Sonata: Spectral-Oriented Neural Approximation with Transformable Activations for Dynamic Gaussian Splatting

Anonymous submission

Abstract

Dynamic scenes often exhibit spatially varying and non-stationary motion patterns, resulting in distinct frequency responses across different regions. This poses a significant challenge for accurate modeling. Traditional MLP-based deformation fields, limited by fixed activation functions and poor adaptability, struggle to capture local details of nonlinear motion, thereby restricting the modeling capability of dynamic Gaussian splatting. To address these limitations, we propose Sonata, a novel Gaussian splatting framework that introduces the hierarchical modeling principle of Kolmogorov–Arnold Networks (KANs), opening a new paradigm for dynamic scene reconstruction. Unlike MLPs with fixed activation mechanisms, KANs replace each neuron with a transformable activation functions, enabling more flexible modeling of nonlinear transformations. Moreover, to effectively decouple high- and low-frequency components in the transformation field, we design a frequency-guided learning strategy. By incorporating supervisory signals from both the wavelet and Fourier domains during training, our method explicitly guides the model to perceive and separate frequency components, avoiding convergence to local minima. This results in improved structural reconstruction accuracy, detail fidelity, and training stability. Extensive experiments on multiple public dynamic novel view synthesis datasets demonstrate that Sonata achieves state-of-the-art performance across various metrics. Anonymous project page: https://sonatags.github.io/.

Introduction

Recent advances in 3D Gaussian Splatting (3DGS) (Li et al. 2024b; Luiten et al. 2024) have brought transformative progress to scene reconstruction by replacing traditional volumetric rendering (Park et al. 2021b; Fridovich-Keil et al. 2023; Cao and Johnson 2023; Wang et al. 2023a; Shao et al. 2023) with tile-based rasterization, significantly improving rendering efficiency and visual fidelity. While these methods have enabled efficient and photorealistic modeling for static scenes, extending 3DGS to dynamic scenarios remains challenging due to complex motion patterns, occlusions, and temporal sparsity. To address these issues, 4D Gaussian Splatting (4DGS) (Huang et al. 2024; Wu et al. 2024; Yang et al. 2024; Luiten et al. 2024) has emerged as a promising solution, offering fast training and high-quality rendering for dynamic scenes, surpassing conventional NeRF-based approaches (Fridovich-Keil et al. 2023; Shao et al. 2023; Wang et al. 2023a).

Existing 4DGS approaches can be roughly categorized into two groups. The first directly models the temporal evolution of Gaussian parameters via predefined time functions (Li et al. 2024c; Park et al. 2025), achieving high visual fidelity but suffering from limited generalization across scenes, and significant manual design overhead (Li et al. 2024c; Park et al. 2025). The second group employs deformation-based representations (Wu et al. 2024; Yang et al. 2024), where a learnable deformation field warps a canonical Gaussian field over time, typically parameterized by a multilayer perceptron (MLP) with latent embeddings. While these methods reduce per-frame redundancy, they still struggle to capture complex high-frequency dynamics due to the limited adaptability of fixed activation functions and the globally coupled nature of MLP architectures.

Dynamic scene modeling poses significant challenges due to the spatially varying and non-stationary nature of motion patterns. Different regions often exhibit distinct nonlinear behaviors and frequency responses, particularly around occlusions, object boundaries, or fast-moving areas. Traditional deformation-field representations based on MLPs suffer from limited adaptability and fixed activation functions, making them inadequate for capturing local high-frequency dynamics in such complex scenarios. To overcome these limitations, we propose a novel dynamic Gaussian Splatting framework that addresses the shortcomings of existing deformation-based 4DGS methods. Specifically, we construct a transformation field based on the Kolmogorov–Arnold Network (KAN), where each neuron is replaced with a learnable univariate function parameterized by radial basis functions (RBFs). To enable explicit separation of high- and low-frequency components in this design, we introduce a frequency-guided learning strategy that incorporates spectral supervision from wavelet (high-frequency) and Fourier (low-frequency) transforms during training. These structured frequency cues help guide the deformation field to better recognize and disentangle frequency components. This enables more accurate reconstruction of both global structures and fine-grained details, helps avoid local minima, and ultimately improves training stability and reconstruction fidelity.

Our main contributions are as follows:

- We propose a novel perspective for modeling the tempo-

ral evolution of dynamic Gaussian primitives, introducing a neural approximation scheme with transformable activation functions. This opens a new pathway for dynamic scene reconstruction.

- We design a transformation field based on a KAN parameterized by RBFs, replacing the conventional MLP with fixed activations. This enhances the modeling capacity for nonlinear motion patterns and improves representation flexibility.
- A frequency-guided learning paradigm is introduced, leveraging both wavelet and Fourier transforms as structured spectral supervision. This design facilitates frequency decomposition and promotes more stable convergence with improved detail fidelity.
- Extensive experiments show that our method outperforms state-of-the-art baselines in monocular dynamic scene reconstruction, delivering superior quantitative metrics and finer visual details across complex motion scenarios.

Related Work

Novel view synthesis for static scenes

In recent years, NeRF has achieved remarkable success in rendering quality and novel view synthesis (Fridovich-Keil et al. 2023; Shao et al. 2023). However, it heavily relies on dense ray tracing, making the overall process extremely time-consuming. To address these limitations, researchers have proposed various methods to accelerate both the training and rendering processes of NeRF (Hu et al. 2022; Wang et al. 2023b; Barron et al. 2021), and to improve rendering quality. However, these improvements still fall short of meeting the demands of real-time rendering.

With the emergence of 3DGS, a novel framework has been introduced for real-time, high-fidelity novel view synthesis in complex scenes. In static scenes, numerous improvements to 3DGS have rapidly emerged. GeoGaussian (Li et al. 2024b), for instance, addresses the issue of geometric degradation during the optimization of Gaussian splatting by initializing thin Gaussians aligned with surface normals, and applying dense strategies and geometric constraints to enhance the generation of structured regions. This leads to outstanding performance in both novel view synthesis and geometric reconstruction. SmileSplat (Li et al. 2024a) proposes a novel generalizable Gaussian splatting method that reconstructs pixel-aligned Gaussian surfels from unconstrained sparse multi-view images. By jointly optimizing Gaussian parameters and camera poses, it achieves high-quality novel view synthesis.

Novel view synthesis for dynamic scenes

Extending novel view synthesis from static to dynamic scenes is a highly challenging task, as it requires accurately modeling temporal correlations and variations within the 3D space. Some NeRF-based approaches have made progress by introducing deformation field techniques (Park et al. 2021b; Wang et al. 2023a; Guo et al. 2023), which represent the scene as a combination of a canonical field and a corresponding deformation field. These deformation fields describe how sampled points vary relative to the static scene across different time frames. Additionally, other studies (Fridovich-Keil et al. 2023; Cao and Johnson 2023; Shao et al. 2023) simplify the handling of 4D data by dividing the spatiotemporal domain into planar or hash grids. Such spatiotemporal grid-based methods effectively capture temporal correlations. However, NeRF-based rendering techniques still require heavy per-ray sampling during inference, which limits their potential for real-time applications.

Building upon the 3DGS framework, various dynamic scene reconstruction techniques have been proposed (Yang et al. 2024; Wu et al. 2024; Luiten et al. 2024), introducing the temporal dimension into traditional 3DGS for novel modeling and rendering paradigms. For example, Deformable 3d gaussians (D3DGS) (Yang et al. 2024) optimizes a neural deformation field to capture temporal variations, thereby maintaining consistency in the 3D Gaussian representation over time. 4D Gaussian Splatting (4DGS) (Wu et al. 2024) proposes an innovative explicit representation that combines 3D Gaussian functions with 4D neural voxels. It introduces a HexPlane-inspired (Cao and Johnson 2023) factorized voxel encoding scheme to efficiently construct Gaussian features from the 4D neural volume.

Preliminaries

3D Gaussian Splatting

3DGS represents a scene using anisotropic 3D Gaussians initialized from Structure-from-Motion (SfM). Each Gaussian is parameterized by a mean $\boldsymbol{\mu} \in \mathbb{R}^3$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$. Through differentiable rasterization, these parameters are optimized end-to-end, while tile-based GPU rendering enables real-time novel view synthesis. This framework achieves high-fidelity and photorealistic rendering with efficient computation:

$$G(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{1}$$

Its covariance matrix $\boldsymbol{\Sigma}$ is parameterized via a rotation-scaling decomposition $\boldsymbol{\Sigma} = RSS^\top R^\top$, where $R \in \mathbb{R}^{3\times3}$ is an orthogonal rotation matrix and $S \in \mathbb{R}^{3\times3}$ is a diagonal scaling matrix. This formulation ensures that $\boldsymbol{\Sigma}$ is positive semi-definite.

During the rendering stage, each 3D Gaussian is first projected into the 2D image space, and then composited using α blending to compute the final pixel color $C \in \mathbb{R}^3$, enabling real-time photorealistic rendering from arbitrary viewpoints:

$$C = \sum_{i\in N} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) \tag{2}$$

here $\alpha_i \in \mathbb{R}$ denotes the opacity of the i-th Gaussian after 2D projection, and $c_i \in \mathbb{R}^3$ represents its view-dependent color modeled using spherical harmonics. The index i iterates over the sorted set of Gaussians involved in the rendering process.

