian Updating

Input: Observed Gaussian: $(\mu_{t-1}, \Sigma_{t-1}), (\mu_t, \Sigma_t)$

Output: Predicted Gaussian: $(\mu_{t+1}^P, \Sigma_{t+1}^P)$

- 1: **Step 1: Velocity Computation via Logarithmic Map**
 $\Delta\mu_t = \mu_t - \mu_{t-1}$
 $P = \Sigma_{t-1}^{1/2} (\Sigma_{t-1}^{1/2} \Sigma_t \Sigma_{t-1}^{1/2})^{1/2} \Sigma_t^{-1/2}$
 $\Delta\Sigma_t = 2\Sigma_{t-1} - P - P^\top$
 - 2: **Step 2: State Prediction via Exponential Map**
 $\mu_{t+1}^P = \mu_t + \Delta\mu_t$
 $\Sigma_{t+1}^P = \Sigma_t + \Delta\Sigma_t + \Gamma_{\Sigma_t}(\Delta\Sigma_t) \Sigma_t \Gamma_{\Sigma_t}(\Delta\Sigma_t)^\top$, where
 $\Gamma_{\Sigma_t}(\Delta\Sigma_t)$ is the root of Sylvester equation.
 - 3: **return** $\mu_{t+1}^P, \Sigma_{t+1}^P$
-

3.4. Overall Loss Function

The total loss function combines the State-Observation Alignment Loss, the Wasserstein regularization, and the rendering loss $\mathcal{L}_{\text{render}}$, which measures the discrepancy between the rendered image and the ground truth:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{render}} + \lambda_{\text{SOA}} \mathcal{L}_{\text{SOA}} + \lambda_{\text{WR}} \mathcal{L}_{\text{WR}}, \quad (16)$$

where λ_{SOA} and λ_{WR} are hyperparameters controlling the importance of each term. Algorithm 3 describes the updating process of Gaussian parameters, combining the neural observation and Wassers in our framework.

Algorithm 3 Wasserstein Gaussian Updating

Require: Initial Gaussians $\{\mathcal{N}^{c(i)}\}_{i=1}^N$, Deform Net f_θ ,

- 1: **for** each t to T **do**
- 2: **for** each Gaussian $\mathcal{N}^{(i)}$ **do**
- 3: Compute velocity $\mathbf{v}_t^{(i)} = -\log_{\mathcal{N}_t^{(i)}} \mathcal{N}_{t-1}^{(i)}$ with previous states,
- 4: Get the observation states by inference of Deform Net $\mathcal{N}_t^{Ob(i)} = f_\theta(\mathcal{N}^{c(i)}, t)$
- 5: Get the prediction states $\mathcal{N}_t^{P(i)} = \exp_{\mathcal{N}_{t-1}^{(i)}} \mathbf{v}_{t-1}^{(i)}$,
- 6: Merge predictions and observations
 $K^{(i)} = \Sigma_t^{Ob(i)} (\Sigma_t^{Ob(i)} + \Sigma_t^{P(i)})^{-1}$,
 $\hat{\mathcal{N}}_t^{(i)} = \mathcal{N}_t^{Ob(i)} + K^{(i)} (\mathcal{N}_t^{P(i)} - \mathcal{N}_t^{Ob(i)})$,
Update $\mathcal{N}_t^{(i)} \leftarrow \hat{\mathcal{N}}_t^{(i)}$
- 7: **end for**
- 8: **end for**
- 9: **return** $\hat{\mathcal{N}}_t^{(i)}$

4. Experiments

We evaluate our method on two datasets: a synthetic dataset from D-NeRF [25] and a real-world dataset from Plenoptic Video [26]. The synthetic dataset provides controlled dynamic scenes with ground truth motions, such as moving digits and animated characters, while the real-world dataset captures more complex dynamic scenes, including people performing actions and objects moving in cluttered environments. Our experiments compare our approach against state-of-the-art dynamic scene rendering methods.

4.1. Training Settings

Following [11], we train for 150k iterations on an NVIDIA A800 GPU. The first 3k iterations optimize only 3D Gaussians for stable initialization. We then jointly train 3D Gaussians and deformation field using Adam [66] with $\beta = (0.9, 0.999)$. The 3D Gaussians' learning rate follows the official implementation, while the deformation network's learning rate decays from 8e-4 to 1.6e-6. The Filter module is introduced after 6k iterations, with SOA Loss and

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FPS \uparrow
DyNeRF [26]	29.58	-	0.080	0.015
StreamRF [27]	28.16	0.850	0.310	8.50
HyperReel [30]	30.36	0.920	0.170	2.00
NeRFPlayer [29]	30.69	-	0.110	0.05
K-Planes [31]	31.05	0.950	0.040	1.5
4D-GS [13]	31.8	0.958	0.032	87
Def-3D-Gauss [11]	32.0	0.960	0.030	118
4D-Rotor-Gauss [67]	34.25	0.962	0.048	1250
Ours	34.45	0.970	0.026	581

Table 1. Comparison on D-NeRF dataset.

