SPARSE2VIEW: ROTATORY NOVEL VIEW SYNTHESIS FROM SCARCE X-RAY INPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The X-ray computed tomography (CT) requires a large number of X-ray projections for construction, exposing patients to radiation and a prolonged imaging time, To address these concerns, this work proposes Sparse2View, a novel rotational X-ray view synthesis framework, able to generate full-view projections from only a few sparse view inputs. Sparse2View features two key innovations: (1) intra-view feature refinement, which employs a tailored anisotropic spectral operator to capture the frequency information within the projection sequences; and (2) inter-view geometric fusion, which enhances geometric representation with cross-view information integration. Additionally, we curate a large-scale, deep learning–ready dataset comprising 1,614 object-level (mouse skeleton) scans, yielding a total of 419,640 medical expert–annotated 2D X-ray images. Extensive experiments demonstrate that Sparse2View can synthesize rotatory novel views on-demand under extremely sparse settings. For instance, taking only 4 views as inputs, it can generate the full 260-view of X-ray views over a 208-degree rotating angle (dramatically reducing the radiation dose by 65 folds), with an average PSNR of 36.72 and an SSIM of 0.931. Moreover, results under the zero-shot settings validate that the Sparse2View can be well generalized to other anatomical regions, unseen fractured bones, and human organs, underscoring its potential medical applications.

1 INTRODUCTION

X-ray computed tomography (CT) has revolutionized medical imaging and transformed patient care across radiology, surgery, and oncology (Withers et al., 2021). It enables high-resolution imaging of entire organs, and is indispensable for identifying structural abnormalities, diagnosing and staging diseases, and guiding surgical interventions. However, each CT scan requires a large number of X-ray projections (typically around 600 or more), which exposes the patient to ionizing radiation and prolongs the imaging time. A recent study (Smith-Bindman et al., 2025) estimated that 103,000 future cancers could be attributed to CT exposure over patients' lifetimes—representing roughly 5% of new cancer diagnoses in a given year. To this end, a natural question arises: *Can we leverage advanced AI techniques to synthesize novel rotational X-ray views with only a few sparse inputs?*

Recent efforts to address this question generally fall into two categories: 1) Neural Radiance Fields (NeRF)-based methods (Mildenhall et al., 2021; Barron et al., 2021; Müller et al., 2022; Chen et al., 2022; Xu et al., 2022) and 2) 3D Gaussian Splatting (3DGS)-based approaches (Kerbl et al., 2023; Yu et al., 2024). NeRF-based methods (e.g., SAX-NeRF, NAF, and MedNeRF) model a continuous 3D scene as a 5D function, parameterized by the 3D spatial location and 2D viewing direction, by employing deep learning architectures such as multilayer perceptrons, LineFormer, or hash-encoded implicit neural networks. Given these inputs, the network predicts the corresponding volume density and radiance. However, they often struggle to synthesize high-quality projections from extremely sparse input views, and the learned implicit representations often generalize poorly to unseen objects. The 3DGS-based approaches (e.g., X-Gaussian and R^2-Gaussian) represent a 3D scene explicitly as a set of 3D Gaussian primitives. During rendering, these Gaussians are projected onto the 2D image plane and then rapidly alpha-blended in depth-sorted order to synthesize the photorealistic X-ray images. While these methods provide the improved rendering speed and performance, they remain limited in their ability to generalize to unseen objects or anatomical structures. Additionally, both prior NeRF- and 3DGS-based approaches applied the simulated data generated from open-source tomographic toolbox TIGRE (Biguri et al., 2016), that assumes an idealized pure-attenuation

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

imaging model. This creates a critical gap with the real-world imaging conditions. Hence, it is in urgent need of a real-world dataset and a novel framework that can performs robustly under extremely sparse inputs, and generalize effectively across new scenes or objects.

To comprehensively investigate this problem, this work will first introduce **MiceBoneX**, the first large-scale, deep learning–oriented X-ray projections dataset. It contains 1,614 object-level scans, each with 260 rotatory X-ray projections, resulting in 419,640 2D X-ray images. Subsequently, we develop a novel rotational X-ray view synthesis framework, **Sparse2View**, capable of generating high-fidelity full-view projections from only a few sparse view inputs (e.g., 4 or 8 views). Distinct from NeRF and 3DGS-based approaches, our Sparse2View adopts a transformer backbone and introduces two new innovations. The first is the intra-view feature refinement, which enhances the generalization of Vision Transformer (ViT) (Dosovitskiy et al., 2021) by introducing a learnable Fourier Transform-based anisotropic spectral operator to capture the frequency information within each projection sequence. The second is the inter-view geometric fusion, which introduces a novel Hybrid Rotary Position Embedding Multi-head Self-Attention (HRoPE-MSA) mechanism, to capture geometric representation with cross-view information integration through the fusion of projective positional encoding and a Plücker Raymap representation.