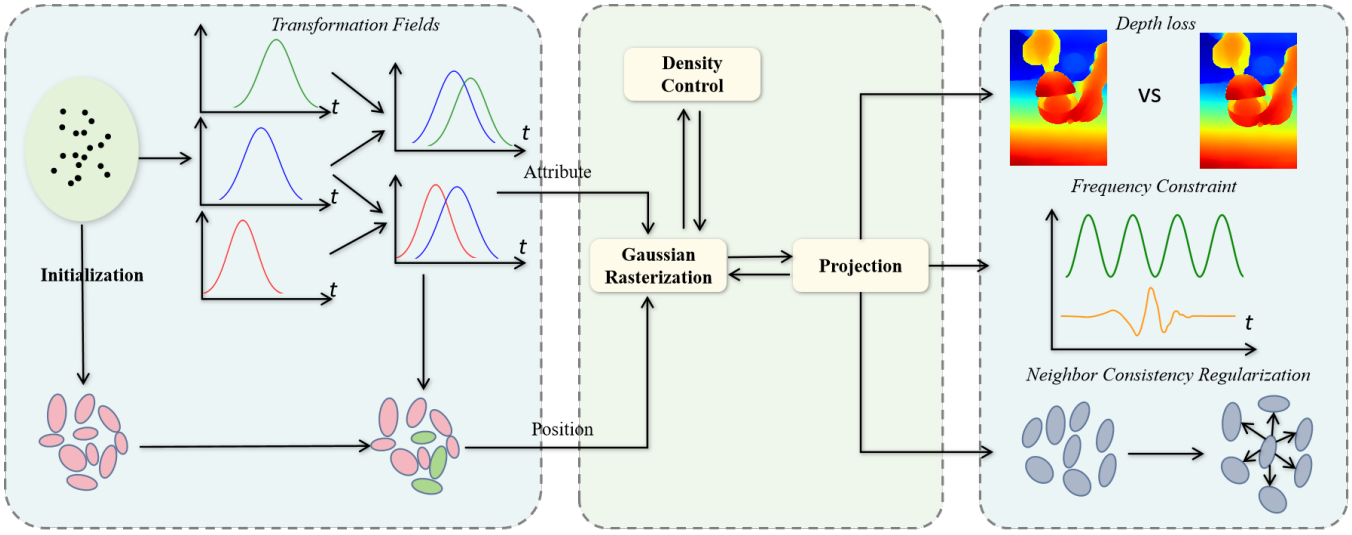

Figure 1: **Overview of the proposed framework**. We propose a novel perspective for modeling the temporal evolution of dynamic Gaussian primitives through a neural approximation-based transformation field with transformable activations. Guided by frequency-aware supervision, the transformation field dynamically modulates its spectral response during training, thereby enhancing reconstruction quality and mitigating convergence to local minima.

Theory

Fundamental Theory of KANs

KANs (Liu et al. 2024) are based on the Kolmogorov–Arnold representation theorem, **which states that any continuous multivariate function can be expressed as a finite sum of hierarchically composed univariate functions**. Building on this principle, KANs approximate complex high-dimensional functions by composing simpler functions in a structured, layered manner.

Formally, a KANs approximates a multivariate function $f(\mathbf{x})$ via the decomposition:

$$\mathcal{G}(\mathbf{x}) = \sum_q \Phi_q \left(\sum_p \phi_{q,p}(x_p) \right), \tag{3}$$

where $\phi_{q,p}$ and Φ_q are univariate basis functions, often implemented as spline functions,e.g., Basis splines(B-splines). This structure reflects the Kolmogorov–Arnold theorem's strategy of **hierarchical decomposition** and enables KANs to isolate **variable-specific transformations** before integrating them through outer-layer mappings.

B-spline basis function KANs constructs the function families Φ and ϕ using B-spline basis functions (Liu et al. 2024), which are theoretically capable of approximating any smooth univariate function to arbitrary precision.

B-splines are a family of piecewise-defined basis functions widely used for function approximation, interpolation, and geometric modeling. A B-spline of degree k is defined over a non-decreasing sequence of real numbers called the knot vector: $t = t_0, t_1, \ldots, t_{n+k}$ where k is the order of the spline (degree = $k-1$) and n is the number of control points minus one.

Cubic B-spline The cubic B-spline (degree = 3, $k = 4$) is one of the most commonly used spline types due to its balance between smoothness and locality. A B-spline curve using control points $\{P_i\}$ can be written as:

$$S(x) = \sum_{i=0}^{n} P_i B_{i,3}(x) \tag{4}$$

However, during training, the dynamic expansion of input variable ranges often causes inputs to fall outside the predefined spline domain. To preserve approximation accuracy, it becomes necessary to frequently invoke the de Boor–Cox algorithm (Gordon and Riesenfeld 1974) to recompute the basis functions and rescale the spline grids. This iterative rescaling mechanism introduces significant computational overhead in practice, forming a critical bottleneck that limits the efficiency of KANs in real-world deployment.

How to Make KANs Efficient and Stable

KANs aim to replace traditional weight-centric neural networks with a function-centric paradigm, learning univariate transformations via adaptive basis functions. To scale this idea to practical, large-scale scenarios, KANs require a design that is both **computationally efficient** and **optimization-stable**—particularly under dynamic input distributions and deep network compositions. This section introduces a basis function strategy that fulfills these requirements.

Why Use RBFs in KANs? Existing implementations of KANs typically employ **B-spline basis functions** to construct learnable univariate mappings. While B-splines offer good locality and approximation capabilities, they are defined only within fixed input grids, making them highly sensitive to input shifts and prone to domain overflow. More-

over, their piecewise structure leads to non-smooth transitions and increased implementation complexity, often resulting in numerical instability and training difficulties.

To address these limitations, we replace B-splines with **Gaussian Radial Basis Functions (RBFs)**. RBFs exhibit significant advantages over B-splines. As they are supported across the entire input space, RBFs are particularly well-suited for handling high-dimensional data and complex distributions, effectively avoiding the boundary-related limitations commonly associated with B-splines. RBFs possess high-order smoothness—especially Gaussian RBFs, which are infinitely differentiable—helping to reduce numerical instability, particularly during gradient computation. The high degree of continuity offered by RBFs contributes to maintaining smooth gradient flow in deep networks, thereby enhancing training stability. Furthermore, RBFs are relatively simple to implement, and their inherent smoothness and continuity facilitate faster convergence of optimization algorithms, ultimately shortening training time. Additionally, the mathematical foundations of commonly used RBF kernels—such as Gaussian and polynomial kernels—make model behavior more interpretable and analytically grounded.

This substitution preserves the theoretical approximation power of KANs while significantly improving their scalability and robustness. It enables the construction of deeper, more expressive, and easier-to-train KANs architectures that are viable in practical settings.

Transformable Activations. Let each learnable univariate function $f : \mathbb{R} \to \mathbb{R}$ in KANs be approximated as a sum of RBFs:

$$\begin{aligned} f(x) &= \sum_{i=1}^{N} \alpha_i \cdot \psi_i(x) \\ \psi_i(x) &= \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right) \end{aligned} \tag{5}$$

Where $\alpha_i \in \mathbb{R}$ is the weight of the i-th basis function; $\mu_i \in \mathbb{R}$ is the center of the RBF; $\sigma > 0$ controls the width of the basis function. **The fundamental paradigm of transformable activations lies in the activation function undergoing a transformable operation with the scaling weights, thereby endowing it with shape plasticity in the function space.**

(Smooth Gradients). The function is infinitely differentiable, with gradient given by:

$$\frac{df(x)}{dx} = \sum_{i=1}^{N} \alpha_i \cdot \left(-\frac{x-\mu_i}{\sigma^2}\right) \cdot \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right) \tag{6}$$

$$f^{(k)}(x) = \sum_{i=1}^{N} \alpha_i \frac{d^k}{dx^k}\psi_i(x) \tag{7}$$

where $f^{(k)}(x)$ denotes the k-th order derivative of $f(x)$ with respect to x. Since $\psi_i(x)$ is a Gaussian function, it is infinitely differentiable over $\mathbb{R}$:

$$\frac{d^k}{dx^k}\psi_i(x) = \psi_i(x) \cdot P_k\left(\frac{x-\mu_i}{\sigma}\right) \tag{8}$$

where P_k is a polynomial related to the Hermite polynomial family. This guarantees smooth and stable gradient flow through layers.

(Parameter Trainability). All RBF parameters are trainable and differentiable:

$$\frac{\partial f(x)}{\partial \mu_i} = \alpha_i \cdot \left(\frac{x-\mu_i}{\sigma^2}\right) \cdot \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right) \tag{9}$$

$$\frac{\partial f(x)}{\partial \sigma} = \alpha_i \cdot \left(\frac{(x-\mu_i)^2}{\sigma^3}\right) \cdot \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right) \tag{10}$$

This makes the RBF-based module fully compatible with gradient-based optimization.

Transformation Field Modeling

In dynamic scene reconstruction, accurately modeling object motion and maintaining temporal consistency requires learning a deformation field that maps coordinates from a canonical space to their time-dependent, deformed positions. To achieve both precision and interpretability in motion modeling, we propose a deformation field parameterized by RBF-driven KANs. This section outlines the construction of the deformation field $\mathcal{K}_\theta(\cdot)$.

Specifically, we define the deformation function $\mathcal{K}_\theta(\cdot)$, which takes two types of learnable embeddings as input: A per-Gaussian embedding $\mathbf{x}_e \in \mathbb{R}^{32}$, which encodes the canonical geometric and appearance characteristics of each Gaussian; A per-frame temporal encoding $\hat{t}$, which captures the scene dynamics at frame t.

The two embeddings are concatenated and passed into the deformation network:

$$\delta_x, \delta_s, \delta_r = \mathcal{K}_\theta(\hat{t}, \mathbf{x}_e) \tag{11}$$

where $\delta_x \in \mathbb{R}^3$, $\delta_s \in \mathbb{R}^3$ and $\delta_r \in \mathbb{R}^3$ represent the predicted residuals for position, scale and rotation, respectively. The deformation network $\mathcal{K}_\theta$ is constructed using RBF-driven KANs, which ensure that the resulting transformation field is not only infinitely differentiable, but also interpretable.