Wasserstein Regularization Loss activated at 20k iterations ($\lambda_{\text{SOA}} = 0.1$, $\lambda_{\text{WR}} = 0.01$). We conduct experiments on synthetic datasets at 800×800 resolution with white background, and real-world Dataset at 1352×1014 pixels.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FPS \uparrow
DyNeRF [26]	29.58	-	0.080	0.015
StreamRF [27]	28.16	0.850	0.310	8.50
HyperReel [30]	30.36	0.920	0.170	2.00
NeRFPlayer [29]	30.69	-	0.110	0.05
K-Planes [31]	30.73	0.930	0.070	0.10
MixVoxels [32]	30.85	0.960	0.210	16.70
4D-GS [13]	29.91	0.928	0.168	76.2
4D-Rotor-Gauss [67]	31.80	0.935	0.142	289.32
Ours	31.62	0.940	0.140	220.47

Table 2. Quantitative comparison on the Plenoptic Video Dataset.

4.2. Experimental Validation and Analysis

We conduct comprehensive experiments to validate our approach against state-of-the-art methods on both synthetic and real-world scenarios, using PSNR [68], SSIM [69], LPIPS [70], and Frames Per Second (FPS) metrics.

On the D-NeRF dataset, our method achieves **34.45 dB** PSNR and **0.970 SSIM** while maintaining **581 FPS**, significantly outperforming previous methods (Table 1). Figure 3 demonstrates our superior detail preservation and motion handling capabilities.

For real-world evaluation on the Plenoptic Video Dataset, we achieve **31.62 dB** PSNR and **0.940 SSIM** at **220.47 FPS** (Table 2). Figure 4 shows our method's effectiveness in handling complex non-rigid deformations.

4.3. Per-Scene Results

We provide detailed per-scene quantitative comparisons on the D-NeRF dataset to demonstrate the effectiveness of our method across various dynamic scenes. Table 3 and 4 presents the results for each scene, comparing our method with several state-of-the-art approaches. We provide video demonstrations in the supplementary material, which are rendered from fixed camera viewpoints using interpolated continuous timestamps.

