VRVVC: Variable-Rate NeRF-Based Volumetric Video Compression

Qiang Hu[1*], Houqiang Zhong[2*], Zihan Zheng[1], Xiaoyun Zhang[1†], Zhengxue Cheng[2], Li Song[2], Guangtao Zhai[2], Yanfeng Wang[3]

[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[3]School of Artificial Intelligence, Shanghai Jiao Tong University
{qiang.hu,zhonghouqiang,1364406834,xiaoyun.zhang,zxcheng,song_li,zhaiguangtao,wangyanfeng}@sjtu.edu.cn

Abstract

Neural Radiance Field (NeRF)-based volumetric video has revolutionized visual media by delivering photorealistic Free-Viewpoint Video (FVV) experiences that provide audiences with unprecedented immersion and interactivity. However, the substantial data volumes pose significant challenges for storage and transmission. Existing solutions typically optimize NeRF representation and compression independently or focus on a single fixed rate-distortion (RD) tradeoff. In this paper, we propose VRVVC, a novel end-to-end joint optimization variable-rate framework for volumetric video compression that achieves variable bitrates using a single model while maintaining superior RD performance. Specifically, VRVVC introduces a compact tri-plane implicit residual representation for inter-frame modeling of long-duration dynamic scenes, effectively reducing temporal redundancy. We further propose a variable-rate residual representation compression scheme that leverages a learnable quantization and a tiny MLP-based entropy model. This approach enables variable bitrates through the utilization of predefined Lagrange multipliers to manage the quantization error of all latent representations. Finally, we present an end-to-end progressive training strategy combined with a multi-rate-distortion loss function to optimize the entire framework. Extensive experiments demonstrate that VRVVC achieves a wide range of variable bitrates within a single model and surpasses the RD performance of existing methods across various datasets.

Introduction

Photorealistic volumetric video provides an immersive experience in virtual reality and telepresence, demonstrating significant potential to become the next-generation video format. Traditional approaches to volumetric video reconstruction have primarily relied on point cloud-based methods (Graziosi et al. 2020) and depth-based techniques (Boyce et al. 2021), which often hinder realistic rendering quality. Recently, both Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) have shown considerable promise in representing photorealistic volumetric video. However, challenges remain in the storage and transmission of volumetric video with NeRF or 3DGS. Compared to NeRF, 3DGS utilizes an explicit point cloud representation, which is less conducive to efficient compression. In summary, NeRF's compact representation and implicit modeling capabilities make it inherently suitable for volumetric video compression.

NeRF and its variants (Müller et al. 2022; Reiser et al. 2023) have achieved remarkable success in synthesizing novel views, inspiring a multitude of derivative research studies focused on dynamic scenes. Some techniques (Park et al. 2021; Pumarola et al. 2020) employ deformation fields to capture voxel movements relative to a canonical space, while others (Fang et al. 2022; Işık et al. 2023; Fridovich-Keil et al. 2023) introduce temporal voxel features or apply joint training across multiple frames to achieve superior temporal reconstructions. However, most existing studies primarily focus on improving the reconstruction quality of NeRF representations, frequently neglecting the critical need to minimize storage size and transmission bandwidth. This oversight poses substantial challenges for practical applications, especially in streaming volumetric video.

To address these problems, several approaches are proposed to compress explicit features of dynamic NeRF. For instance, ReRF (Wang et al. 2023) uses a grid-based explicit representation to model the spatial-temporal feature space of dynamic scenes and adopts traditional image encoding techniques to compress the representation after training. TeTriRF (Wu et al. 2024) utilizes a hybrid representation with tri-plane to model dynamic scenes and employs a traditional video codec to reduce redundancy. However, these methods optimize representation and compression independently, neglecting the rate-distortion (RD) tradeoff during the training phase, which limits their compression performance. To close this gap, JointRF (Zheng et al. 2024b) introduces an end-to-end combined training approach for dynamic NeRF representation and compression, but it is fixed-rate only and suffers from slow rendering speed.

In this paper, we propose VRVVC, a novel variable-rate compression framework tailored for NeRF-based volumetric video. Our key idea involves estimating the bitrate of NeRF representations during end-to-end training and controlling it using the RD tradeoff parameter λ. By incorporating both bitrate and distortion terms into the loss function, we achieve optimal RD performance across a wide range of variable bitrates using a single model, as illustrated in Fig.

[*]These authors contributed equally.
[†]Corresponding author.

Figure 1: **Left**: Our proposed VRVVC efficiently compresses volumetric video at variable bitrates using a single model. **Middle**: We demonstrate two examples of reconstruction quality at a bitrate of 60 KB per frame. **Right**: The RD performance of our approach surpasses prior work (e.g. ReRF (Wang et al. 2023), TeTriRF (Wu et al. 2024))

1. We realize this through three main innovations. First, we introduce a compact tri-plane implicit residual representation for inter-frame modeling within long sequences. For each frame, VRVVC decomposes the radiance field into a tri-plane and models the residual information between adjacent timestamps within this feature space. This representation effectively captures high-dimensional appearance features within compact planes.

Second, we propose a variable-rate residual representation compression scheme that leverages a learnable quantization step and a tiny MLP-based entropy model, combined with a predefined set of Lagrange multipliers, to facilitate variable bitrates. Third, we present an end-to-end progressive learning scheme to jointly optimize both the representation and compression. This approach yields temporally consistent and low-entropy 4D sequential representations that can be effectively compressed, significantly enhancing RD performance. Experimental results show that our VRVVC achieves variable bitrates by a single model while maintaining state-of-the-art RD performance across various datasets. Compared to the previous leading method, TeTriRF (Wu et al. 2024), our approach achieves approximately **-81%** BD-rate savings on the DNA-Rendering (Cheng et al. 2023) dataset and an **-46%** BD-rate reduction on the ReRF dataset.