The predicted offsets are subsequently applied to the canonical Gaussian properties to derive their dynamic counterparts at each frame, allowing each Gaussian primitive to evolve smoothly and coherently over time. Furthermore, we propose a frequency-guided training strategy that shapes each RBF-driven KANs activation function through frequency-domain supervision. By leveraging frequency-aware losses, the transformation field is guided to dynamically adapt its spectral response during training, effectively learning to capture both low-frequency structural components and high-frequency details.

How does the transformation field decompose frequencies? The transformation field incorporates two key mechanisms: a hierarchical nested structure inspired by the Kolmogorov–Arnold representation theorem, and an input-variable decoupling strategy. The former employs a recursive nesting of outer and inner functions to incrementally

Table 1: Quantitative results on the NeRF-DS (Yan, Li, and Lee 2023) dataset. The best result is shown in **bold**, and the second-best is highlighted with a gray background. LPIPS-V is based on VGG network.

Method	As			Basin			Bell			Cup		
	PSNR↑	SSIM↑	LPIPS-V↓	PSNR↑	SSIM↑	LPIPS-V↓	PSNR↑	SSIM↑	LPIPS-V↓	PSNR↑	SSIM↑	LPIPS-V↓
3D-GS (Kerbl et al. 2023)	22.69	0.802	0.299	18.42	0.717	0.315	21.01	0.789	0.250	21.71	0.830	0.255
TiNeuVox (Fang et al. 2022)	21.26	0.829	0.397	**20.66**	0.815	0.269	23.08	0.824	0.257	19.71	0.811	0.364
HyperNeRF (Park et al. 2021b)	25.58	**0.895**	0.178	20.41	**0.820**	0.191	23.06	0.810	0.205	24.59	0.877	0.165
NeRF-DS (Yan, Li, and Lee 2023)	25.13	0.878	0.174	19.96	0.817	0.186	23.19	0.821	0.187	**24.91**	0.874	0.174
D3DGS (Yang et al. 2024)	26.31	0.884	0.178	19.67	0.793	0.190	25.74	0.850	0.154	24.86	0.891	0.153
SCGS (Huang et al. 2024)	26.20	-	**0.142**	19.60	-	**0.154**	25.10	-	0.117	24.50	-	**0.115**
ADCGS (Huang et al. 2025)	26.26	0.854	0.183	19.75	0.768	0.180	24.61	0.827	0.179	24.37	0.878	0.163
4DGS (Wu et al. 2024)	25.58	0.861	0.165	19.62	0.783	0.156	25.55	0.861	0.132	24.40	0.874	0.147
Ours	**26.39**	0.874	0.152	19.84	0.793	0.159	**26.02**	**0.883**	**0.104**	24.73	**0.893**	0.126

	Plate			Press			Sieve			Mean		
	PSNR↑	SSIM↑	LPIPS-V↓	PSNR↑	SSIM↑	LPIPS-V↓	PSNR↑	SSIM↑	LPIPS-V↓	PSNR↑	SSIM↑	LPIPS-V↓
3D-GS (Kerbl et al. 2023)	16.14	0.697	0.409	22.89	0.816	0.290	23.16	0.820	0.225	20.29	0.782	0.292
TiNeuVox (Fang et al. 2022)	20.58	0.803	0.332	24.47	0.861	0.300	21.49	0.827	0.318	21.61	0.823	0.278
HyperNeRF (Park et al. 2021b)	18.93	0.771	0.294	26.15	**0.890**	0.196	25.43	0.880	0.165	23.45	0.849	0.199
NeRF-DS (Yan, Li, and Lee 2023)	20.54	0.804	**0.200**	25.72	0.862	0.205	25.78	**0.890**	0.147	23.60	0.849	0.182
D3DGS (Yang et al. 2024)	20.48	**0.812**	0.222	26.01	0.865	0.191	25.70	0.872	0.150	24.11	0.852	0.177
SCGS (Huang et al. 2024)	20.20	-	0.202	**26.60**	-	**0.135**	26.00	-	**0.114**	24.11	-	**0.140**
ADCGS (Huang et al. 2025)	20.56	0.784	0.244	25.86	0.827	0.208	25.59	0.857	0.164	23.86	0.828	0.189
4DGS (Wu et al. 2024)	20.03	0.762	0.250	26.41	0.873	0.138	25.46	0.842	0.177	23.86	0.837	0.166
Ours	**21.10**	0.811	**0.200**	26.43	0.862	0.146	**26.44**	0.877	0.134	**24.42**	**0.856**	0.146

enhance the model's expressivity for high-dimensional functions, while also helping to mitigate overfitting. **The latter assigns each input variable a set of transformable RBF basis functions, enabling independent modeling and frequency decoupling. Specifically, each input variable is modeled using a group of RBF kernels with centers uniformly distributed over the input domain. The scaling weights of these kernels determine their response shapes (e.g., sharp or smooth), corresponding to different frequency bands. From a frequency-domain perspective, sharper (narrower) kernels exhibit broader spectral responses and better capture high-frequency components, whereas smoother (wider) kernels respond more strongly to low-frequency content. Through backpropagation, the network automatically adjusts the scaling weights, thereby selecting appropriate frequency components for each input variable.**

Learning Frequency Features in Dual Domains

We introduce learnable gating mechanisms that adaptively modulate frequency cues derived from wavelet and Fourier representations, guiding the transformation field to focus on both local details and global structures during training.

Wavelet for Localized Detail Enhancement. We apply a differentiable discrete wavelet transform (DWT) to both the predicted image $\hat{I}$ and the ground truth I, extracting high-frequency subbands W^h:

$$\_, W^h = \text{DWT}(\hat{I}), \quad \_, W^h_{gt} = \text{DWT}(I) \tag{12}$$

The wavelet loss is then computed as the Mean Squared Error (MSE) over high-frequency subbands at each decomposition level:

$$\mathcal{L}_{\text{wave}} = \sum_{l=1}^{L} \sum_{j=1}^{3} \text{MSE}(W_l^h[j], W_{gt,l}^h[j]) \tag{13}$$

where L is the number of wavelet decomposition levels, and $j = 1, 2, 3$ denote the HL, LH, and HH subbands.

Fourier for Global Spectral Consistency. To encourage structural coherence, we compute the amplitude spectra using FFT:

$$\hat{F} = \text{FFT}(\hat{I}), \quad F_{gt} = \text{FFT}(I) \tag{14}$$

$$\mathcal{L}_{\text{fourier}} = \text{MSE}(|\hat{F}|, |F_{gt}|) \tag{15}$$

Learnable Gating. We use learnable parameters to dynamically balance the two losses:

$$\lambda_{\text{wave}} = \sigma(g_{\text{wave}}) \cdot \lambda_{\text{wave}}^{\max}, \quad \lambda_{\text{fourier}} = \sigma(g_{\text{fourier}}) \cdot \lambda_{\text{fourier}}^{\max} \tag{16}$$

This enables adaptive frequency supervision, benefiting both fine detail and global consistency. The total frequency-aware loss is computed as:

$$\mathcal{L}_{\text{freq}} = \lambda_{wave} \cdot \mathcal{L}_{\text{wave}} + \lambda_{fourier} \cdot \mathcal{L}_{\text{fourier}} \tag{17}$$

Neighbor Consistency Regularization

To encourage locally consistent deformations in dynamic objects—where neighboring Gaussians typically undergo similar transformations—we introduce a Neighbor Consistency Regularization on the per-Gaussian embedding $\mathbf{z}_g$, inspired by (Luiten et al. 2024):

$$\mathcal{L}_{\text{reg}} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \sqrt{w_{i,k} \cdot \|\mathbf{x}_{ei} - \mathbf{x}_{ek}\|_2^2 + \varepsilon} \tag{18}$$