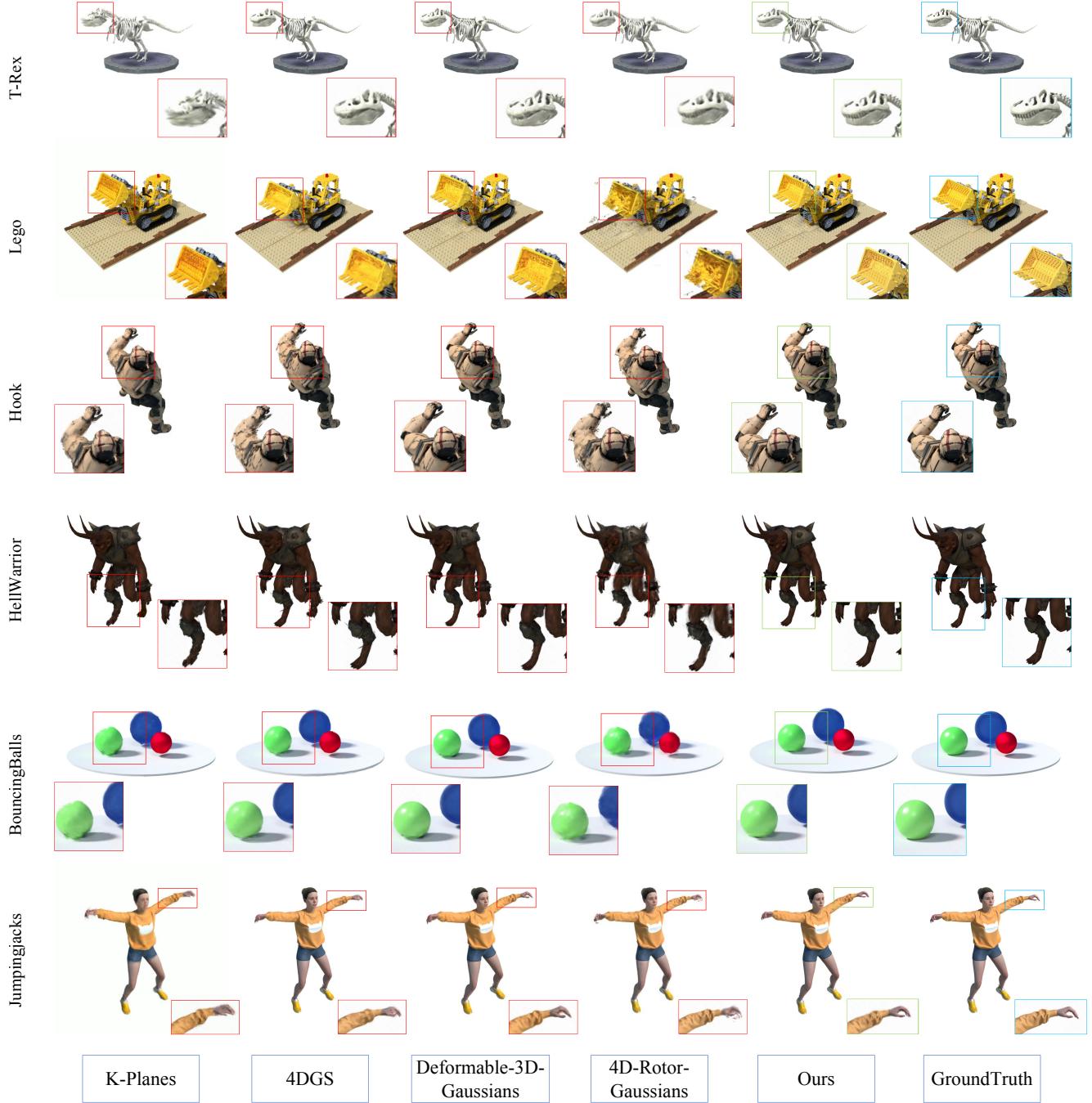


Figure 3. Qualitative results on the synthetic dataset. Zoom in for details.

4.3.1. Analysis of Results

Our method demonstrates strong performance across most scenes in terms of PSNR, SSIM, and LPIPS metrics. In the *Hell Warrior* scene, *Def-3D-Gauss* achieves the highest PSNR of **41.54**, while our method follows closely with **39.06**. This close performance demonstrates the effectiveness of our Wasserstein-constrained state-space modeling in

capturing complex dynamic motions.

In the *Mutant* scene, *Def-3D-Gauss* attains a PSNR of **39.26**, while our method achieves a superior PSNR of **40.77**. Our method also demonstrates better perceptual quality with the lowest LPIPS of **0.0048**, indicating both higher reconstruction quality and better visual results.

For scenes with rapid motions like *Bouncing Balls* and

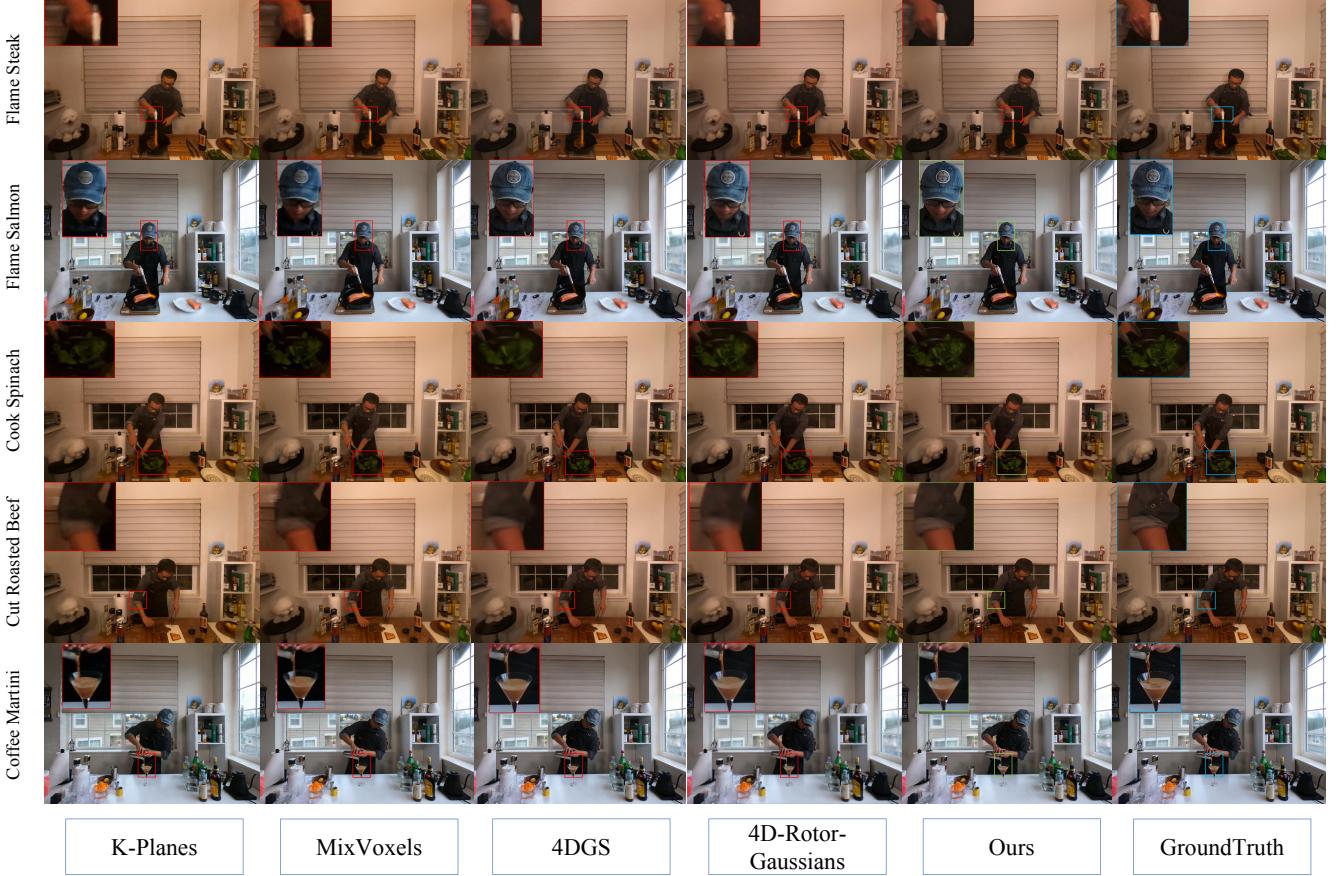


Figure 4. Qualitative results on the real-world dataset. Zoom in for details.