In summary, our contributions are as follows:

- We propose VRVVC, a novel approach for variable-rate compression of NeRF-based volumetric video. Our VRVVC achieves variable bitrates within a single model while delivering improved RD performance.
- We introduce a compact and compression-friendly representation that models volumetric video as a tri-plane residual radiance field, effectively minimizing temporal redundancy for inter-frame modeling of extended dynamic scenes.
- We present an end-to-end progressive training scheme that jointly optimizes representation and compression through a multi-rate-distortion loss function, significantly improving compression performance compared to post-training methods.

Related Work

Dynamic Radiance Field Representation. NeRF (Mildenhall et al. 2021) employs implicit representations to synthesize realistic novel views. Its advancements (Müller et al. 2022; Rabich, Stotko, and Klein 2024; Martin-Brualla et al. 2021; Barron et al. 2021, 2022) in static scenes have catalyzed research into dynamic scenes, particularly in volumetric video. Deformation field (Du et al. 2021; Li et al. 2022b; Pumarola et al. 2020; Song et al. 2023) recovers temporal features by warping real-time frames to a canonical space. However, these methods struggle with large motions and deformations, leading to slower training and rendering. Conversely, other approaches (Fang et al. 2022; Işık et al. 2023; Fridovich-Keil et al. 2023; Cao and Johnson 2023a; Li et al. 2022a; Shao et al. 2023) extend the radiance field into a 4D spatio-temporal domain, facilitating faster training and rendering at the cost of increased storage demands. Several studies (Wang et al. 2023, 2024; Wu et al. 2024; Zheng et al. 2024b,a) use residual radiance fields to represent long-sequence dynamic scenes, leveraging compact motion grids and residual feature grids to exploit inter-frame feature similarity. Our compact tri-plane residual-based dynamic modeling method is designed for inter-frame modeling in extended sequences, which effectively captures high-dimensional appearance features within compact planes.

NeRF Compression. Recently, deep learning-based image and video compression methods have demonstrated strong RD performance for 2D video (Lu et al. 2024a; Ballé,

Laparra, and Simoncelli 2016; Ballé et al. 2018; Ballé, Laparra, and Simoncelli 2017; Guo et al. 2020; Choi, El-Khamy, and Lee 2019; Cui et al. 2021; Lu et al. 2024b, 2022). Efforts are now being made to extend these compression techniques to the NeRF domain (Li et al. 2023; Lee et al. 2023; Peng et al. 2023; Rho et al. 2023). VQRF (Li et al. 2023) and ECRF (Lee et al. 2023) have made strides by employing entropy encoding and frequency domain mapping, respectively, for compressing static radiance fields. However, these methods are limited to static scenes and do not address dynamic scenarios. Recent studies like ReRF, VideoRF (Wang et al. 2024), and TeTriRF (Wu et al. 2024) focus on dynamic scenes. They integrate traditional image and video encoding techniques for feature compression but fail to jointly optimize the representation and compression of the radiance field, resulting in a loss of dynamic details and compression efficiency. Our approach estimates the bitrate of representations during training and controls it using the RD tradeoff parameter λ, enabling end-to-end training. This allows our model to achieve a wide range of variable bitrates, unlike JointRF, which is restricted to a fixed bitrate.

Method

In this section, we introduce the details of the proposed VRVVC. Fig. 2 illustrates the overall framework of our method. We model the inter-frame relationships of long dynamic scenes using a compact tri-plane residual representation. Additionally, we propose a variable-rate entropy coding scheme to achieve a wide range of variable bitrates within a single model. We also introduce a fast progressive training strategy that jointly optimizes representation and compression, greatly improving compression efficiency while preserving high rendering quality.

Tri-plane Residual Dynamic Modeling

Recall that a NeRF models a 3D volumetric scene using a 5D function Ψ, which maps the spatial coordinate $\mathbf{x} = (x, y, z)$ and view direction $\mathbf{d} = (\theta, \phi)$ to color $\mathbf{c}$ and density σ, formulated as $(\mathbf{c}, \sigma) = \Psi(\mathbf{x}, \mathbf{d})$. Then, volume rendering is employed for photo-realistic novel view synthesis. To enhance training and rendering efficiency, we employ a feature tri-plane $\mathbf{P} = \{\mathbf{P}^l \mid l \in L\}$, $L = \{xy, yz, xz\}$ along with a 3D density grid $\mathbf{V}$ as our static representation $\mathbf{F} = (\mathbf{P}, \mathbf{V})$. Specifically, the radiance field of a static scene is:

$$\begin{aligned} \mathbf{f} &= \bigcap_{l \in L} \varphi\left(\pi_l(\mathbf{x}, \mathbf{P}^l)\right) \\ \mathbf{c} &= \Phi(\mathbf{f}, \omega(\mathbf{d})) \\ \sigma &= \varphi\left(\mathbf{x}, \mathbf{V}\right) \end{aligned} \tag{1}$$

where φ denotes the interpolation function, π_l projects the 3D point $\mathbf{x}$ onto feature plane l, and $\bigcap$ represents concatenating the features from three planes. The MLP Φ decodes the color at point $\mathbf{x}$ based on the concatenated feature $\mathbf{f}$ and the encoded view direction $\omega(\mathbf{d})$. The density of point $\mathbf{x}$ is derived through interpolation on the density grid.