Both quantitative and qualitative results demonstrate that Sparse2View can synthesize high-fidelity full-view X-ray projections under extremely sparse settings. For instance, with only 4 and 8 input views, the model attains average PSNR values of 36.72 dB and of 38.65 dB, respectively. Furthermore, Sparse2View shows strong zero-shot generalization to other anatomical regions, unseen fractured bones and human organs, highlighting its robustness and adaptability.

2 Related Works

We discuss the recent studies of X-ray novel view synthesis and highlights how our approach differs from them.

Neural Radiance Fields (NeRF) was first introduced in (Mildenhall et al., 2021), which employs a Multi-Layer Perceptron (MLP) to represent a continuous 3D scene as a 5D function. This function takes a 3D spatial coordinate and a 2D viewing direction as inputs, and outputs the corresponding volume density and radiance at that point. Building upon this framework, a number of studies (Zha et al., 2022; Cai et al., 2024b; Corona-Figueroa et al., 2022; Zang et al., 2021) have been proposed to enhance the synthesis quality of X-ray views. For example, SAX-NeRF (Cai et al., 2024b) replaces the standard MLP in NeRF with a Lineformer to improve structural learning and achieve novel-view rendering from 50 input views. NAF (Zha et al., 2022) proposes a fast, self-supervised method that leverage a hash-encoded implicit neural network to reconstruct high-quality 3D CT volumes from sparse-view projections. MedNeRF (Corona-Figueroa et al., 2022) adapts generative radiance fields by introducing a powerful discriminator enhanced with self-supervised learning and a specialized data augmentation (DAG) framework for lifting medical image synthesis quality. However, NeRF-based approaches require extensive ray sampling to achieve the satisfactory synthesis quality, and their learned 3D representations often fail to generalize well to unseen objects.

3D Gaussian Splatting (3DGS (Kerbl et al., 2023)) represents a 3D scene as an explicit set of 3D Gaussian primitives. Each Gaussian is characterized by its special position, covariance, opacity, and color. The covariance dictates the shape, scale, and orientation of the Gaussian, while Spherical Harmonics are employed to model view-dependent color variations. During rendering, these 3D Gaussians are projected onto the 2D image plane and then rapidly alpha-blended in depth-sorted order to synthesize novel views. Building upon this, several subsequent works (Li et al., 2023; Cai et al., 2024a; Zha et al., 2024; Wang et al., 2025; Wu et al., 2025b) have adapted 3DGS framework for X-ray projection synthesis. For instance, X-Gaussian (Cai et al., 2024a) extends 3DGS framework for X-ray novel view synthesis, by introducing a radiative Gaussian model with differentiable rasterization tailored to X-ray isotropy, along with an angle-pose cuboid initialization strategy for circular cone beam scanning. R^2-Gaussian (Zha et al., 2024) addresses integration bias inherent in standard Gaussian splatting by reformulating the X-ray projection formulas and implementing a CUDA-based, differentiable voxelizer tool to convert optimized Gaussians into a 3D volume. However, they typically require 25, 50, or 75 views to optimize 3D Gaussian representation. X-Field (Wang et al., 2025) tailors 3D representation to X-ray imaging by modeling physical properties of penetration and attenuation, instead of optical reflection. It exhibits relatively stronger performance under sparse settings, but suffering from limited generalization ability to unseen anatomical structures.

Diffusion and Transformer-based models have been explored for novel view synthesis of nature images (Zhou et al., 2025; Jin et al., 2025), but they have not yet been applied to the novel X-ray view synthesis. Given the fundamental differences between X-ray imaging and nature image acquisition, particularly the rotational nature of X-ray imaging, we omit their discussions here.

3 OUR MICEBONEX DATASET

We curate a new deep learning–oriented X-ray dataset, namely **MiceBoneX**, which was based on micro computed tomography (microCT) scanning of hindlimbs of anesthetized mice using a live animal scanner (Bruker® Skyscan 1276). For each scan, sequential 2D X-ray projections were taken at different views while the gantry assembly consisting of the X-ray source and detector rotated around the animal (rotation center) with an incremental angle. The Distance from Source to Object (d_1) and Distance from Source to Detector (d_2) as well as the detector size determine the resolution of the projections. We used the settings recommended by the vendor, including d_1 (88.02 mm), d_2 (158.85 mm), scanning resolution (7 microns), incremental angle (0.8 degrees), and a total of 260 projections spanning from the anterior to the lateral and posterior views of the limbs. The 260 projections cover a swing angle of 208 degrees, adequate for 3D reconstruction of the hindlimbs.

In total, this dataset consists of 1,614 object-level (mouse skeleton) scans and 260 rotatory X-ray projections per scan, resulting in 419,640 medical expert-annotated 2D X-ray images. To the best of our knowledge, this dataset is the largest publicly available X-ray dataset in the community. Since the original images are of ultra-high resolution (7 microns per pixel) with a large image size (4032 × 2688 pixels), we have performed data preprocessing to facilitate model training and validation. Specifically, all images are resized to 256 × 256 pixels and augmented with metadata of the scanning parameters. These parameters are then stored as key–value pairs and specifically include information such as the DSO d_1 and DSD d_2. For each scan, both the camera's intrinsics (the camera internal properties) and extrinsics (the 3D poses, e.g., position and orientation of the camera for every view) were computed and recorded along with each image. The specifications of our MiceBoneX dataset and a sample case where sequential projections are generated are shown in Appendix A.1.