Jumping Jacks, our method maintains robust performance. In *Bouncing Balls*, we achieve a PSNR of **42.79**, surpassing *Def-3D-Gauss*'s **41.01**. In *Jumping Jacks*, our method leads with a PSNR of **37.91**, showcasing our capability in handling challenging dynamic content. The incorporation of Wasserstein geometry allows for smooth and consistent updates of Gaussian parameters, effectively reducing artifacts and ensuring temporal coherence.

Compared to previous methods like *4D-GS* and *4D-Rotor-Gauss*, our method shows consistent improvements across most scenes. For example, in the *Lego* scene, our method achieves a PSNR of **34.74**, surpassing *4D-Rotor-Gauss* by approximately **9.5 dB** and exceeding *Def-3D-Gauss* by **1.67 dB**.

Overall, these results indicate that our method achieves superior average performance while maintaining competitive or leading metrics in most scenes. This confirms the effectiveness of integrating Wasserstein geometry and state-space modeling for dynamic scene rendering.

4.3.2. Comparison with Baseline Methods

Compared to methods like *DyNeRF* and *StreamRF*, which primarily rely on Euclidean metrics for parameter updates,

our approach offers a more geometrically meaningful way to model Gaussian dynamics. The consistent performance improvements illustrate the advantages of our geometric approach over traditional methods.

Methods like *Def-3D-Gauss* and *4D-Rotor-Gauss* improve upon traditional approaches by considering deformations and rotations, and our method builds upon these advances by incorporating Wasserstein geometry and state-space modeling. This comprehensive framework leads to more robust and consistent results across various dynamic scenes.

4.4. Ablation Studies

4.4.1. Effect of the State Consistency Filter

We compare against a baseline that relies solely on observations without the Filter, using Average EndPoint Error (AEPE) [71] as the metric.

Our Filter improves motion estimation by reducing noise in observed flow (left) and producing clearer motion patterns (middle). The residual map (right) shows minimal differences between observation and prediction after training, validating that our Filter successfully balances physical

Method	Hell Warrior			Mutant			Hook			Bouncing Balls		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DyNeRF	24.58	0.9240	0.0807	29.31	0.9472	0.0492	28.02	0.9395	0.0646	29.58	0.9490	0.0523
StreamRF	23.16	0.9150	0.0910	28.31	0.9372	0.0592	27.02	0.9295	0.0746	28.16	0.9350	0.0623
HyperReel	25.36	0.9354	0.0735	30.11	0.9551	0.0401	29.63	0.9433	0.0536	30.36	0.9520	0.0423
NeRFPlayer	24.69	0.9283	0.0810	29.69	0.9451	0.0510	28.69	0.9383	0.0610	30.69	0.9483	0.0510
K-Planes	24.58	0.9520	0.0824	32.50	0.9713	0.0362	28.12	0.9489	0.0662	40.05	0.9934	0.0322
4D-GS	40.02	0.9155	0.1056	37.53	0.9336	0.0580	32.71	0.8876	0.1034	40.62	0.9591	0.0600
Def-3D-Gauss	41.54	0.9873	0.0234	38.10	0.9951	0.0052	42.06	0.9867	0.0144	41.01	0.9953	0.0093
4D-Rotor-Gauss	34.25	0.9620	0.0480	39.26	0.9670	0.0380	33.33	0.9570	0.0420	34.25	0.9620	0.0480
Ours	39.06	0.9863	0.0244	40.77	0.9941	0.0048	41.52	0.9857	0.0154	42.79	0.9943	0.0103
Method	Lego			T-Rex			Stand Up			Jumping Jacks		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DyNeRF	28.58	0.9384	0.0607	29.58	0.9439	0.0487	29.58	0.9401	0.0585	29.58	0.9397	0.0528
StreamRF	27.16	0.9284	0.0707	28.16	0.9339	0.0587	28.16	0.9301	0.0685	28.16	0.9297	0.0628
HyperReel	29.36	0.9484	0.0507	30.36	0.9539	0.0387	30.36	0.9501	0.0485	30.36	0.9497	0.0428
NeRFPlayer	29.69	0.9383	0.0610	30.69	0.9451	0.0510	30.69	0.9483	0.0510	30.69	0.9479	0.0510
K-Planes	33.10	0.9695	0.0331	33.43	0.9737	0.0343	33.09	0.9793	0.0310	31.11	0.9708	0.0402
4D-GS	25.10	0.9384	0.0607	21.93	0.9539	0.0487	38.11	0.9301	0.0785	34.23	0.9297	0.0828
Def-3D-Gauss	33.07	0.9794	0.0183	44.16	0.9933	0.0098	44.62	0.9951	0.0063	37.72	0.9897	0.0126
4D-Rotor-Gauss	25.24	0.9570	0.0530	39.26	0.9595	0.0505	38.89	0.9620	0.0480	33.75	0.9595	0.0505
Ours	34.74	0.9784	0.0193	43.66	0.9940	0.0088	41.24	0.9945	0.0073	37.91	0.9887	0.0136