When expanding from static to dynamic scenes, a straightforward approach is to utilize individual per-frame features to represent a dynamic scene composed of M frames, denoted as $\{\mathbf{F}_t\}_{t=1}^{M}$. However, this approach neglects temporal coherence, resulting in substantial temporal redundancy. Conversely, other methods (Cao and Johnson 2023b; Fridovich-Keil et al. 2023) that directly model entire dynamic scenes using NeRF representation may lead to suboptimal performance for long sequences and are unsuitable for streaming applications. To address these challenges, we extend the current static NeRF representation to dynamic scenes by employing a frame-by-frame tri-plane residual inter-frame modeling strategy.

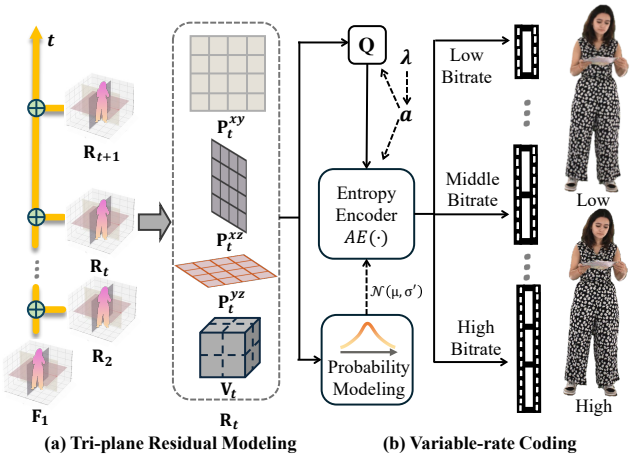

Figure 2: Illustration of our VRVVC framework. We employ a compact tri-plane residual representation for inter-frame modeling of long-duration dynamic scenes. The residuals are encoded into several bitstreams in an MLP-based entropy model that utilizes the RD tradeoff parameter λ to achieve variable bitrates within a single model.

Our tri-plane residual modeling method divides the sequence into equal-length groups of features (GoFs), each containing N frames. In each GoF, the first frame is modeled independently as an I-feature $\mathbf{F}_1$, and the subsequent frames are P-features $\{\mathbf{R}_t\}_{t=2}^{N}$, representing the residuals relative to the preceding frame. Frames within the same GoF share a compact global MLP Φ as the feature decoder, reducing bitrate consumption while maintaining performance. Finally, our VRVVC represents a GoF with N frames as Φ and $\mathbf{G} = \{\mathbf{F}_1, \mathbf{R}_2, \cdots \mathbf{R}_t \cdots \mathbf{R}_N\}$, as shown in Fig. 2.

Our VRVVC enables highly efficient sequential modeling of P-features by leveraging inter-frame feature similarities. Specifically, we retrieve the reconstructed feature of the previous frame $\hat{\mathbf{F}}_{t-1}$ from the decoded buffer and combine it with the input images of the current frame to learn the residual for the current frame $\mathbf{R}_t$, as shown in Fig. 3. Then, we can reconstruct the entire feature of the current frame $\hat{\mathbf{F}}_t$ by applying the residual compensation:

$$\begin{aligned} \hat{\mathbf{F}}_t &= \hat{\mathbf{F}}_{t-1} + \hat{\mathbf{R}}_t \\ &= (\bigcup_{l \in L} (\hat{\mathbf{P}}^l_{t-1} + \hat{\mathbf{R}}^l_t), \hat{\mathbf{V}}_{t-1} + \hat{\mathbf{R}}^\sigma_t) \end{aligned} \tag{2}$$

where $\bigcup$ represents the union of tri-plane features, and $\hat{\mathbf{R}}_t = \{\hat{\mathbf{R}}^{xy}_t, \hat{\mathbf{R}}^{yz}_t, \hat{\mathbf{R}}^{xz}_t, \hat{\mathbf{R}}^{\sigma}_t\}$ denotes the reconstruction residual

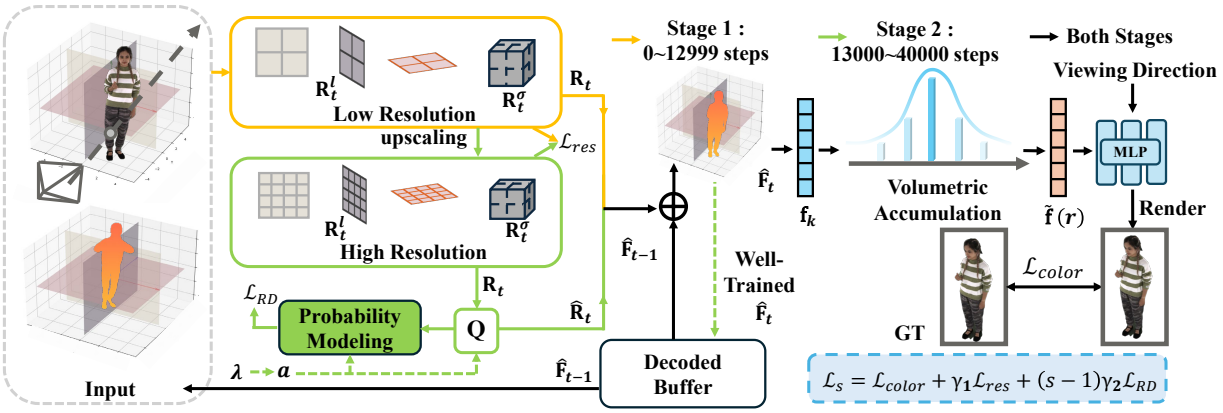

Figure 3: Overview of our progressive training. In the first stage, we adopt the reconstructed features $\hat{\mathbf{F}}_{t-1}$ from the previous frame, retrieved from the decoded buffer, to train the current frame's low-resolution residual features. In the second stage, these features are reused as an effective initialization for further training, where they are integrated with a variable-rate entropy coding model for joint optimization. The entire training process is supervised by the multi-rate-distortion loss $\mathcal{L}_s$.

for tri-plane and density grid. Finally, $\hat{\mathbf{F}}_t$ is stored in the decode buffer for the reconstruction of the next frame.