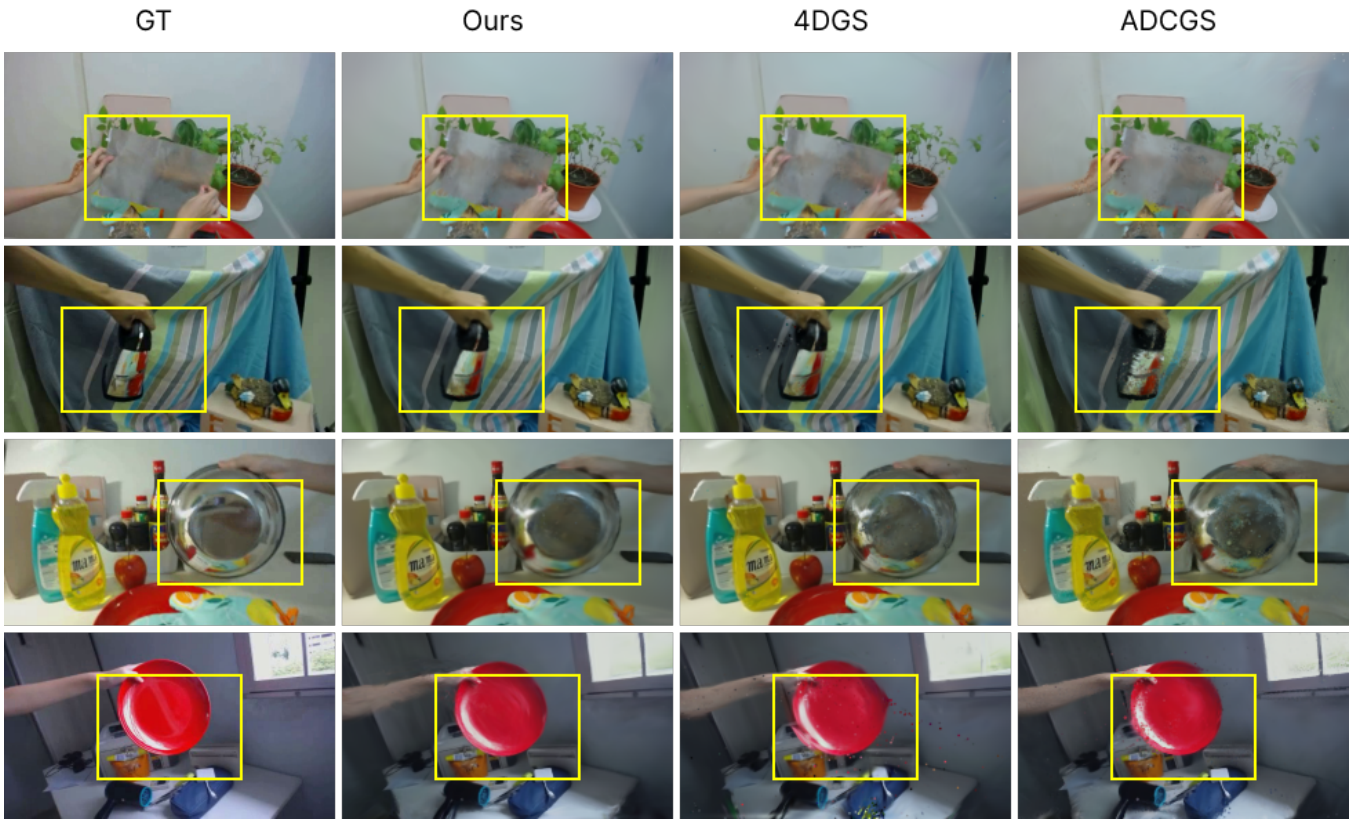

Figure 2: Qualitative comparison of novel view synthesis on the NeRF-DS (Yan, Li, and Lee 2023) dataset, with problem regions highlighted in boxes. **More results can be found in the supplementary material**.

where the weight $w_{i,k} = \exp(-\gamma_\theta |d_{ki}|_2)$ depends on the spatial distance between Gaussian centers i and k. We set $\varepsilon = 1e-20$, $\gamma_\theta = 2000$ and use $K = 20$ nearest neighbors following (Luiten et al. 2024). To reduce computational overhead, the KNN neighborhood is computed only during densification.

Experimental

In this section, we present the implementation details of our proposed method and conduct quantitative comparisons with state-of-the-art (SOTA) approaches on two widely used dynamic scene datasets. Additionally, we perform ablation studies and in-depth analyses to comprehensively evaluate the performance and effectiveness of our method.

Experimental Settings

Our framework is implemented using PyTorch, allowing for streamlined training procedures and seamless integration of all model components. To obtain high-quality initial point clouds and camera parameters, we apply COLMAP (Schonberger and Frahm 2016) for sparse reconstruction on each scene. The RAdam (Liu et al. 2020) optimizer is employed, with an adaptive learning rate tailored to the complexity of each task during model optimization.

Datasets and Evaluation Protocol

We evaluate our method on two dynamic datasets: HyperNeRF (Park et al. 2021b) and Nerf-Ds (Yan, Li, and Lee 2023). The HyperNeRF dataset includes four distinct scenes characterized by rich non-rigid deformations and challenging view-dependent occlusions. The Nerf-Ds dataset contains seven dynamic scenes with complex deformations, occlusions, and illumination changes, making it well-suited for assessing the spatiotemporal generalization capabilities of neural rendering methods.

To evaluate performance, we adopt three standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), which respectively assess pixel accuracy, structural consistency, and perceptual similarity via deep features. Together, they offer a comprehensive evaluation of our method's effectiveness in dynamic monocular scene synthesis.

Quantitative Analysis of Novel View Synthesis

We evaluate the proposed method on two challenging real-world dynamic scene benchmarks: **NeRF-DS** (Yan, Li, and Lee 2023) and **HyperNeRF** (Park et al. 2021b). Our method demonstrates consistent superiority over existing approaches across various metrics, validating both its fidelity and scalability.

Performance on NeRF-DS. On the NeRF-DS dataset, as shown in Table 1, our method achieves the best overall performance, attaining the highest average PSNR (24.42) and SSIM (0.856), along with a competitive LPIPS (0.146). These results demonstrate superior reconstruction fidelity and structural consistency, outperforming prior methods such as 4DGS, Deformable-GS, and HyperNeRF in both pixel-level quality and perceptual similarity. Furthermore, across various scenes—*As, Bell, Plate*, and *Sieve*—our method achieves top-1 PSNR performance, highlighting its effectiveness in recovering sharp structural details and preserving fine textures under complex deformations. In *Bell*, and *Plate*, it consistently ranks first or second across all metrics, with especially strong results in LPIPS-V, demonstrating robustness to varying frequency distributions.

Performance on HyperNeRF. Our method achieves the highest PSNR of 25.78, surpassing all competing baselines including recent dynamic scene representations like E3DGS (25.43) and ADCGS (Huang et al. 2025) (25.30), as shown in Table 2, indicating superior pixel-level reconstruction fidelity. In terms of SSIM, our approach reaches a score of 0.700, which is the highest among all evaluated methods, demonstrating enhanced structural consistency and preservation of fine scene details. Although our LPIPS of 0.265 is slightly higher than HyperNeRF (Park et al. 2021b) (which obtains the lowest at 0.153), it remains competitive and notably better than most other baselines such as TiNeuVox (0.393) and 4DGS (Wu et al. 2024) (0.282). This trade-off suggests that our model effectively balances perceptual similarity while maintaining high PSNR and SSIM.

Table 2: Quantitative comparison on the HyperNeRF (Park et al. 2021b) dataset. LPIPS-A is based on Alex network.

Method	PSNR↑	SSIM↑	LPIPS-A↓
Nerfies (Park et al. 2021a)	22.23	-	0.170
HyperNeRF (Park et al. 2021b)	22.29	0.598	**0.153**
TiNeuVox (Fang et al. 2022)	24.20	0.616	0.393
EDGS (Kong, Yang, and Wang 2025)	25.70	-	-
D3DGS (Yang et al. 2024)	22.40	0.598	0.275
E3DGS (Bae et al. 2024)	25.43	0.697	0.231
ADCGS (Huang et al. 2025)	25.31	0.696	0.284
4DGS (Wu et al. 2024)	25.17	0.686	0.282
Ours	**25.78**	**0.700**	0.265

Qualitative Analysis of Novel View Synthesis

The qualitative results in Figure. 2 further highlight the advantages of our method. Compared with existing baselines, our model generates more realistic and detail-preserving images. While 4DGS (Wu et al. 2024) is generally effective, it often suffers from speckling, artifacts, and trailing effects in dynamic regions, and is highly sensitive to light reflections, resulting in significantly degraded reconstruction quality near reflective surfaces. ADCGS (Huang et al. 2025) performs poorly in reconstructing dynamic details, frequently exhibiting speckling and artifacts, and shows noticeable reconstruction failures and lighting inconsistencies in fast-moving areas, indicating considerable room for improvement in modeling instantaneous dynamics. **More comparisons and analyses are provided in the supplementary material**.

Table 3: Ablation study on the NeRF-DS (Yan, Li, and Lee 2023) dataset. LPIPS-V is computed using features from a pretrained VGG network. **More ablation studies are provided in the appendix.**

Method	PSNR↑	SSIM↑	LPIPS-V↓
w/o Position	22.93	0.804	0.216
w/o Scales	24.16	0.848	0.151
w/o Rotations	23.99	0.840	0.165
Ours Full	**24.42**	**0.856**	**0.146**

Ablation Study

To evaluate the effectiveness of each component in our framework, we conducted an ablation study on the NeRF-DS (Yan, Li, and Lee 2023) dataset, as shown in Table 3. Removing intrinsic attributes of Gaussian primitives (such as *w/o Position*, *w/o Scales*, and *w/o Rotations*) consistently degrades performance across all metrics, indicating that modeling attributes like *Position* is critical for dynamic scene reconstruction. In particular, *w/o Position* and *w/o Rotations* lead to a significant increase in LPIPS-V to **0.216** and **0.165** respectively (compared to **0.146** in the full model), suggesting a notable decline in perceptual quality.

In addition to perceptual quality, the reconstruction fidelity in terms of spatial detail and structural coherence also deteriorates when omitting these attributes. Notably, *w/o Position* fails to capture fine-grained motion trajectories, leading to spatial drift and blurred boundaries, while *w/o Scales* impairs the adaptability to multi-scale geometry variations. These findings emphasize that accurate modeling of both spatial and geometric properties is crucial for achieving high-fidelity dynamic scene reconstruction.

Conclusion

This study investigates the challenges inherent in modeling dynamic scenes that exhibit non-stationary behavior and multi-frequency motion patterns. To address the expressive limitations of conventional MLPs with fixed activation functions, we propose Sonata, a novel modeling framework for dynamic Gaussian primitives. By integrating the hierarchical modeling strengths of KANs with transformable activation mechanisms, Sonata enables more flexible and expressive representations of complex motion. Additionally, a frequency-guided learning scheme leveraging both wavelet and Fourier transforms is introduced to steer the decoupling of high- and low-frequency components, enhancing both training stability and the fidelity of fine-detail reconstruction. Experimental results across multiple monocular dynamic datasets demonstrate that Sonata consistently outperforms existing methods on quantitative benchmarks and delivers superior qualitative reconstruction, underscoring its effectiveness in dynamic scene modeling.