Table 3. Quantitative comparison on D-NeRF dataset across different scenes. For all metrics, PSNR↑, SSIM↑ indicate higher is better, while LPIPS↓ indicates lower is better. Red, orange and yellow indicate the best, second-best and third-best results respectively.

Method	Hell Warrior			Mutant			Hook			Bouncing Balls		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
4D-GS	29.82	0.9160	0.0856	30.44	0.9340	0.0780	34.67	0.8880	0.0834	39.11	0.9595	0.0600
Def-3D-Gauss	38.55	0.9870	0.0234	39.20	0.9950	0.0053	42.06	0.9865	0.0144	40.74	0.9950	0.0093
4D-Rotor-Gauss	31.77	0.9515	0.0471	33.35	0.9665	0.0297	32.85	0.9565	0.0395	35.89	0.9615	0.0480
Ours	37.77	0.9715	0.0261	40.40	0.9940	0.0045	40.31	0.9710	0.0148	41.79	0.9630	0.0260
Method	Lego			T-Rex			Stand Up			Jumping Jacks		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
4D-GS	24.29	0.9380	0.0507	38.74	0.9535	0.0487	31.77	0.9300	0.0485	24.29	0.9295	0.0428
Def-3D-Gauss	25.38	0.9790	0.0183	44.16	0.9930	0.0099	38.01	0.9950	0.0063	39.21	0.9895	0.0126
4D-Rotor-Gauss	24.93	0.9365	0.0541	31.77	0.9490	0.0511	30.33	0.9430	0.0479	33.40	0.9190	0.0521
Ours	24.74	0.9680	0.0191	44.66	0.9940	0.0088	37.24	0.9730	0.0162	23.93	0.9700	0.0129

Table 4. Quantitative comparison on D-NeRF dataset with **black background**. For all metrics, PSNR↑, SSIM↑ indicate higher is better, while LPIPS↓ indicates lower is better. Red, orange, and yellow indicate the best, second-best, and third-best results respectively. We change the default background colors as set in their official released code.

consistency with observational accuracy.

Method	AEPE ↓
Only Observation (No Filter)	1.45
With State Consistency Filter	1.02

Table 5. AEPE comparison on the Plenoptic dataset, where ground truth optical flow is obtained from the original video sequences.

4.4.2. Effect of Wasserstein Regularization

To evaluate our Wasserstein Regularization (\mathcal{L}_{WR}) and State-Observation Alignment Loss (\mathcal{L}_{SOA}), we generate

continuous video sequences with fixed viewpoints. We evaluate our Wasserstein Regularization by comparing three approaches: without regularization, with Linear Regularization, and with Wasserstein Regularization. The Linear Regularization baseline uses MSE losses:

$$\begin{aligned} \mathcal{L}_{SOA-Linear} &= \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^{\text{Ob}}\|^2 + \|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t^{\text{Ob}}\|_F^2 \\ \mathcal{L}_{R-Linear} &= \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\|^2 + \|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{t-1}\|_F^2 \end{aligned} \quad (17)$$

Quantitatively, the Filter reduces AEPE by 29.7% (from 1.45 to 1.02) on the Plenoptic dataset. This evaluation is particularly meaningful as the Plenoptic dataset provides

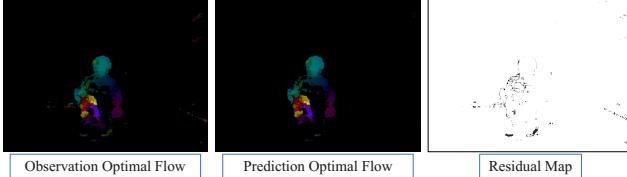


Figure 5. Optical Flow Visualization. Our method naturally derives a speed field by computing 3D motions for all Gaussian points and projecting them to 2D optical flow. Left: Raw observed flow with noticeable noise. Middle: Predicted flow with Filter showing clearer motion boundaries and better dynamic-static separation. Right: Residual map indicating the consistency between observation and prediction.

Method	Quality Metrics			Training Time(h)
	PSNR	SSIM	LPIPS	
<i>D-NeRF Dataset</i>				
w/o Wasserstein Reg.	32.45	0.962	0.032	3.5
w/ Linear Reg.	33.45	0.966	0.029	2.8
w/ Wasserstein Reg.	34.45	0.970	0.026	1.5
<i>Plenoptic Dataset</i>				
w/o Wasserstein Reg.	30.79	0.932	0.145	4.5
w/ Linear Reg.	31.79	0.938	0.141	3.8
w/ Wasserstein Reg.	32.79	0.945	0.138	2.2

Table 6. Comparison of regularization methods. Wasserstein Regularization achieves best quality while reducing training time.