Our tri-plane residual representation offers several key advantages. Firstly, it is both effective and compact, capable of capturing high-dimensional appearance features and decomposing them into three orthogonal feature planes. Secondly, it is highly compression-friendly, as it leverages the simplicity of the residual data distribution to efficiently reduce spatio-temporal redundancy between frames. Thirdly, it facilitates efficient training and rendering by incorporating an explicit density grid, which enables rapid retrieval of density values. This allows for the swift removal of sample points in empty space without the need for network inference, thus accelerating both training and inference processes.

Variable-rate Entropy Coding

We also propose a variable-rate entropy coding scheme for residual representation, enabling flexible adjustments between different bitrates and reconstruction quality within a single model. Unlike traditional methods (Yang et al. 2020; Lin et al. 2021) that adjust the interval of the fixed Lagrange multiplier in universal quantization, our method integrates λ with a univariate quantization regulator a to control the quantization error of the overall latent representation, achieving a wide range of variable bitrates.

A shared two-layer MLP is first used to extract high-dimensional latent representation $\mathbf{y}_t$ from the residual $\mathbf{R}_t$, aggregating feature information while mitigating compression-induced information loss. This is followed by a CNN with five 3x3 layers, which refines the features and generates the final context feature $\mathbf{z}_t$. We then estimate the Gaussian entropy $p(\hat{\mathbf{y}}_t|\hat{\mathbf{z}}_t)$ of the quantized latent representation $\hat{\mathbf{y}}_t$ on condition of quantized context feature $\hat{\mathbf{z}}_t$. This estimation guides the arithmetic entropy coding of $\hat{\mathbf{y}}_t$ into a bitstream. In this paper, we use a tiny MLP to predict $p(\hat{\mathbf{y}}_t|\hat{\mathbf{z}}_t)$ as follows:

$$p(\hat{\mathbf{y}}_t|\hat{\mathbf{z}}_t) = \mathcal{N}(\mu, \sigma') * \mathcal{U}(-\frac{1}{2}, \frac{1}{2})(\hat{\mathbf{y}}_t) \tag{3}$$

where $\mathcal{N}(\mu, \sigma')$ denotes the Gaussian distribution.

When $\mathbf{y}_t$ undergoes different quantization operations, its probability distribution can vary significantly, leading to substantial quantization errors. To mitigate this, we introduce a set of learnable quantization parameters $\mathbf{A} = \{a_1, a_2, \ldots, a_n\}$, coupled with predefined Lagrange multipliers $\Lambda = \{\lambda_1, \lambda_2 \cdots \lambda_n\}$ to control these errors and enable variable bitrates. The learnable quantization parameter a_i adjusts the quantization bin size and impacts bitrate, while the Lagrange multiplier λ_i controls the trade-off between bitrate and distortion, creating a coupling relationship between a_i and λ_i. In learning-based image codecs, $\sqrt{\lambda_i}$ is nearly proportional to a_i, whereas in video codecs, QP is proportional to $\ln(\lambda)$. Thus, pairing a_i with λ_i better balances the RD trade-off, and the values are averaged across different scenes to ensure broad applicability. The latent representation $\mathbf{y}_t$ is initially scaled by its corresponding parameter a_i before being quantized into $\hat{\mathbf{y}}_t$ as follows:

$$\hat{\mathbf{y}}_t = \text{round}\left(\frac{\mathbf{y}_t}{a_i}\right) \cdot a_i, \quad a_i \in \mathbf{A}. \tag{4}$$

The entropy model of $\hat{\mathbf{y}}_t$ in Eq. 3 is then rewritten as:

$$p(\hat{\mathbf{y}}_t|\hat{\mathbf{z}}_t, a_i) = \mathcal{N}(\mu_i, \sigma_i') * \mathcal{U}\left(-\frac{1}{2a_i}, \frac{1}{2a_i}\right)(\hat{\mathbf{y}}_t) \tag{5}$$

Since the quantization operation is inherently non-differentiable, we also apply a straight-through estimator (STE) to approximate the gradient during backpropagation. The STE facilitates gradient flow through the quantization step by approximating the gradient as $\frac{\partial \hat{\mathbf{y}}_t}{\partial \mathbf{y}_t} \approx 1$. This approximation enables effective optimization of the learnable quantization parameters a_i during training, allowing the model to dynamically adjust the quantization step size and optimize