4 METHODOLOGY

Consider a set of N samples $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N}$, where $\boldsymbol{x} \in \mathbb{R}^{V_\text{i} \times H \times W \times C}$ and $\boldsymbol{y} \in \mathbb{R}^{V_\text{t} \times H \times W \times C}$ represent the input and target 2D X-ray projections, respectively. Here, V_i denotes the number of input X-ray projections, while V_t denotes the corresponding number of generated X-ray views. H, W, and C respectively indicate the height, width, and number of channels of an X-ray projection. We derive pixel-wise Plücker ray embeddings (Plucker, 1865) $\boldsymbol{r} \in \mathbb{R}^{H \times W \times 6}$ from camera poses and intrinsic parameters. Given the learnable target query (with a hidden dimension size of d_i), denoted as $\boldsymbol{q} \in \mathbb{R}^{V_\text{t} \times d_i}$, our task is to train a model (with its parameters $\boldsymbol{\theta}$) $f_{\boldsymbol{\theta}} : (\mathbb{R}^{V_\text{t} \times d_i}, \mathbb{R}^{V_\text{i} \times H \times W \times C}, \mathbb{R}^{H \times W \times 6}) \rightarrow \mathbb{R}^{V_\text{t} \times H \times W \times C}$ that generates the V_t views of target X-ray images by conditioning on the V_i views of input X-ray images and the Plucker ray embeddings, expressed as $f_{\boldsymbol{\theta}}(\boldsymbol{q} \mid \boldsymbol{x}, \boldsymbol{r}) = \boldsymbol{y}$. We focus on rotatory novel view synthesis, in which both the input and target X-ray images are obtained from the same object but at rotated angles, under the challenging setting for scarce input views.

4.1 MODEL OVERVIEW

Figure 1 illustrates the workflow of our Sparse2View architecture, which consists of four key components: *Tokenization*, *Anisotropic Processing*, *Transformer Encoding*, and *Output Decoding*. The *Tokenization* employs two parallel tokenizers: i) an image tokenizer that processes input views; and ii) a ray tokenizer that converts Plücker ray embeddings into ray tokens. The *Anisotropic Processing* processes the concatenation of image tokens and of target query tokens, and its goal is to incorporate 2D inductive bias (learned from intra-view spatial dependencies) into the Vision Transformer (ViT) (Dosovitskiy et al., 2021; Jin et al., 2025). Its output is concatenated with ray tokens before applying our novel Multi-head Self-Attention (MSA) method within the *Transformer Encoding*. Finally, the *Output Decoding* selectively utilizes the encoded tokens associated with the target query to generate the target views of X-ray images. The detailed descriptions are elaborated as below.

4.2 SPARSE2VIEW

Tokenization. It processes two types of inputs: X-ray images and Plücker ray embeddings, using dedicated tokenizers for each. We first divide the X-ray images and Plücker ray embeddings

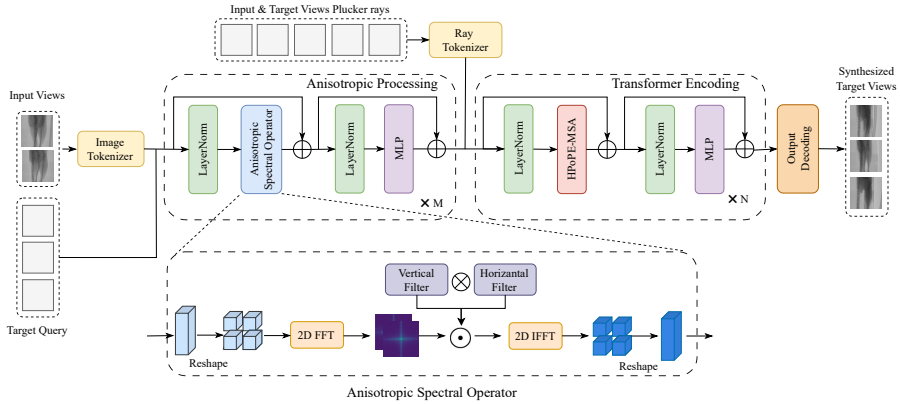

Figure 1: Sparse2View model architecture.

into non-overlapping patches with a patch size of p. These patches are then projected into high-dimensional tokens via the linear projection, yielding image tokens $\boldsymbol{z}_i$ and ray tokens $\boldsymbol{z}_r$, i.e.,

$$\boldsymbol{z}_i = \text{Proj}_i\left(\tau(\boldsymbol{x})\right) \quad \in \quad \mathbb{R}^{V_i \times P \times d_i}, \quad \boldsymbol{z}_r = \text{Proj}_r\left(\tau(\boldsymbol{r})\right) \quad \in \quad \mathbb{R}^{V \times P \times d_r}. \tag{1}$$

Here, $\tau(\cdot)$ denotes the patchify process, $V = V_i + V_t$ represents the total views of input and target X-ray images, and $P = \frac{H}{p} \times \frac{W}{p}$ indicates the total number of patches. The linear projection layers Proj_i and Proj_r map the patches to token spaces of dimensions d_i and d_r, respectively.