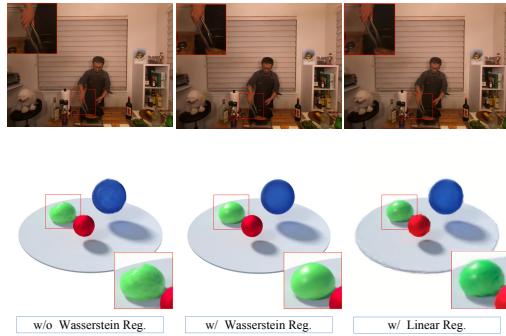


Figure 6. Wasserstein regularization effect on temporal consistency.

continuous frame sequences from single camera views, allowing us to use the optical flow from original videos as Ground Truth for accurate assessment.

Our Wasserstein Regularization improves PSNR by **2.0 dB** over the baseline and **1.0 dB** over Linear Regularization on both datasets, while reducing training time by **57.1%** (D-NeRF) and **51.1%** (Plenoptic). Figure 6 shows how it effectively reduces flickering artifacts by properly handling both positional and shape changes of Gaussians, outperforming the simple Euclidean metrics of Linear Regularization.

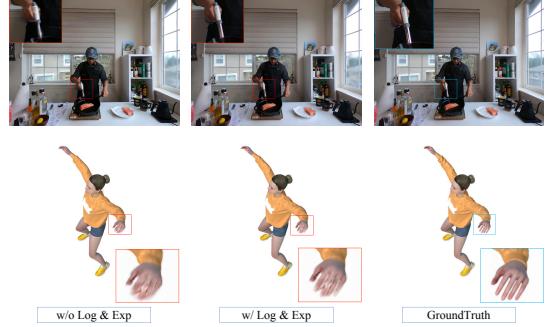


Figure 7. Effect of Wasserstein geometry modeling.

4.4.3. Effect of Modeling Gaussian Dynamics with Wasserstein Geometry

We evaluate our Wasserstein geometry-based dynamics modeling against a baseline using simple Euclidean differences on Gaussian parameters. As shown in Table 7, incorporating Wasserstein geometry modeling (Method 4) improves rendering quality compared to using only Filter and Regularization (Method 3). Figure 7 demonstrates how our log/exp mappings in Wasserstein space better preserve shape deformations, particularly evident in complex motions like hand movements.

4.4.4. Ablation on Model Components

We conduct ablation studies to evaluate each component’s contribution. The results show:

- State Consistency Filter improves PSNR by **0.80 dB** (D-NeRF) and **0.66 dB** (Plenoptic), reducing training time by **20.0%**
- Wasserstein Regularization adds **0.70 dB** PSNR gain on both datasets with **21.4%** further training time reduction
- Full model with Log/Exp maps achieves total PSNR gains of **2.00 dB**, while reducing training time by **57.1%** (D-NeRF) and **51.1%** (Plenoptic)

5. Conclusion

We have introduced *Gaussians on Their Way*, a novel framework that enhances 4D Gaussian Splatting by integrating state-space modeling with Wasserstein geometry. Our approach achieves accurate and temporally coherent dynamic scene rendering by guiding Gaussians along their natural trajectories in the Wasserstein space while maintaining state consistency. This work establishes a promising foundation for dynamic scene representation by combining optimal transport theory with state-space modeling. Future directions include extending to larger-scale scenes, exploring advanced state estimation techniques, and incorporating learning-based methods for improved performance.

Method	Filter	W. Reg.	Log	D-NeRF			Plenoptic			D-NeRF Eff.		Plen. Eff.	
				PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	FPS	Train(h)	FPS	Train(h)
1. Only Obs.	✗	N/A	N/A	32.45	0.962	0.032	30.79	0.932	0.145	650	3.5	250	4.5
2. + Filter	✓	✗	✗	33.25	0.965	0.030	31.45	0.938	0.142	600	2.8	235	3.8
3. + W. Reg.	✓	✓	✗	33.95	0.968	0.028	32.15	0.942	0.140	600	2.2	235	3.0
4. + Log & Exp	✓	✓	✓	34.45	0.970	0.026	32.79	0.945	0.138	581	1.5	220	2.2

Table 7. Ablation study results. Filter: State Consistency Filter; W. Reg.: Wasserstein Regularization; Log: Log & Exp Maps.