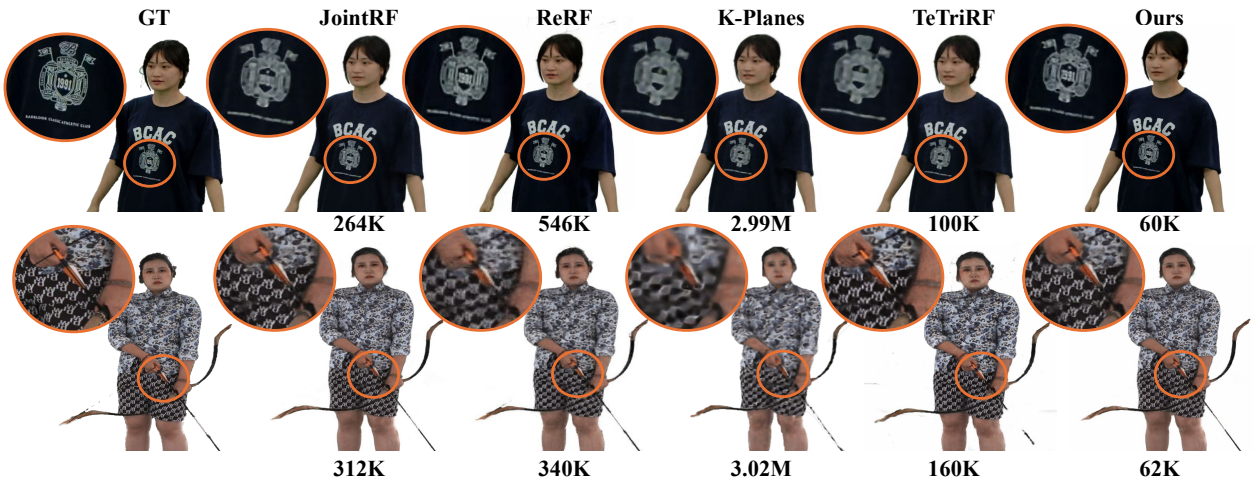

Figure 4: Qualitative comparison against volumetric video coding methods K-planes (Fridovich-Keil et al. 2023), ReRF (Wang et al. 2023), TeTrirf (Wu et al. 2024) and JointRF (Zheng et al. 2024b).

the bitrate. The RD loss function for the variable-rate model is formulated as:

$$\mathcal{L}_{RD} = \sum_{i=1}^{n} (\mathbb{E}[-\log p(\hat{\mathbf{y}}_t|\hat{\mathbf{z}}_t, a_i)] + \lambda_i \cdot D(\mathbf{R}_t, \hat{\mathbf{R}}_t)) \quad (6)$$

where $\mathbb{E}[-\log p(\hat{\mathbf{y}}_t|\hat{\mathbf{z}}_t, a_i)]$ is the estimated bitrate for encoding $\hat{\mathbf{y}}_t$, while $D(\mathbf{R}_t, \hat{\mathbf{R}}_t)$ measures the distortion between the original residual $\mathbf{R}_t$ and its reconstruction $\hat{\mathbf{R}}_t$. The Lagrange multiplier λ_i paired with a_i balances the trade-off between bitrate and distortion. Our approach integrates a univariate quantization regulator into the quantization and entropy coding process to control quantization error, applying rate-distortion supervision to achieve variable bitrates within a single model.

Progressive Training Strategy

Here, we introduce an end-to-end progressive training scheme that jointly optimizes both the representation and compression to further improve RD performance. An overview of our progressive training, which incorporates a two-stage coarse-to-fine strategy, is illustrated in Fig. 3. In the first stage, we train the density grid and feature planes at a low resolution, enabling rapid exploration of the scene's core structure. In the second stage, we leverage the low-resolution feature planes from the first stage as an effective initialization for subsequent training, combining them with a variable-rate entropy coding model for joint optimization. Our approach dramatically accelerates training while improving both rendering quality and compression efficiency.

Stage 1. The inputs for this stage include the multi-view images of the current frame and the reconstructed features $\hat{\mathbf{F}}_{t-1}$ of the previous frame obtained from the decoded buffer. These reconstructed features are downsampled to a low resolution, serving as the initialization for the coarse training stage. The outputs of this stage are the residual triplane features and the density grid, both at a low resolution. The density grid provides a rough approximation of the scene's geometry, which is essential for identifying and eliminating idle spaces during the preliminary reconstruction of the density field, thereby reducing unnecessary computational overhead. This training stage not only accelerates convergence but also establishes a solid foundation for more detailed optimization in stage 2.

Stage 2. The residual features generated in stage 1 are upsampled to a higher resolution and used as initialization for the second training stage. By reusing these features instead of starting from scratch, we greatly reduce the training time and enhance convergence speed. Additionally, we employ a learnable variable-rate entropy coding model that is jointly trained with the residual dynamic modeling. During training, multiple λ are used within the entropy model to optimize the quantization parameters $\mathbf{A}$, enabling variable-rate bitstreams. This joint training approach effectively captures high-dimensional appearance features with low entropy, significantly enhancing compression efficiency while maintaining high rendering quality.

Training Object. The multi-rate-distortion loss function of the entire framework is formulated as follows:

$$\mathcal{L}_s = \mathcal{L}_{color} + \gamma_1 \mathcal{L}_{res} + (s-1)\gamma_2 \mathcal{L}_{RD}, s \in \{1, 2\} \quad (7)$$

where $\mathcal{L}_s$ is the loss for stage s, γ_1 and γ_2 are the weights for our regular terms. $\mathcal{L}_{res} = \|\mathbf{R}_t\|_1$ serves as a residual regularization term, designed to ensure temporal continuity and minimize the magnitude of $\mathbf{R}_t$. $\mathcal{L}_{RD}$ represents variable-rate compression loss. $\mathcal{L}_{color}$ is the photometric loss,

$$\mathcal{L}_{color} = \sum_{r \in \Re} \|\mathbf{c}_g(\mathbf{r}) - \hat{\mathbf{c}}(\mathbf{r})\|^2 \quad (8)$$

where $\Re$ is the set of training pixel rays, $\mathbf{c}_g(\mathbf{r})$ and $\hat{\mathbf{c}}(\mathbf{r})$ are the ground truth and reconstructed colors of a ray $\mathbf{r}$, respectively.