Anisotropic Processing. The ViT architecture (Dosovitskiy et al., 2021; Jin et al., 2025) can reduce the dependency on 3D inductive biases, but lacking important 2D inductive biases, Hence, we propose a new block, namely Anisotropic Processing, to introduce essential 2D inductive biases into the ViT architecture, substantially improving the zero-shot generalizability. Specifically, we introduce a per-view Anisotropic Spectral Operator $S_{\boldsymbol{\phi}}(\cdot)$, which operates on image tokens of each view to learn intra-view spatial dependencies. Let $\boldsymbol{z}_{i,q} \in \mathbb{R}^{V \times P \times d_i}$ represent the concatenation of input image tokens or of query target tokens, which is reshaped to a 2D grid $\boldsymbol{z}^{\text{grid}}_{i,q} \in \mathbb{R}^{V \times P^{\frac{1}{2}} \times P^{\frac{1}{2}} \times d_i}$, corresponding to the spatial arrangement of image patches. This operator leverages the 2D Convolution Theorem (Cooley & Tukey, 1965) to perform efficient global convolution within this 2D spatial domain. Let $\mathcal{F}_{2D}$ and $\mathcal{F}^{-1}_{2D}$ denote the 2D Real Discrete Fourier Transform (Sorensen et al., 2003) and its inverse, respectively. This operation can be expressed as:

$$S_{\boldsymbol{\phi}}(\boldsymbol{z}^{\text{grid}}_{i,q}) = \mathcal{F}^{-1}_{2D}\left(\boldsymbol{w}_{\boldsymbol{\phi}} \odot \mathcal{F}_{2D}(\boldsymbol{z}^{\text{grid}}_{i,q})\right). \tag{2}$$

Note that after the inverse process $\mathcal{F}^{-1}_{2D}$, we reshape the 2D grid back to its original 1D sequence. We parameterize $\boldsymbol{w}_{\boldsymbol{\phi}}$ as an anisotropic filter (Freeman et al., 1991), factorized into two independent learnable filters, i.e., a vertical frequency filter $\boldsymbol{w}_{\boldsymbol{\phi},v} \in \mathbb{R}^{P^{\frac{1}{2}} \times 1 \times d_i}$ and a horizontal frequency filter $\boldsymbol{w}_{\boldsymbol{\phi},h} \in \mathbb{R}^{1 \times P^{\frac{1}{2}} \times d_i}$. The final filter is their outer product, yielding $\boldsymbol{w}_{\boldsymbol{\phi}} = \boldsymbol{w}_{\boldsymbol{\phi},v} \otimes \boldsymbol{w}_{\boldsymbol{\phi},h}$. To further engage viewpoint information in the per-view processing, we condition the filter $\boldsymbol{w}_{\boldsymbol{\phi}}$ on the relative angle distance between each input view and the target view. Let $\boldsymbol{p}_i \in \mathbb{R}^5$ denote a pose descriptor comprising the relative rotation angle, the camera-center baseline, and the unit direction between the input view i and the target view. A lightweight hypernetwork $g_{\boldsymbol{\psi}}$ learnt to map $\boldsymbol{p}_i$ to vertical and horizontal frequency gains that rescale the shared filter, i.e.,

$$\boldsymbol{w}_{\boldsymbol{\phi},i} = \boldsymbol{w}_{\boldsymbol{\phi}} \odot (\boldsymbol{g}_{i,v} \otimes \boldsymbol{g}_{i,h}), \quad (\boldsymbol{g}_{i,v}, \boldsymbol{g}_{i,h}) = 1 + \tanh\left(g_{\boldsymbol{\psi}}(\boldsymbol{p}_i)\right), \tag{3}$$

Here $g_{\boldsymbol{\psi}}$ is zero-initialized so that $\boldsymbol{w}_{\boldsymbol{\phi},i} = \boldsymbol{w}_{\boldsymbol{\phi}}$ at the beginning of training.