A.Theoretical Analysis

When evaluating whether a model possesses the ability to capture both high- and low-frequency components, observations conducted solely in the temporal or spatial domain often fail to effectively reveal its sensitivity and selectivity to spectral components. The Fourier Transform, as a tool for mapping functions from the spatial domain to the frequency domain, provides valuable insight into the frequency distribution characteristics embedded in a model's representational function. For neural networks, achieving spectral selectivity requires that a given module can actively enhance or suppress specific frequency bands through structural design or parameter modulation. Analyzing modeling capacity from a frequency-domain perspective is thus a key pathway toward understanding the model's spectral modulation mechanisms and its representational upper bound.

A.1. Frequency Representation

We adopt RBF-based Transformable Activations, formulated as follows:

$$\begin{aligned} f(x) &= \sum_{i=1}^{N} \alpha_i \cdot \psi_i(x) \\ \psi_i(x) &= \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right) \end{aligned} \tag{19}$$

Notably, the Fourier transform of each individual Gaussian RBF function also results in a Gaussian function. The Fourier transform of a single Gaussian RBF $\psi_j(x)$ is given by:

$$\mathcal{F}[\psi_j](\omega) = \sigma\sqrt{2\pi} \cdot \exp\left(-2\pi^2\sigma^2\omega^2\right) \cdot e^{-2\pi i\omega\mu_j},$$

where $\exp(-2\pi^2\sigma^2\omega^2)$ defines a Gaussian-shaped low-pass envelope, common to all kernels due to shared σ, $e^{-2\pi i\omega\mu_j}$ encodes the phase shift associated with the kernel's center location μ_j.

The overall function $f(x)$ has the Fourier transform:

$$\mathcal{F}[f](\omega) = \sqrt{2\pi}\,\sigma \exp\left(-\frac{1}{2}\sigma^2\omega^2\right) \cdot \sum_{j=1}^{N} w_j \exp(-i\omega\mu_j), \tag{20}$$

The full basis $\{\psi_i\}$, due to their shifts μ_j, introduces phase diversity via $\exp(-i\omega\mu_j)$, allowing constructive and destructive interference in the frequency domain when linearly combined. This interference pattern modulated by weights w_j forms the core mechanism for frequency shaping.

Specifically, $\mathcal{F}[f]$ consists of a fixed Gaussian envelope and a trainable trigonometric polynomial:

$$P(\omega) = \sum_{j=1}^{N} w_j e^{-i\omega\mu_j}. \tag{21}$$

This polynomial $P(\omega)$ acts as a frequency selector—its amplitude and phase can be controlled via w_j, $P(\omega)$ **is a weighted sum of complex exponentials, which is essentially equivalent to performing a discrete Fourier transform (DFT) on the weight sequence w_j. Therefore, although all Gaussian basis functions share the same spectral envelope—determined by a fixed kernel width σ—the learnable weights w_j allow flexible control over the shape of the spectral response through their combinations, enabling the synthesis of output functions with varying frequency characteristics**. Specifically, *When the weights w_j vary smoothly, the spectrum $P(\omega)$ is concentrated in the low-frequency range;When w_j exhibits rapid oscillations (e.g., alternating signs), $P(\omega)$ introduces more high-frequency components; More generally, carefully designed or optimized weight patterns can amplify or suppress specific frequency bands, thereby achieving spectral selection.*

Thus, even without explicitly introducing Fourier bases or frequency modulation terms, the structure still possesses strong capabilities for spectral modulation and selection. This is fundamentally enabled by the positional distribution of RBF kernel centers and the coherent spectral interference mechanism mechanism between their linear weights in the frequency domain, allowing the model to adaptively capture multi-frequency information and effectively represent both high-frequency details and low-frequency structures.

A.2. Coherent Spectral Interference Mechanism

The weighted sum term in the frequency-domain representation: $P(\omega) = \sum_{j=1}^{n} w_j \cdot e^{-2\pi i\omega\mu_j}$ forms a typical **spectral interference pattern**, where each term $e^{-2\pi i\omega\mu_j}$ corresponds to a phase shift in the frequency spectrum associated with a basis centered at μ_j. By adjusting the weights w_j, the model can induce constructive or destructive interference among these phase components.

More specifically, constructive interference occurs when multiple components are in phase at a given frequency ω_0, leading to additive amplification and thus enhancing that frequency band. Destructive interference arises when phase misalignment causes cancellation across components, resulting in a suppressed spectral response in that band. Through learning the weights w_j, the network enables adaptive constructive interference control in the frequency domain, allowing it to selectively enhance or suppress frequency components without modifying the kernel scale σ. This facilitates flexible spectral selection and decomposition.

This mechanism plays a pivotal role in our method: it offers a **structurally stable yet spectrally expressive** approach to frequency modeling, supporting our ability to capture both local details and global structures across multi-scale dynamic scenes.

A.3. Summary

Adaptive modulation of the kernel width σ for each RBF basis function leads to nonlinear variations in the bandwidths of the corresponding components ψ_i, thereby disrupting the coherent interference structure among basis functions in the frequency domain. This disruption

Table 4: Ablation study on the NeRF-DS (Yan, Li, and Lee 2023) dataset. The best results are marked in **bold**, while the second-best are highlighted with a gray background. LPIPS-V is computed using features from a pretrained VGG network.

Method	PSNR↑	SSIM↑	LPIPS-V↓
w/ MLP	22.87	0.802	0.214
w/o Freq Loss	23.89	0.839	0.160
w/o Depth Loss	24.29	0.852	0.151
Ours Full	**24.42**	**0.856**	**0.146**

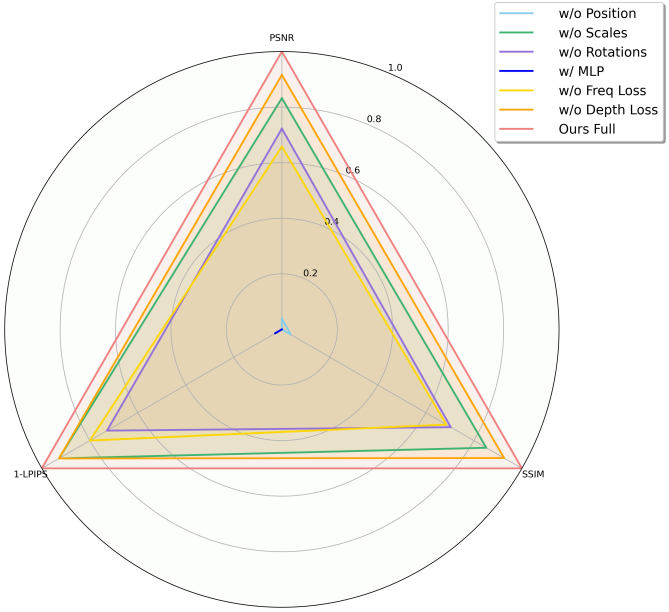

Figure 3: Radar chart showing the effects of ablating key components on overall reconstruction quality.

results in irregular spectral overlaps, reducing the clarity and decoupling of the frequency structure, and ultimately impairing the model's ability to identify local frequency components. Specifically, if the kernel widths σ_i are inconsistent (i.e., each basis function has a different bandwidth), the overlap regions in the frequency domain become irregular. Such disorder in spectral overlap undermines the composability of different frequency responses, leading to blurred separation between high and low frequencies, weakened frequency decoupling, and the breakdown of constructive or destructive interference patterns—thus degrading the model's sensitivity to frequency variations.

In contrast, using a set of RBF functions with fixed kernel widths and uniformly distributed centers results in Gaussian functions that share the same shape but are centered at different input locations. In the spectral domain, these give rise to a frequency-localized and phase-rich basis system, where each basis contributes a shared magnitude envelope but distinct phase shift. Within this structure, frequency modulation is entirely governed by the learnable linear weights w_j, which directly influence the spectral response $P(\omega)$ (as analyzed in Appendix A.1). These weights regulate the superposition of Gaussian spectral components with identical envelopes but varying phase offsets, and through coherent interference, they effectively enhance or suppress responses in specific frequency regions, thereby enabling explicit frequency selection.

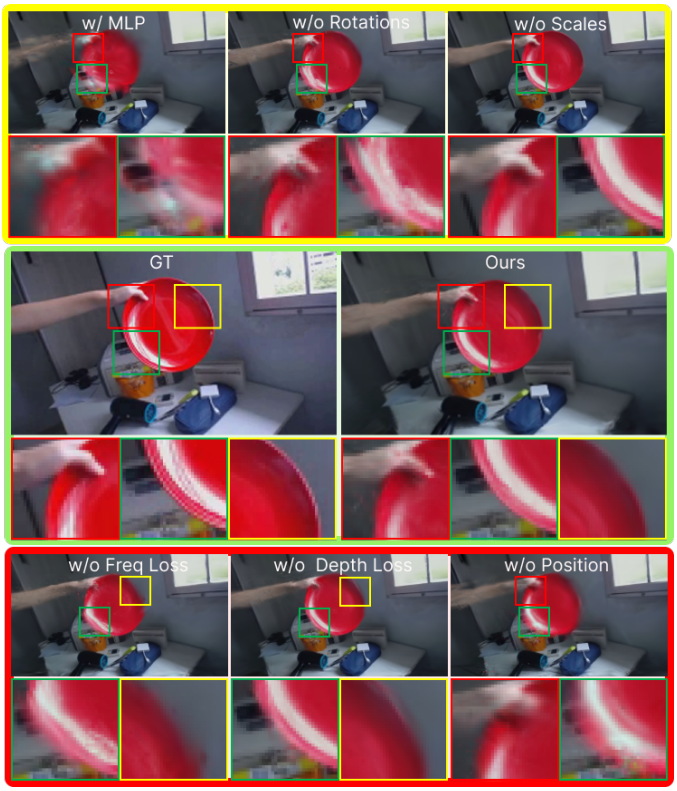

Figure 4: Visual Comparisons on the *Plate* Subset of the NeRF-DS (Yan, Li, and Lee 2023) Dataset Highlighting the Impact of Module Ablations

This design preserves the **stability** of RBF representations while ensuring **disentanglement in the frequency space**, which is critical for modeling multi-scale and hierarchical motion structures.