Rendering Acceleration. In addition to utilizing a progressive training strategy, we also employ a deferred rendering model to further accelerate both the training and render-

	ReRF Dataset					DNA-Rendering Dataset				
	Training View		Testing View		SIZE ↓	Training View		Testing View		SIZE ↓
Methods	PSNR↑	SSIM↑	PSNR↑	SSIM↑	(KB)	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	(KB)
K-Planes	35.18	0.982	29.96	0.951	2992	31.98	0.971	27.81	0.946	3085
ReRF	35.20	0.982	30.88	0.962	496	30.20	0.968	29.59	0.950	314
TeTriRF	35.94	0.986	32.05	0.974	101	32.33	0.976	29.48	0.950	160
JointRF	35.62	0.983	31.94	0.970	227	33.74	0.979	30.27	0.962	269
Ours (Low)	35.93	0.986	32.16	0.975	**40**	33.24	0.978	30.25	0.962	**30**
Ours (Mid)	<u>36.31</u>	<u>0.988</u>	<u>32.45</u>	<u>0.976</u>	<u>62</u>	<u>34.45</u>	<u>0.981</u>	<u>31.37</u>	<u>0.968</u>	<u>56</u>
Ours (High)	**38.73**	**0.990**	**33.52**	**0.978**	223	**35.95**	**0.984**	**32.45**	**0.977**	240

Table 1: Quantitative comparison against volumetric video encoding methods. Bold data indicate the best performance, while underlined data indicate the second best.

ing processes. Specifically, We begin by accumulating the features along the ray:

$$\tilde{\mathbf{f}}(\mathbf{r}) = \sum_{k=1}^{n_s} T_k \left(1 - \exp(-\sigma_k \delta_k)\right) \mathbf{f}_k, \quad T_k = \exp\left(-\sum_{j=1}^{k-1} \sigma_j \delta_j\right), \tag{9}$$

where n_s represents the number of sample points along the ray $\mathbf{r}$, δ_k denotes the interval between adjacent samples. The density $\sigma_k = \varphi(\mathbf{k}, \mathbf{V})$ is interpolated from the density grid $\mathbf{V}$. We also leverage the density grid to eliminate points in empty space, thus reducing unnecessary computations. The composed feature $\mathbf{f}_k$ is formed by concatenating the appearance features $\mathbf{f}_k^l, l \in L$ from the tri-planes. The reconstructed color of the ray $\mathbf{r}$ is then computed using a tiny global MLP Φ that is shared across frames in the same GoF: $\hat{\mathbf{c}}(\mathbf{r}) = \Phi(\tilde{\mathbf{f}}(\mathbf{r}), \omega(\mathbf{d}))$ where $\omega(\mathbf{d})$ denotes the positional encoding of the viewing direction. This approach significantly reduces computational complexity, as each ray requires only a single MLP decoding.

Experiment

Configurations

Datasets. We validate our method on two datasets: ReRF (Wang et al. 2023) and DNA-Rendering (Cheng et al. 2023), using 2 views for testing and the rest for training.

Setups. Our experimental setup includes an Intel E5-2699 v4 and a V100 GPU. We train 40,000 iterations, with each GoF lasting 30 frames. The Lagrange multipliers Λ are initialized as $\{0.0018, 0.0035, 0.0067, 0.0130, 0.025, 0.0483, 0.0932, 0.18\}$, and the quantization parameters A are set to $\{1.0000, 1.3944, 1.9293, 2.6874, 3.7268, 5.1801, 7.1957, 10.0\}$. The weights γ_1 and γ_2 are 0.0001 and 0.001.

Evaluation Metrics. We use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) (Wang et al. 2004) as quality metrics. Bitrate is measured in KB per frame. For overall compression efficiency, we calculate the Bjontegaard Delta Bit-Rate (BDBR).

	K-Planes	ReRF	TeTriRF	JointRF	Ours
ReRF	2.20	0.44	0.13	0.78	0.13
DNA-Rendering	2.20	0.47	0.12	0.82	0.14

Table 2: Comparison of rendering time per frame in seconds.

Comparison

We demonstrate the effectiveness of VRVVC through comparisons with state-of-the-art methods qualitatively and quantitatively: K-planes (Fridovich-Keil et al. 2023), ReRF (Wang et al. 2023), TeTriRF (Wu et al. 2024), and JointRF (Zheng et al. 2024b). Fig. 4 shows qualitative results for the *kpop* sequence from ReRF and the *Archer* sequence from DNA-Rendering. Our VRVVC achieves finer detail reconstruction at a lower bitrate, such as the clothing in *kpop* and the hand in *Archer*, highlighting its superior subjective quality.

Tab. 1 shows the quantitative results on the ReRF and DNA-Rendering datasets. "Ours (Low)", "Ours (Middle)", and "Ours (High)" represent our method with variable-rate bitstreams from a single model. "Ours (Middle)" achieves higher reconstruction quality at a lower bitrate than other methods, while "Ours (High)" offers significantly better reconstruction quality with much lower bitrate than K-planes and ReRF. "Ours (Low)" outperforms TeTriRF and JointRF in bitrate, maintaining comparable quality. Additionally, both our method and TeTriRF offer rendering times at least twice as fast as ReRF and JointRF as shown in Tab. 2. VRVVC takes about 2.6 min for training and 0.13 s, 1.23 s, and 0.88 s for rendering, encoding, and decoding a frame.