The Anisotropic Processing comprises an Anisotropic Spectral Operator and a MLP. Each component incorporates Layer Normalization (LN) (Ba et al., 2016) before the block and employs a residual connection (He et al., 2016) afterward. So, at the ℓ-th block, it can be expressed as:

$$\boldsymbol{z}^{\ell}_{i,q} = S_{\boldsymbol{\phi}}\left(\text{LN}(\boldsymbol{z}^{\ell-1}_{i,q})\right) + \boldsymbol{z}^{\ell-1}_{i,q}, \quad \boldsymbol{z}^{\ell+1}_{i,q} = \text{MLP}\left(\text{LN}(\boldsymbol{z}^{\ell}_{i,q})\right) + \boldsymbol{z}^{\ell}_{i,q}. \tag{4}$$

Transformer Encoding. To facilitate the rotatory X-ray images synthesis, we take into account both the token-level absolute geometry representation and the sequence-level relative positional information for preserving the rotatory properties. Specifically, at the token level, we incorporate the Plücker

ray embeddings by concatenating the image tokens and ray tokens, *i.e.*, $\boldsymbol{z} = [\boldsymbol{z}_{i,q}, \boldsymbol{z}_r] \in \mathbb{R}^{V\times P\times d}$ with $d = d_i + d_r$, guiding the model to capture the absolute geometry information. At the sequence level, we utilize the full set of camera parameters (intrinsics and extrinsics) as a relative positional encoding, inspired from PRoPE (Li et al., 2025). To capture the mixture of absolute geometry representation and relative positional information, we develop a new Hybrid Rotary Position Embedding Multi-head Self-Attention (HRoPE-MSA) mechanism, yielding:

$$\begin{aligned} \boldsymbol{A} &= \text{softmax}\left(\frac{\boldsymbol{T}_Q(\boldsymbol{Q},\boldsymbol{C})\boldsymbol{T}_K(\boldsymbol{K},\boldsymbol{C})^T}{\sqrt{d_k}}\right), \\ \text{HRoPE-MSA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{C}) &= \boldsymbol{T}_O(\boldsymbol{A}\cdot\boldsymbol{T}_V(\boldsymbol{V},\boldsymbol{C}),\boldsymbol{C}), \\ \boldsymbol{Q} = \boldsymbol{W}_Q\cdot\boldsymbol{z},\quad \boldsymbol{K} &= \boldsymbol{W}_K\cdot\boldsymbol{z},\quad \boldsymbol{V} = \boldsymbol{W}_V\cdot\boldsymbol{z},\quad \text{and } \boldsymbol{C} = \boldsymbol{W}_C\cdot\boldsymbol{z}_c. \end{aligned} \tag{5}$$

Here, $\boldsymbol{T}_Q(\cdot,\cdot)$, $\boldsymbol{T}_K(\cdot,\cdot)$, $\boldsymbol{T}_V(\cdot,\cdot)$, and $\boldsymbol{T}_O(\cdot,\cdot)$ are PRoPE transformations (Li et al., 2025) applied to the query, key, value, and output token, respectively. $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$, and $\boldsymbol{W}_C$ are four learnable projection matrices. $\boldsymbol{z}_c$ represents the full set of camera parameters, which is explicitly decoupled into the spatial extrinsic transformations and intrinsic projection models. Specifically, the extrinsic component is formulated as the world-to-camera view matrix $\boldsymbol{M}_{ext} \in \mathbb{R}^{B\times V\times 4\times 4}$, where B represents the batch size and V denotes the number of involved views. This matrix is computed by inverting the concatenated camera-to-world matrices of the involved views, while the intrinsic matrix $\boldsymbol{M}_{int} \in \mathbb{R}^{B\times V\times 3\times 3}$ is constructed from the focal lengths and principal points (f_x, f_y, c_x, c_y).

Our design enables HRoPE-MSA to dynamically map the decoupled camera parameters into block-diagonal rotation matrices, rather than merely relying on static positional indices or additive bias terms. These matrices are then multiplied with the normalized queries and keys, modulating their phase within a continuous rotary manifold. Crucially, during the scaled dot-product computation, the inner product between the rotated queries and keys implicitly evaluates the relative projective transformation between different camera views. This mathematically elegant design ensures that the attention weights are explicitly conditioned on real-world relative camera geometry, without altering the underlying self-attention architecture.

Similar to Anisotropic Processing, our Transformer Encoding is composed of an HRoPE-MSA and an MLP, with Layer Normalization and the residual connection applied before and after each block:

$$\begin{aligned} \boldsymbol{z}^{\ell} &= \text{HRoPE-MSA}\left(\text{LN}(\boldsymbol{z}^{\ell-1})\right) + \boldsymbol{z}^{\ell-1}, \\ \boldsymbol{z}^{\ell+1} &= \text{MLP}\left(\text{LN}(\boldsymbol{z}^{\ell})\right) + \boldsymbol{z}^{\ell}. \end{aligned} \tag{6}$$

Output Decoding. Finally, the encoded tokens $\boldsymbol{z}_t$ corresponding to V_t views of the target query are selected from the output sequence. A decoding layer, consisting of a LN and a linear projection layer, maps these feature-rich tokens back to the pixel values of synthetic X-ray images.