References

- [1] C. Wang, M. A. Reza, V. Vats, Y. Ju, N. Thakurde-sai, Y. Wang, D. J. Crandall, S.-h. Jung, and J. Seo, “Deep learning-based 3d reconstruction from multiple images: A survey,” *Neurocomputing*, vol. 597, p. 128018, 2024. [1](#) [2](#)
- [2] M. T. [HUANG] Jiahui, “A survey of dynamic 3d scene reconstruction,” *Journal of Graphics*, vol. 45, no. 1, pp. 14–25, 2024. [1](#)
- [3] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” in *ECCV*, 2020. [1](#) [2](#)
- [4] J. J. Park and et al., “Nerfies: Deformable neural radiance fields,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4480–4490. [1](#) [2](#)
- [5] Z. Li, S. Niklaus, N. Snavely, and O. Wang, “Neural Scene Flow Fields for Space-time View Synthesis of Dynamic Scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6498–6508.
- [6] Z. Yan, C. Li, and G. H. Lee, “Nerf-ds: Neural radiance fields for dynamic specular objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 8285–8295.
- [7] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, “Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields,” *ACM Trans. Graph.*, vol. 40, no. 6, dec 2021. [1](#) [2](#)
- [8] R. Shao, Z. Zheng, H. Tu, B. Liu, H. Zhang, and Y. Liu, “Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [9] B. Park and C. Kim, “Point-dynrf: Point-based dynamic radiance fields from a monocular video,” in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 3159–3169. [1](#)
- [10] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, “4d gaussian splatting for real-time dynamic scene rendering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20310–20320. [1](#) [2](#)
- [11] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, “Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20331–20341. [6](#)
- [12] Y. Duan, F. Wei, Q. Dai, Y. He, W. Chen, and B. Chen, “4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes,” in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.
- [13] Z. Yang, H. Yang, Z. Pan, and L. Zhang, “Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting,” in *The Twelfth International Conference on Learning Representations*. [6](#)
- [14] Z. Li, Z. Chen, Z. Li, and Y. Xu, “Spacetime gaussian feature splatting for real-time dynamic view synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8508–8520.
- [15] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, “Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis,” in *3DV*, 2024. [1](#) [2](#)
- [16] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/> [1](#) [2](#)
- [17] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao, “Recent advances in 3d gaussian splatting,” *Computational Visual Media*, vol. 10, no. 4, pp. 613–642, 2024. [1](#)
- [18] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, “3d gaussian splatting as new era: A survey,” *IEEE Transactions on Visualization and Computer Graphics*, 2024. [1](#)
- [19] D. E. Catlin, *Estimation, control, and the discrete Kalman filter*. Springer Science & Business Media, 2012, vol. 71. [1](#)

- [20] G. H. Givens and R. W. Shortt, “Class of wasserstein distances for probability measures on euclidean spaces,” *The Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984. [1](#) [4](#)
- [21] V. M. Panaretos and Y. Zemel, “Statistical aspects of wasserstein distances,” *Annual review of statistics and its application*, vol. 6, no. 1, pp. 405–431, 2019. [1](#) [4](#)
- [22] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008. [1](#)
- [23] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, “Neural volumes: Learning dynamic renderable volumes from images,” *ACM Trans. Graph.*, vol. 38, no. 4, pp. 65:1–65:14, Jul. 2019. [2](#)
- [24] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, “Local light field fusion: Practical view synthesis with prescriptive sampling guidelines,” *ACM Transactions on Graphics (TOG)*, 2019.
- [25] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, “D-nerf: Neural radiance fields for dynamic scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327. [2](#) [6](#)
- [26] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe *et al.*, “Neural 3d video synthesis from multi-view video,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5521–5531. [2](#) [6](#)
- [27] L. Li, Z. Shen, zhongshu wang, L. Shen, and P. Tan, “Streaming radiance fields for 3d video synthesis,” in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [6](#)
- [28] A. Cao and J. Johnson, “Hexplane: A fast representation for dynamic scenes,” *CVPR*, 2023.
- [29] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger, “Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields,” *IEEE Transactions on Visualization and Computer Graphics*, 2023. [6](#)
- [30] B. Attal, J.-B. Huang, C. Richardt, M. Zollhoefer, J. Kopf, M. O’Toole, and C. Kim, “HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [6](#)
- [31] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, “K-planes: Explicit radiance fields in space, time, and appearance,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 479–12 488. [6](#)
- [32] F. Wang, S. Tan, X. Li, Z. Tian, Y. Song, and H. Liu, “Mixed neural voxels for fast multi-view video synthesis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 19 706–19 716. [2](#) [6](#)
- [33] E. H. Adelson, J. R. Bergen *et al.*, *The plenoptic function and the elements of early vision*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991, vol. 2. [2](#)
- [34] M. LEVOY, “Light field rendering,” in *SIGGRAPH*, 1996. [2](#)
- [35] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, “Unstructured lumigraph rendering,” in *SIGGRAPH*, 2001. [2](#)
- [36] G. Riegler and V. Koltun, “Free view synthesis,” in *ECCV*, 2020. [2](#)
- [37] ———, “Stable view synthesis,” in *CVPR*, 2021. [2](#)
- [38] T. Wu, F. Zhong, A. Tagliasacchi, F. Cole, and C. Oztireli, “D²nerf: Self-supervised decoupling of dynamic and static objects from a monocular video,” in *NeurIPS*, 2022. [2](#)
- [39] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, “Humannerf: Free-viewpoint rendering of moving people from monocular video,” in *CVPR*, 2022. [2](#)
- [40] C. Gao, A. Saraf, J. Kopf, and J.-B. Huang, “Dynamic view synthesis from dynamic monocular video,” in *ICCV*, 2021. [2](#)
- [41] T. Alldieck, H. Xu, and C. Sminchisescu, “imghum: Implicit generative models of 3d human shape and articulated pose,” in *ICCV*, 2021. [2](#)
- [42] T. Jiang, X. Chen, J. Song, and O. Hilliges, “Instavatator: Learning avatars from monocular video in 60 seconds,” in *CVPR*, 2023.
- [43] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu, “Rignerf: Fully controllable neural 3d portraits,” in *CVPR*, 2022.
- [44] Y. Bai, Y. Fan, X. Wang, Y. Zhang, J. Sun, C. Yuan, and Y. Shan, “High-fidelity facial avatar reconstruction from monocular video with generative priors,” in *CVPR*, 2023.
- [45] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao, “Animatable neural radiance fields for modeling dynamic human bodies,” in *ICCV*, 2021. [2](#)
- [46] Z. Yang, H. Yang, Z. Pan, and L. Zhang, “Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting,” in *ICLR*, 2023. [2](#)
- [47] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and W. Xinggang, “4d gaussian splat-