The RD performance of our VRVVC compared to ReRF, TeTriRF, and JointRF is presented in Tab. 3. Our VRVVC consistently outperforms these methods in terms of RD performance. For instance, compared to TeTriRF, our method achieves average BDBR reductions of **-46.25%** for training views and **-48.27%** for testing views on the ReRF dataset. Similarly, on the DNA-Rendering dataset, we observe average BDBR savings of **-81.86%** for training views and **-83.51%** for testing views. The RD curves in Fig. 5 further illustrate that our VRVVC achieves superior RD performance across a wide range of bitrates. Unlike JointRF requires training multiple fixed-bitrate models to achieve

Dataset	Method	Training View	Testing View
		BDBR(%) ↓	BDBR(%) ↓
ReRF	ReRF	424.97	346.58
	JointRF	127.70	90.33
	Ours	**-46.25**	**-48.27**
DNA-Rendering	ReRF	177.71	103.56
	JointRF	2.99	0.32
	Ours	**-81.86**	**-83.51**

Table 3: The BDBR results of our VRVVC, ReRF and JointRF when compared with TeTriRF on different datasets.

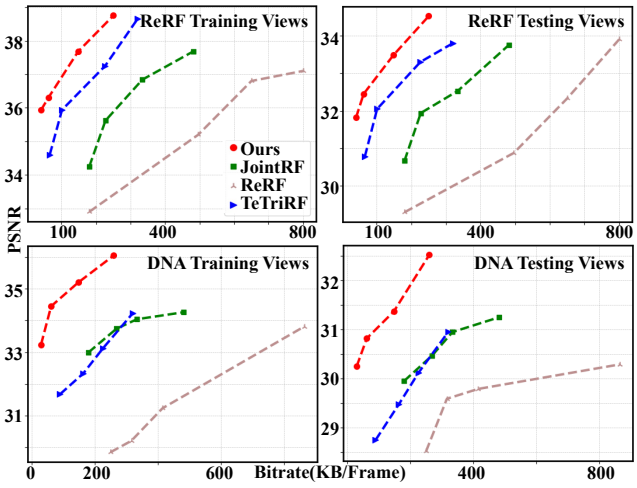

Figure 5: The RD performance comparison results on the ReRF and DNA-Rendering datasets.

different rate-distortion trade-offs, our method provides a broader range of RD performance with just a single model, offering greater flexibility and efficiency.

Ablation Studies

We perform three ablation studies on residual dynamic modeling, progressive training, and joint optimization by disabling each component individually. In the first study, we model volumetric video frame by frame without residual dynamic modeling. In the second study, we skip the initial stage and train the entire framework directly. In the last study, we train the residual representation and entropy model separately instead of optimizing them jointly.

The ablation study results in Fig. 6 show that disabling either residual dynamic modeling or progressive training leads to an increase in bitrate, underscoring the effectiveness of these modules. Additionally, joint optimization produces temporally consistent and low-entropy 4D sequential representations, which are more efficiently compressed, thereby significantly enhancing RD performance. Fig. 7 presents a qualitative comparison of the complete VRVVC at different bitrates against its variants. These findings highlight the advantages of our residual dynamic modeling, progressive training, and joint optimization strategy in volumetric video compression.

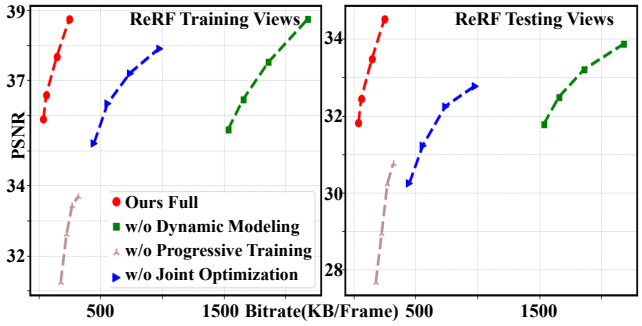

Figure 6: RD curves. This figure illustrates the efficiency of various components within our method.

Figure 7: Qualitative results of complete VRVVC and its variants. Excluding any module results in lower reconstruction quality and an increase in bitrate.

Conclusion

In this paper, we present a novel variable-rate compression framework tailored for NeRF-based volumetric video. Our tri-plane residual representation in VRVVC is compact and compression-friendly, effectively reducing spatio-temporal redundancy between frames in a sequential manner. Our residual representation compression scheme employs an implicit entropy model coupled with RD tradeoff parameters to enable variable bitrates. Our end-to-end training strategy jointly optimizes both representation and compression, significantly improving compression performance. Experimental results demonstrate that VRVVC not only achieves a wide range of variable bitrates within a single model but also surpasses state-of-the-art fixed-rate methods, greatly advancing the transmission capabilities of volumetric video.

Acknowledgements

This work was supported by National Natural Science Foundation of China (62271308), STCSM (24ZR1432000,24511106902,22511105700,22DZ2229005), 111 plan (BP0719010), Open Project of National Key Laboratory of China (23Z670104657) and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

Ballé, J.; Laparra, V.; and Simoncelli, E. P. 2016. End-to-end optimization of nonlinear transform codes for perceptual quality. In *2016 Picture Coding Symposium (PCS)*, 1–5. IEEE.

Ballé, J.; Laparra, V.; and Simoncelli, E. P. 2017. End-to-end Optimized Image Compression. In *International Conference on Learning Representations*.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations*.

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. *ICCV*.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR*.

Boyce, J. M.; Doré, R.; Dziembowski, A.; Fleureau, J.; Jung, J.; Kroon, B.; Salahieh, B.; Vadakital, V. K. M.; and Yu, L. 2021. MPEG immersive video coding standard. *Proceedings of the IEEE*, 109(9): 1521–1536.

Cao, A.; and Johnson, J. 2023a. HexPlane: A Fast Representation for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 130–141.

Cao, A.; and Johnson, J. 2023b. HexPlane: A Fast Representation for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 130–141.

Cheng, W.; Chen, R.; Fan, S.; Yin, W.; Chen, K.; Cai, Z.; Wang, J.; Gao, Y.; Yu, Z.; Lin, Z.; et al. 2023. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19982–19993.