Moreover, we incorporate a frequency-guided training strategy by introducing supervisory signals from both the wavelet and Fourier domains. These frequency-aware cues guide the transformation field to effectively distinguish between high- and low-frequency components during training, helping the model avoid suboptimal local minima and early-stage instability, while enhancing overall spectral sensitivity and improving convergence robustness.

B. Extended Ablation Results

To further assess the contribution and necessity of each design component in our framework, we present additional ablation experiments in both quantitative and qualitative forms.

As shown in Table 4 and Figure 3, 4, these extended evaluations offer deeper insights into how individual modules affect the model's performance across various metrics and visual outcomes. This comprehensive analysis not only reinforces the effectiveness of our core components but also highlights their complementary roles in enhancing the reconstruction quality of dynamic scenes.

B.1 Quantitative Results of Ablation Studies

We conduct ablation experiments on the NeRF-DS (Yan, Li, and Lee 2023) dataset to evaluate the individual contribution of each component in our framework. The results are presented in Table 4. Specifically, *"w/ MLP"* denotes using a standard MLP (Yang et al. 2024) as the transformation field, *"w/o Freq Loss"* refers to removing the frequency-aware losses (including both wavelet and Fourier losses), and *"w/o Depth Loss"* indicates the exclusion of the depth supervision term.

For *"w/ MLP"*: When using an MLP as the transformation field, PSNR drops to 22.87 while LPIPS-V increases to 0.214, confirming that lightweight MLPs, due to their fixed activation functions and limited adaptability, underperform in capturing high-frequency details. For *"w/o Freq Loss"*: Removing the frequency loss leads to a decrease in PSNR (-0.53) and SSIM (-0.017), and an increase in LPIPS-V (+0.014) compared to our full model. This demonstrates that the proposed frequency loss plays a guiding role within the Sonata framework, effectively encouraging frequency disentanglement and helping the model avoid local minima. For *"w/o Depth Loss"*: Removing the depth loss results in a 0.13 drop in PSNR compared to the full model, suggesting that depth cues contribute meaningfully to our framework and serve as a key driver for geometric accuracy.

B.2 Multi-Dimensional Performance Analysis

To comprehensively evaluate and compare the impact of each model component, we conduct a multi-dimensional analysis based on key perceptual metrics, including PSNR, SSIM, and LPIPS (Zhang et al. 2018). Given the diversity and scale differences among these metrics, all values are first normalized to a common range to ensure fair comparison. To intuitively illustrate the performance differences across ablation variants, we adopt a radar chart (as shown in Figure 3), which offers a unified view highlighting the strengths and weaknesses of each design choice. This analysis clearly demonstrates the effectiveness of our full model, which consistently outperforms others across all evaluation dimensions.

As shown in the radar chart in Figure 3, all metrics are normalized to the [0, 1] range, where values closer to the outer ring (1.0) indicate better performance, and those nearer to the center (0.0) indicate worse performance. The overall trends are consistent with the quantitative results reported in Tables 3 and 3, and lead to the following observations:

High-frequency fidelity:Along the 1-LPIPS axis, *Ours Full* reaches the outermost ring (1.0), while *w/ MLP* and *w/o Position* are noticeably contracted inward, indicating that using a standard MLP (Yang et al. 2024) or removing positional encoding significantly degrades the reconstruction of high-frequency details, leading to a clear collapse of fine structures.

Geometric accuracy:In both the PSNR and SSIM directions, *w/o Rotations* and *w/o Freq Loss* deviate substantially from the outer ring, suggesting that rotation modeling and frequency loss are equally crucial for accurate geometry reconstruction—more so than *w/o Depth Loss* and *w/o Scales*. Additionally, the result's proximity to the center further underscores that both positional encoding and our constructed transformation field are central to precise geometric modeling.

Component complementarity: Comparing the six ablated variants with *Ours Full* reveals that removing any individual component results in a visible "dent" along at least one axis. For example, *w/o Freq Loss* exhibits a significant drop along the PSNR axis, while *w/o Scales* shows a slight reduction in SSIM. This indicates that each loss term and architectural element plays a non-redundant, complementary role in jointly enhancing frequency modeling and geometric alignment.

B.3 Qualitative Results and Detailed Visual Comparisons of Ablation Studies

In addition to numerical comparisons, we present qualitative ablation results to visually assess the impact of each component. For clearer insight, we further highlight fine-grained differences by zooming in on critical regions with rich structural or textural details. These comparisons reveal that our full model reconstructs sharper boundaries and more faithful geometry, especially in challenging dynamic scenarios. As shown in Figure 4:

For *w/ MLP*, the red "plate" shows noticeable edge blurring and visible artifacts; the hand region also suffers from detail loss, and the green box highlights further artifacts. The plate's edge is heavily blurred. When replacing our transformable activation functions with standard MLPs, the lack of frequency awareness in MLPs leads to significant high-frequency attenuation, resulting in blurred reconstructions and loss of fine details.

For *w/o Rotations*, the red box reveals inaccurate hand details, while both the red and green boxes show deformation around the plate edges. Removing the rotation parameters causes the model to fail at aligning orientation-dependent features, leading to geometric distortion. This degradation is especially evident in directional structures such as the plate boundary.

For *w/o Scales*, the green box reveals blurred plate edges. In the red box, the reconstruction around the hand-plate contact area is distorted, and the plate edge appears slightly blurred. Without scale modulation, the model loses its ability to capture multi-scale structures, resulting in degraded performance in complex regions such as the hand-plate interaction and the plate thickness.

For *w/o Freq Loss*, the green box highlights severe blurring at the plate edge, especially in reflective regions. In the zoomed-in view, heavy frequency aliasing is observed. The yellow box also shows slight blurring, and the plate contours are not clearly visible. Without frequency loss, the model lacks spectral guidance, making it prone to local min-

ima during frequency decoupling, leading to collapsed high-frequency details.

For *w/o Depth Loss*, minor geometric deviations appear, with softened plate contours and blurred reconstructions. Some regions show slight deformation. Removing depth supervision reduces the model's capability to learn accurate geometry, especially around object boundaries and hierarchical structures.

For *w/o Position*, the moving region of the image becomes significantly aliased, with clear contour distortion and blurring along the plate edges. Some areas even show color misalignment. Removing positional encoding weakens the model's spatial representation, leading to poor scene alignment and structural degradation.

For *Ours*, the boundaries are very close to the ground truth. The texture and edges of the red plate are well preserved, and hand details remain clearly visible. This demonstrates that our proposed model achieves superior performance in both geometric and frequency modeling, effectively restoring fine details and complex structures in dynamic scenes.

C. Zoom-in Visual Comparison of Reconstruction Details

To further demonstrate the effectiveness of our method, we provide additional qualitative results, with a focus on fine-grained detail reconstruction and subtle illumination variations. While the main paper highlights overall scene reconstruction quality, this section emphasizes the model's ability to preserve high-frequency structures such as sharp edges, fine textures, and motion boundaries. **These visualizations showcase how the transformation field in our approach can dynamically adjust its spectral response during training, enabling the model to better capture both low-frequency structures and high-frequency details. Meanwhile, the frequency-aware guidance introduced by wavelet and Fourier transforms helps prevent the model from falling into local minima**. Compared to existing methods, our approach consistently produces sharper details that are closer to real images and achieves more coherent illumination reconstruction, particularly in challenging dynamic regions.

C.1 Qualitative Results on NeRF-DS

As shown in Figure 5, our method achieves superior performance in detail reconstruction, significantly outperforming existing approaches. In contrast, ADCGS (Huang et al. 2025) and 4DGS (Wu et al. 2024) suffer from various reconstruction issues, including illumination inconsistencies, artifacts, and geometric distortions.

First row: Overall, both D3DGS (Yang et al. 2024) and 4DGS (Wu et al. 2024) exhibit subpar texture reconstruction quality. In the red box (hand region), 4DGS (Wu et al. 2024) and ADCGS (Huang et al. 2025) struggle with fine-detail preservation, resulting in noticeable blotchiness and blurring. In the yellow box (thin-sheet region), ADCGS (Huang et al. 2025) fails to generalize to reflective surfaces, leading to severe blurring of background pixels behind the sheet. While 4DGS (Wu et al. 2024) partially recovers the scene, it exhibits considerable geometric artifacts and structural distortions. In contrast, our method effectively suppresses artifacts and blotchiness on reflective materials while accurately reconstructing fine details and preserving scene geometry.

Second row: In both the red and yellow boxes, visible blotches and artifacts appear in 4DGS (Wu et al. 2024) and ADCGS (Huang et al. 2025) results. ADCGS (Huang et al. 2025) , in particular, shows the densest noise, indicating instability in its implicit representation under high-frequency conditions. 4DGS (Wu et al. 2024) exhibits fewer artifacts but still suffers from reconstruction deficiencies and noticeable edge noise. In comparison, our results show no visible noise, producing smooth surfaces with sharp edges.

Third row: For a challenging scene involving a specular object, 4DGS (Wu et al. 2024) presents blotches and heavily blurred highlights in the red and yellow boxes, indicating poor generalization to complex lighting and specular materials. ADCGS (Huang et al. 2025) also shows blotches (in both boxes) and lighting distortions (in the yellow box), leading to distorted reflections. Our method, however, produces specular highlights and surrounding details that appear much closer to reality.

Fourth row: In the red box (plate), both 4DGS (Wu et al. 2024) and ADCGS (Huang et al. 2025) produce large blotches and trailing artifacts, along with missing geometry around the plate rim. The yellow box also shows noticeable blotches and reconstruction distortions. Our method not only achieves superior geometric completeness but also delivers texture and lighting consistency that closely matches the ground truth.