- ting for real-time dynamic scene rendering,” in *CVPR*, 2024. 2
- [48] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, “Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction,” *arXiv preprint arXiv:2309.13101*, 2023. 4
- [49] Y.-H. Huang, Y.-T. Sun, Z. Yang, X. Lyu, Y.-P. Cao, and X. Qi, “Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes,” *arXiv preprint arXiv:2312.14937*, 2023. 4
- [50] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang, “3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting,” in *CVPR*, 2024.
- [51] S. Hu and Z. Liu, “Gauhuman: Articulated gaussian splatting from monocular human videos,” in *CVPR*, 2024.
- [52] Z. Lu, X. Guo, L. Hui, T. Chen, M. Yang, X. Tang, F. Zhu, and Y. Dai, “3d geometry-aware deformable gaussian splatting for dynamic view synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8900–8910. 2
- [53] A. Bewley, Z. Ge, D. Ott, F. Ramos, and A. Upadhyay, “Simple online and realtime tracking,” in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468. 2
- [54] A. J. Davison and et al., “Monoslam: Real-time single camera slam,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2
- [55] B. Wagstaff, E. Wise, and J. Kelly, “A self-supervised, differentiable kalman filter for uncertainty-aware visual-inertial odometry,” 03 2022. 2
- [56] G. Revach, N. Shlezinger, X. Ni, A. L. Escoriza, R. J. Van Sloun, and Y. C. Eldar, “KalmanNet: Neural Network Aided Kalman Filtering for Partially Known Dynamics,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 1532–1547, 2022. 2
- [57] B. Zhang, Y. Cheng, J. Yang, C. Wang, F. Zhao, Y. Tang, D. Chen, and B. Guo, “Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling,” in *Advances in Neural Information Processing Systems*, 2024. 2
- [58] Q. Wang, V. Ye, H. Gao, J. Austin, Z. Li, and A. Kanazawa, “Shape of motion: 4d reconstruction from a single video,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.13764> 2
- [59] Y. Zhan and et al., “Kfd-nerf: Kalman filter-based deformation field for dynamic neural radiance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [60] S. Ramasinghe, V. Shevchenko, G. Avraham, H. Hussain, and A. van den Hengel, “Improving the convergence of dynamic nerfs via optimal transport,” 2024. 2
- [61] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Transactions of the ASME—Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960. 3
- [62] P. J. Bickel and E. Levina, “Regularized estimation of large covariance matrices,” 2008. 4
- [63] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2021, vol. 58. 4
- [64] Y. Luo, S. Zhang, Y. Cao, and H. Sun, “Geometric characteristics of the wasserstein metric on $\text{spd}(n)$ and its applications on data processing,” *Entropy*, vol. 23, no. 9, p. 1214, 2021. 4, 5
- [65] N. J. Higham, “Cayley, sylvester, and early matrix theory,” *Linear Algebra Appl.*, vol. 428, pp. 39–43, 2008. 5
- [66] D. P. Kingma and J. Ba, “Adam: A method for Stochastic Optimization,” *arXiv preprint arXiv:1412.6980*, 2014. 6
- [67] Y. Duan, F. Wei, Q. Dai, Y. He, W. Chen, and B. Chen, “4d-rotor gaussian splatting: Towards efficient novel view synthesis for dynamic scenes,” in *Proc. SIGGRAPH*, 2024. 6
- [68] J. Korhonen and J. You, “Peak signal-to-noise ratio revisited: Is simple beautiful?” in *2012 Fourth International Workshop on Quality of Multimedia Experience*. IEEE, 2012, pp. 37–38. 6
- [69] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 6
- [70] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595. 6
- [71] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, “Performance of optical flow techniques,” *International journal of computer vision*, vol. 12, pp. 43–77, 1994. 8