4.3 Loss Function

The $\mathcal{L}_1$ (or $\mathcal{L}_2$) and SSIM losses have been widely adopted to minimize the prediction error between the synthetic and ground-truth images (Zha et al., 2024; Cai et al., 2024a; Jin et al., 2025). However, the $\mathcal{L}_1$ loss is effective only if the model possesses strong 3D inductive biases, whereas the $\mathcal{L}_2$ loss often suffers from slow convergence due to its smooth gradients. Considering this, we adapt a flexible loss selection mechanism, Huber loss (Huber, 1992), that adaptively switches between $\mathcal{L}_1$ and $\mathcal{L}_2$ losses depending on the error magnitude, i.e.,

$$\mathcal{L}_{\text{Huber}}(\boldsymbol{\delta}) = \frac{1}{d_y}\sum_{j=1}^{d_y}\begin{cases}\frac{1}{2}\delta_j^2, & \text{if } |\delta_j| \le \beta \\ \beta\cdot(|\delta_j| - \frac{1}{2}\beta), & \text{otherwise.}\end{cases} \tag{7}$$

Here, β denote the pre-defined threshold and $\boldsymbol{\delta} = \boldsymbol{y}_{\text{pred}} - \boldsymbol{y}_{\text{true}} \in \mathbb{R}^{d_y}$ represent the prediction error.

During the early training stage, the Huber loss behaves like $\mathcal{L}_1$ loss (with large prediction errors), enabling faster convergence. As training progresses, it transitions to $\mathcal{L}_2$ loss, providing smooth gradients that help capture fine-grained features. As such, our total loss function is defined as below:

$$\mathcal{L} = \mathcal{L}_{\text{Huber}}(\boldsymbol{y}_{\text{pred}} - \boldsymbol{y}_{\text{true}}) + \lambda\cdot\text{Perceptual}(\boldsymbol{y}_{\text{pred}} - \boldsymbol{y}_{\text{true}}). \tag{8}$$

Here, the Huber loss emphasizes pixel-level fidelity, thereby improving synthetic quality, while the Perceptual loss (Johnson et al., 2016) facilitates high-level feature extraction, ultimately enhancing generalization. The hyperparameter λ balances their contributions.

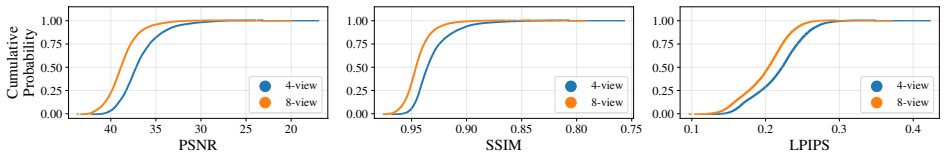

Figure 2: CDFs for per-view rendering quality metrics from a 4-view input and an 8-view input.

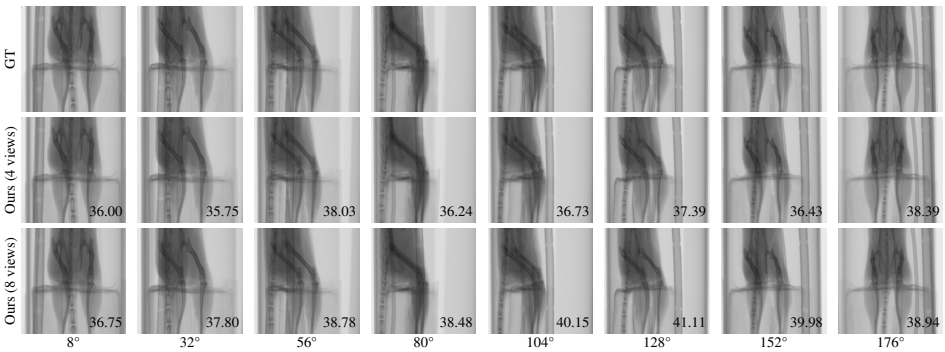

Figure 3: Qualitative results of novel view synthesis with 4-view and 8-view inputs.

5 Experiments

Our Sparse2View was trained on three NVIDIA H200 GPUs, with a batch size of 8 per GPU and a total of $100,000$ steps. We use the AdamW (Loshchilov & Hutter, 2019) optimizer (with $\beta_1 = 0.9$, $\beta_2 = 0.95$) and a cosine learning rate schedule with a peak learning rate of 4×10^{-4}, a weight decay of 0.05, and a warm-up of $1,000$ iterations. The X-ray images are tokenized using an image patch size of 16 and a hidden dimension of 768.

We adopt three widely used evaluation metrics: Peak Signal-to-Noise Ratio (PSNR) (Eskicioglu & Fisher, 2002), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018). Following the conventional approaches (Zha et al., 2024; Cai et al., 2024b), we also evaluate reconstruction quality using PSNR and SSIM, with PSNR calculated in 3D volume and SSIM averaged over 2D slices in axial, coronal, and sagittal directions, for reconstructed 3D volumes.

5.1 Main Results

We train Sparse2View on our curated **MiceBoneX** dataset, taking the splitting ratio of 9:1 for objects for training and testing. Using the trained model, we evenly select 4 or 8 views of each test object as inputs, to synthesize the remaining views, generating a complete sequence of 260 rotated images.