C.2 Qualitative Results on HyperNeRF

We conduct a qualitative analysis on the HyperNeRF (Park et al. 2021b) dataset, focusing on two subsets: Chicken and Banana, as visualized in Figure 6.

In the Chicken subset, all methods achieve high overall reconstruction quality and successfully recover fine details of the target. In the red box region, all approaches demonstrate satisfactory texture restoration. However, 4DGS exhibits blurry reconstruction in the right concavity, while ADCGS shows slight blurring around texture edges. In the yellow box region, 4DGS (Wu et al. 2024) suffers from insufficient sharpness and minor surface detail blurring. Although ADCGS (Huang et al. 2025) shows structural blurriness at the initial protrusion and diffused lighting, our method produces more focused highlights and sharper structural features.

In the Banana scene, the dynamic region poses greater reconstruction challenges. Both 4DGS (Wu et al. 2024) and ADCGS (Huang et al. 2025) show severe artifacts and blurring. In the red box, hand reconstruction significantly degrades in both methods: ADCGS (Huang et al. 2025) completely loses the finger structure, retaining only blurred contours of the skin, while 4DGS (Wu et al. 2024) shows slightly better performance but still lacks discernible hand textures. The motion boundaries lack clear and sharp transitions. In the yellow box, neither method successfully reconstructs the inner banana peel's visible ridges—its texture

Figure 5: Qualitative comparisons on the NeRF-DS (Yan, Li, and Lee 2023) dataset, highlighting the reconstructed image details achieved by different algorithms.

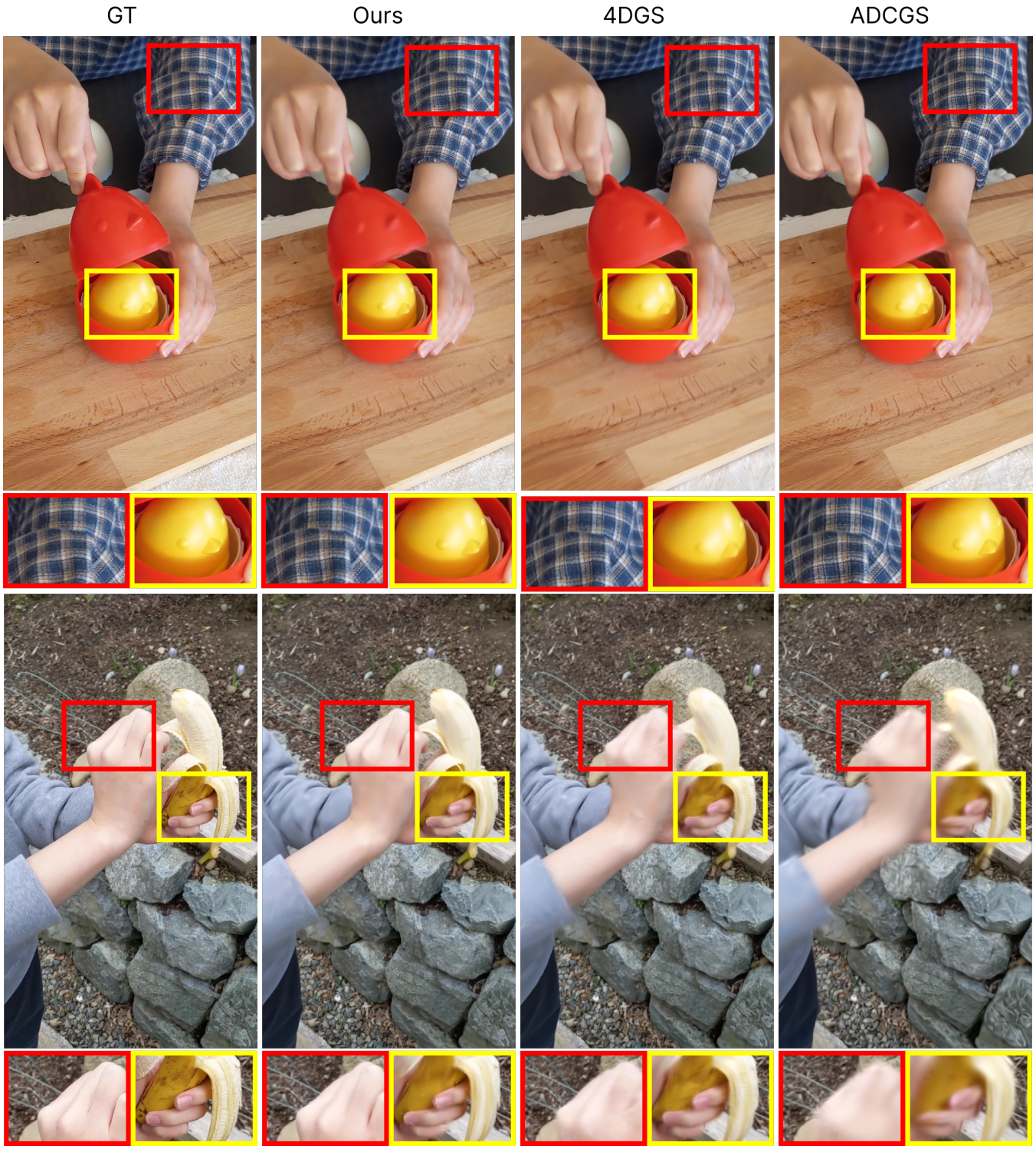

Figure 6: Qualitative comparisons on the hypernerf (Park et al. 2021b) dataset, highlighting the reconstructed image details achieved by different algorithms.

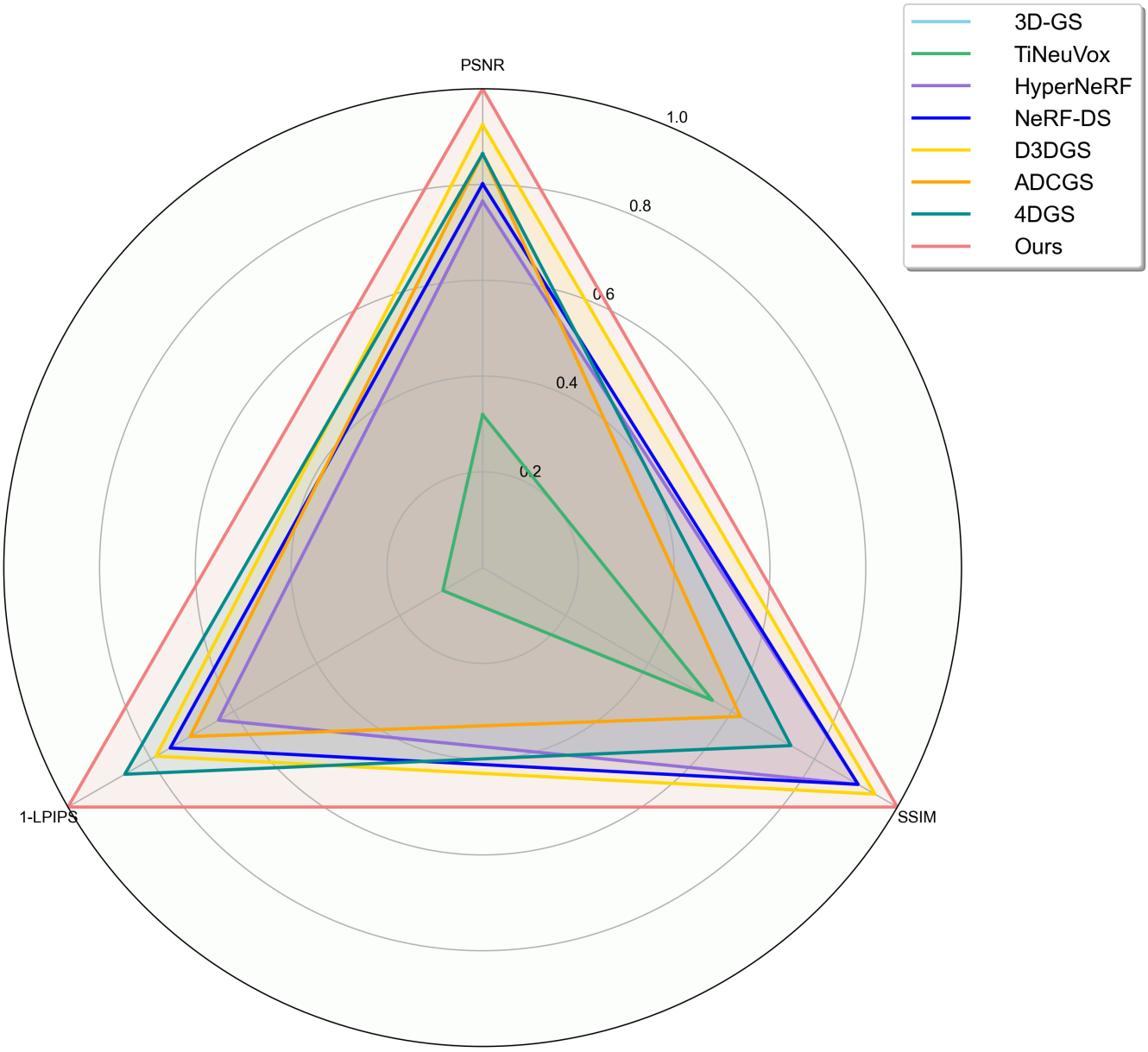

Figure 7: Multi-dimensional Radar Chart of NeRF-DS (Yan, Li, and Lee 2023) dataset Performance