Choi, Y.; El-Khamy, M.; and Lee, J. 2019. Variable Rate Deep Image Compression With a Conditional Autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Cui, Z.; Wang, J.; Gao, S.; Guo, T.; Feng, Y.; and Bai, B. 2021. Asymmetric Gained Deep Image Compression With Continuous Rate Adaptation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10527–10536.

Du, Y.; Zhang, Y.; Yu, H.-X.; Tenenbaum, J. B.; and Wu, J. 2021. Neural Radiance Flow for 4D View Synthesis and Video Processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

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*, SA '22. ACM.

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.

Graziosi, D.; Nakagami, O.; Kuma, S.; Zaghetto, A.; Suzuki, T.; and Tabatabai, A. 2020. An overview of ongoing point cloud compression standardization activities: Video-based (V-PCC) and geometry-based (G-PCC). *APSIPA Transactions on Signal and Information Processing*, 9: e13.

Guo, T.; Wang, J.; Cui, Z.; Feng, Y.; Ge, Y.; and Bai, B. 2020. Variable Rate Image Compression with Content Adaptive Optimization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 533–537.

Işık, M.; Rünz, M.; Georgopoulos, M.; Khakhulin, T.; Starck, J.; Agapito, L.; and Nießner, M. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics (TOG)*, 42(4).

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.

Lee, S.; Shu, F.; Sanchez, Y.; Schierl, T.; and Hellge, C. 2023. ECRF: Entropy-Constrained Neural Radiance Fields Compression with Frequency Domain Optimization. *arXiv preprint arXiv:2311.14208*.

Li, L.; Shen, Z.; Wang, Z.; Shen, L.; and Bo, L. 2023. Compressing volumetric radiance fields to 1 mb. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4222–4231.

Li, L.; Shen, Z.; Wang, Z.; Shen, L.; and Tan, P. 2022a. Streaming radiance fields for 3d video synthesis. *Advances in Neural Information Processing Systems*, 35: 13485–13498.

Li, T.; Slavcheva, M.; Zollhoefer, M.; Green, S.; Lassner, C.; Kim, C.; Schmidt, T.; Lovegrove, S.; Goesele, M.; Newcombe, R.; et al. 2022b. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5521–5531.

Lin, J.; Liu, D.; Liang, J.; Li, H.; and Wu, F. 2021. A deeply modulated scheme for variable-rate video compression. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3722–3726. IEEE.

Lu, G.; Ge, X.; Zhong, T.; Hu, Q.; and Geng, J. 2024a. Preprocessing Enhanced Image Compression for Machine Vision. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Lu, G.; Ge, X.; Zhong, T.; Hu, Q.; and Geng, J. 2024b. Preprocessing Enhanced Image Compression for Machine Vision. *IEEE Transactions on Circuits and Systems for Video Technology*, 1. Publisher Copyright: IEEE.

Lu, G.; Zhong, T.; Geng, J.; Hu, Q.; and Xu, D. 2022. Learning based Multi-modality Image and Video Compression. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6073–6082.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S. M.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. NeRF in

the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021. Nerfies: Deformable Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5865–5874.

Peng, S.; Yan, Y.; Shuai, Q.; Bao, H.; and Zhou, X. 2023. Representing volumetric videos as dynamic mlp maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4252–4262.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2020. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Rabich, S.; Stotko, P.; and Klein, R. 2024. FPO++: efficient encoding and rendering of dynamic neural radiance fields by analyzing and enhancing Fourier PlenOctrees. *The Visual Computer*, 1–12.

Reiser, C.; Szeliski, R.; Verbin, D.; Srinivasan, P.; Mildenhall, B.; Geiger, A.; Barron, J.; and Hedman, P. 2023. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 42(4): 1–12.

Rho, D.; Lee, B.; Nam, S.; Lee, J. C.; Ko, J. H.; and Park, E. 2023. Masked Wavelet Representation for Compact Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20680–20690.

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.

Song, L.; Chen, A.; Li, Z.; Chen, Z.; Chen, L.; Yuan, J.; Xu, Y.; and Geiger, A. 2023. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5): 2732–2742.

Wang, L.; Hu, Q.; He, Q.; Wang, Z.; Yu, J.; Tuytelaars, T.; Xu, L.; and Wu, M. 2023. Neural Residual Radiance Fields for Streamably Free-Viewpoint Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 76–87.

Wang, L.; Yao, K.; Guo, C.; Zhang, Z.; Hu, Q.; Yu, J.; Xu, L.; and Wu, M. 2024. VideoRF: Rendering Dynamic Radiance Fields as 2D Feature Video Streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 470–481.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wu, M.; Wang, Z.; Kouros, G.; and Tuytelaars, T. 2024. TeTriRF: Temporal Tri-Plane Radiance Fields for Efficient Free-Viewpoint Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6487–6496.

Yang, F.; Herranz, L.; Van De Weijer, J.; Guitián, J. A. I.; López, A. M.; and Mozerov, M. G. 2020. Variable rate deep image compression with modulated autoencoder. *IEEE Signal Processing Letters*, 27: 331–335.

Zheng, Z.; Zhong, H.; Hu, Q.; Zhang, X.; Song, L.; Zhang, Y.; and Wang, Y. 2024a. HPC: Hierarchical Progressive Coding Framework for Volumetric Video. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 7937–7946. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Zheng, Z.; Zhong, H.; Hu, Q.; Zhang, X.; Song, L.; Zhang, Y.; and Wang, Y. 2024b. JointRF: End-to-End Joint Optimization for Dynamic Neural Radiance Field Representation and Compression. *arXiv preprint arXiv:2405.14452*.