Overall Quantitative Performance. With 8-view input, our model can achieve the average PSNR, SSIM and LPIPS values of 38.64, 0.944, and 0.200, respectively, across all testing objects. Even with extreme sparse 4-view inputs, the average values remain 36.72, 0.931, and 0.219 for PSNR, SSIM, and LPIPS, respectively. To further analyze the model performance, we plot the cumulative distribution functions (CDF) of synthesized X-ray projections with respect to three performance metrics as shown in Figures 2 (a), (b), and (c). The three plots clearly demonstrate that our model achieves high fidelity for the synthesized X-ray projections. For instance, with 8-view input, 84.89% of projections achieve a PSNR above 37, 84.89% attain an SSIM above 0.93, and 84.89% yield an LPIPS below 0.23.

Qualitative Results. Due to the page limit, we cannot display all generated views. Instead, we randomly select a case of synthesized views to show their quality. Figure 3 showcases the synthesized views with incremental angles (from $8°$ to $176°$) and their corresponding ground truth (GT) under the 4- and 8-view settings. We observe that the synthesized views are strikingly similar to their ground-truths under the scenarios of 4- and 8-input views with PSNR values as indicated at the right bottom of each view. Our results are consistent across all cases, demonstrating that our model can generate accurate and structurally consistent rotatory image sequences with very sparse input views. Notably, a full-view 260 projection for a new subject can be synthesized within seconds.

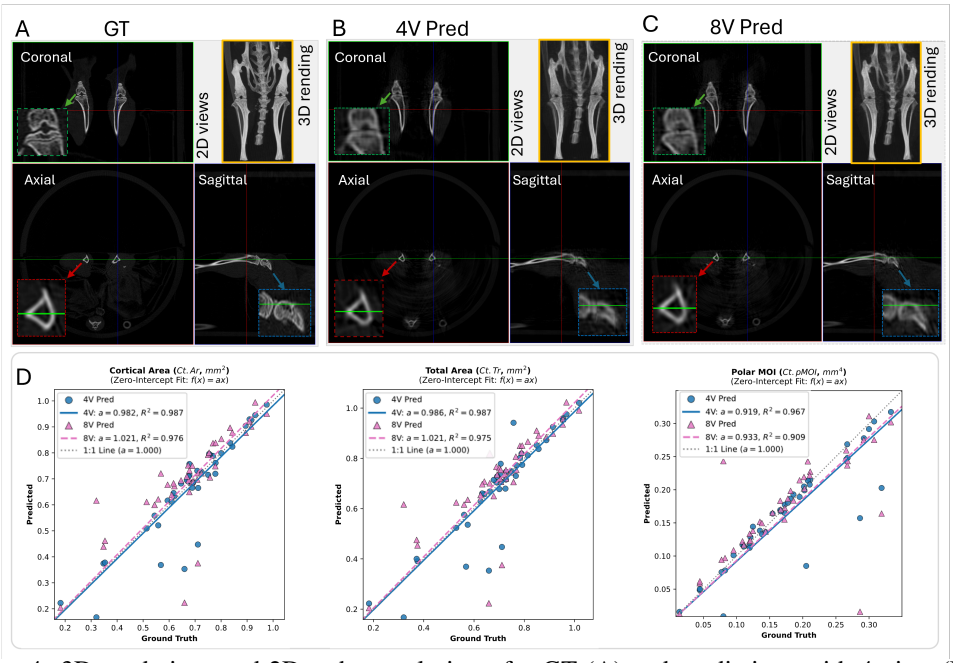

Figure 4: 3D renderings and 2D orthogonal views for GT (A) and predictions with 4-view (B) or 8-view input (C). Blowups show detailed structures. Linear regression of tibial cortical area (D), total area (E) and polar moment of inertia (F) between predictions and GT (n = 20 pairs) suggests the model's potentials in predicting subject-specific features and enabling clinical assessments.

To further exhibit the utility of our synthesized views, we used Bruker's Recon(R) software to generate the 3D renderings from 20 randomly selected cases of GT and predicted X-rays using 4- and 8-view input. Both predicted renderings show consistent anatomical features of knees, hips, pelvis and spines with high fidelity to GT (see the top left of Figure 4 A-C). Quantitatively, the averaged PSNR values are 37.93, 40.44, and 39.48, when comparing 8-view results with GT in the axial, coronal, and sagittal planes, respectively. The 4-view and 8-view predictions' three orthogonal views demonstrate anatomically accurate skeletal shapes/positioning and internal structures in comparison with GT (see blowups in Figure 4). Histomorphometry measurements using CT slices is a standard evaluation tool in orthopedics. Thus, we measured the cortical bone area Ct.Ar, total area (Ct.Tr), and polar moment of inertia (Ct.pMOI) of the two tibiae, one slice per scan, using Bruker's CTAN(R) software with a global threshold of 90/255. Linear regression of the 4-view or 8-view predictions against GT showed a slope between 0.92 and 1.02 with R2 $>$ 0.9 (Figure 4). Please note that the binary operation could contribute to data scattering. Nevertheless, the excellent agreement between the predictions and GT over the healthy variations in the three measures strongly suggest our model's potentials in providing subject-specific clinically relevant assessments, even in the sparse input conditions.