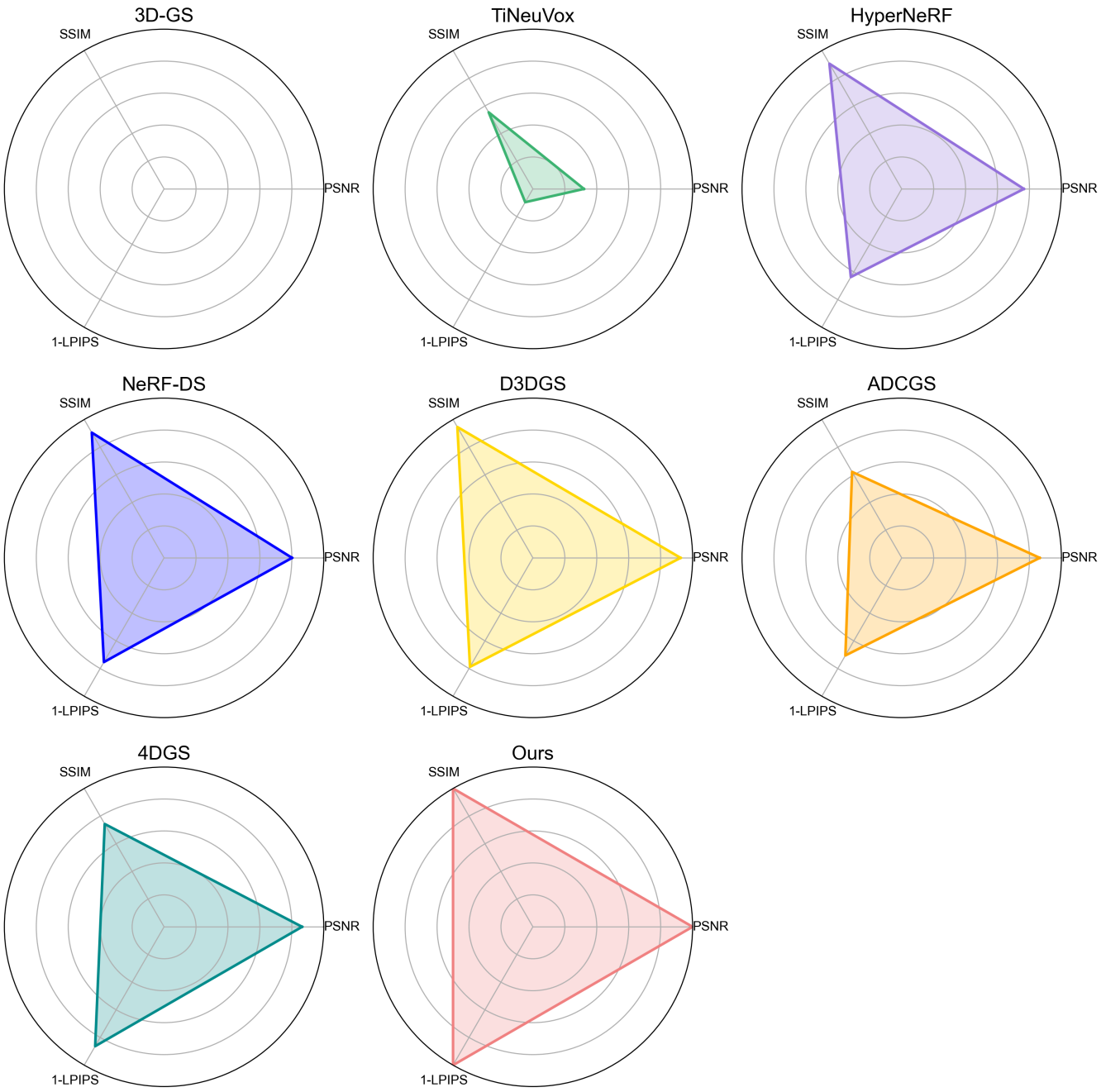

Figure 8: Individual Radar Chart Visualization for Each Model on NeRF-DS (Yan, Li, and Lee 2023) dataset

appears blurry in both. ADCGS (Huang et al. 2025) deteriorates more drastically, with nearly complete edge distortion and indistinct motion boundaries. While 4DGS (Wu et al. 2024) retains some structural information, the overall result remains hazy.

D. Multi-Dimensional Performance Analysis

To gain a comprehensive understanding of model performance across multiple perceptual dimensions, we construct radar charts based on three representative metrics: PSNR, SSIM, and the perceptual LPIPS score (transformed to $1 -$ LPIPS to ensure directional consistency). All metric values are normalized to the $[0, 1]$ range using min-max normalization to allow for fair comparison across scales. The radar charts are generated in two formats: a global chart showing all models overlaid for holistic comparison, and individual charts for per-model analysis, enabling fine-grained inspection of each variant. We conduct a multi-dimensional analysis using the following representative models: 3D-GS (Kerbl et al. 2023), TiNeuVox (Fang et al. 2022), HyperNeRF (Park et al. 2021b), NeRF-DS (Yan, Li, and Lee 2023), D3DGS (Yang et al. 2024), ADCGS (Huang et al. 2025), and 4DGS (Wu et al. 2024).

In the overall radar chart in Figure 7, our proposed method ("Ours") forms the largest enclosed area among all models, indicating superior performance across all three dimensions. Notably, while several methods such as D3DGS (Yang et al. 2024), 4DGS (Wu et al. 2024), and ADCGS (Huang et al. 2025) approach comparable scores in one or two metrics, they fall short in forming a balanced triangle across all dimensions. For instance, HyperNeRF shows strong SSIM performance but lags in LPIPS and PSNR . The visual distinction of triangle sizes and shapes offers an intuitive view of trade-offs between perceptual quality and pixel-level fidelity across the competing methods.

Figure 8 presents individual radar charts for each model. Notably, 3D-GS is mapped entirely to the origin, as it yields the lowest score on all metrics and is therefore normalized to zero across all dimensions. These charts further highlight the specific strengths and weaknesses of each method. For instance, TiNeuVox shows uniformly low performance across all metrics, indicating limited effectiveness on this dataset. In contrast, D3DGS (Yang et al. 2024) and ADCGS (Huang et al. 2025) exhibit relatively balanced yet moderate performance profiles, forming nearly equilateral but smaller triangles. Our method ("Ours") consistently dominates across all three axes, reaffirming the robustness and effectiveness of the proposed design. These per-model visualizations not only support the trends observed in the global chart but also help isolate and interpret the impact of architectural differences among competing approaches.

References

Bae, J.; Kim, S.; Yun, Y.; Lee, H.; Bang, G.; and Uh, Y. 2024. Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In *European Conference on Computer Vision*, 321–335. Springer.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5855–5864.

Cao, A.; and Johnson, J. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 130–141.

Fang, J.; Yi, T.; Wang, X.; Xie, L.; Zhang, X.; Liu, W.; Nießner, M.; and Tian, Q. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.

Gordon, W. J.; and Riesenfeld, R. F. 1974. B-spline curves and surfaces. In *Computer aided geometric design*, 95–126. Elsevier.

Guo, X.; Sun, J.; Dai, Y.; Chen, G.; Ye, X.; Tan, X.; Ding, E.; Zhang, Y.; and Wang, J. 2023. Forward flow for novel view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16022–16033.

Hu, T.; Liu, S.; Chen, Y.; Shen, T.; and Jia, J. 2022. Efficientnerf efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12902–12911.

Huang, H.; Yang, Q.; Liu, M.; Xu, Y.; and Li, Z. 2025. ADCGS: Anchor-Driven Deformable and Compressed Gaussian Splatting for Dynamic Scene Reconstruction. *arXiv preprint arXiv:2505.08196*.

Huang, Y.-H.; Sun, Y.-T.; Yang, Z.; Lyu, X.; Cao, Y.-P.; and Qi, X. 2024. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4220–4230.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kong, H.; Yang, X.; and Wang, X. 2025. Efficient gaussian splatting for monocular dynamic scene rendering via sparse time-variant attribute modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4374–4382.

Li, Y.; Fang, Y.; Tombari, F.; and Lee, G. H. 2024a. SmileSplat: Generalizable Gaussian Splats for Unconstrained Sparse Images. *arXiv preprint arXiv:2411.18072*.

Li, Y.; Lyu, C.; Di, Y.; Zhai, G.; Lee, G. H.; and Tombari, F. 2024b. Geogaussian: Geometry-aware gaussian splatting for scene rendering. In *European Conference on Computer Vision*, 441–457. Springer.

Li, Z.; Chen, Z.; Li, Z.; and Xu, Y. 2024c. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8508–8520.

Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Han, J. 2020. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations, ICLR 2020*.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Luiten, J.; Kopanas, G.; Leibe, B.; and Ramanan, D. 2024. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, 800–809. IEEE.

Park, J.; Bui, M.-Q. V.; Bello, J. L. G.; Moon, J.; Oh, J.; and Kim, M. 2025. Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26866–26875.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021a. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5865–5874.

Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021b. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Shao, R.; Zheng, Z.; Tu, H.; Liu, B.; Zhang, H.; and Liu, Y. 2023. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16632–16642.

Wang, F.; Tan, S.; Li, X.; Tian, Z.; Song, Y.; and Liu, H. 2023a. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19706–19716.

Wang, Z.; Shen, T.; Nimier-David, M.; Sharp, N.; Gao, J.; Keller, A.; Fidler, S.; Müller, T.; and Gojcic, Z. 2023b. Adaptive shells for efficient neural radiance field rendering. *arXiv preprint arXiv:2311.10091*.

Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20310–20320.

Yan, Z.; Li, C.; and Lee, G. H. 2023. Nerf-ds: Neural radiance fields for dynamic specular objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8285–8295.

Yang, Z.; Gao, X.; Zhou, W.; Jiao, S.; Zhang, Y.; and Jin, X. 2024. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20331–20341.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.