Generalization to Clinically Relevant Fractures. We have examined our model's zero-shot generalization capability to unseen anatomical regions, which achieve the promising novel X-ray synthesis performance. The results were presented in Appendix A.3. To better evaluate clinical value, we also evaluate Sparse2View on fractured data. Our fracture dataset having 21 object scans includes a diverse spectrum of fracture patterns, including transverse fractures with large and small gaps and compound fractures (multiple bone fragments). Under the zero-shot setting, 8-view inputs attain an average PSNR of 32.65 and SSIM of 0.911. The model correctly predicts fracture gaps across all views without noticeable blurry. We fine-tune the model on 11 new fracture scans for 100 steps. Fine-tuning improves the average PSNR to 34.08 and SSIM to 0.915. Under the 16-view zero-shot setting, the model further improves the average PSNR to 34.44 and SSIM to 0.926. Figure 5 shows representative qualitative results in femoral and tibial transverse fractures (3 cases each) and compound fractures with multiple broken pieces (2 cases). More qualitative results are shown in Appendix A.4 and the CDFs of PSNR and SSIM for three settings are shown in Appendix A.5.

Generalization to Human Bones. Furthermore, we evaluate the zero-shot performance of Sparse2View on human limbs from MORE dataset (Wu et al., 2025a). We leverage the 8-view input

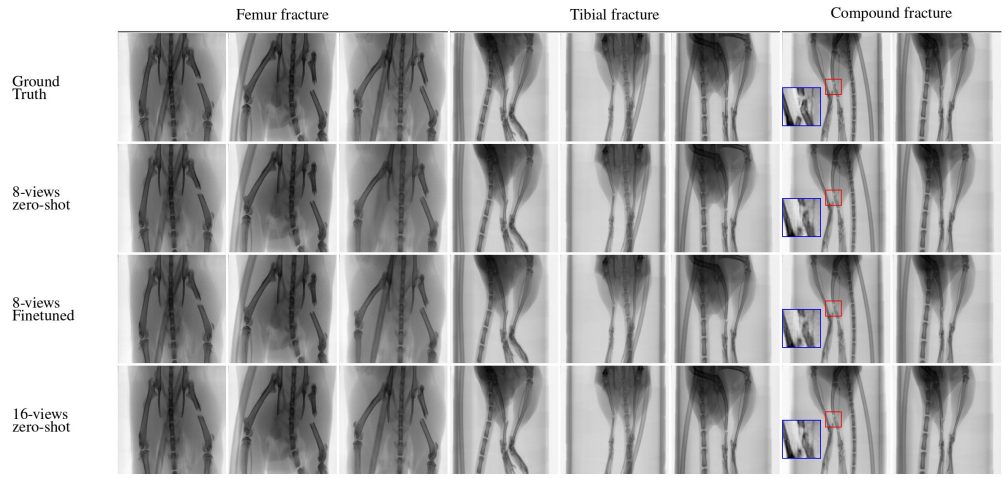

Figure 5: 8-view zero-shot, 8-view finetune, 16-view zero-shot results for synthesizing X-ray projections of femur fracture, tibial fracture, and tibial Compound fracture.

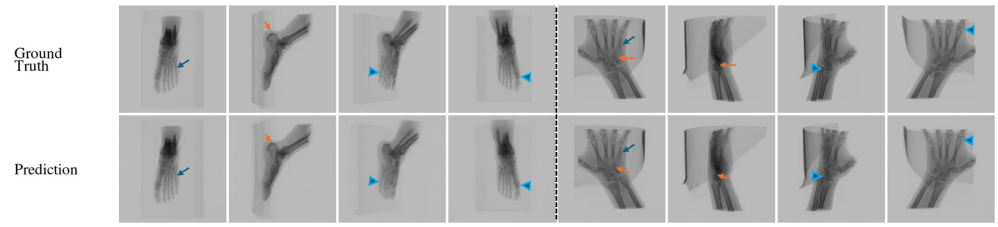

Figure 6: Zero-shot experiments for synthesizing X-ray projections of human organs.

model trained on the MiceBoneX dataset to synthesize novel views of human skeletons, achieving the averaged PSNR of 36.69, SSIM of 0.957, and LPIPS of 0.219. Due to page limits, we only present the synthesized X-ray projections on two human bones (Figure 6). Our model outputs successfully project the many bony segments in human feet/ankles and hands/wrists, revealing not only tubular bones (blue arrows) in feet (metatarsals) and hands (metacarpals) but also those irregular ones (orange arrows) like calcaneus and hamate, as well as the multiple articulating joints (cyan arrow heads) in human skeletons. More qualitative results are presented in Appendix A.6.

5.2 Comparison to State-of-the-Art Methods

We next compare our Sparse2View with three traditional, one NeRF-, two 3DGS-, one diffusion- and two transformer-based methods on our MiceBoneX dataset. Since traditional, NeRF- and 3DGS-based methods take the line integral of attenuation as input rather than transmission images from X-ray projections, we follow their protocols and use the tomography toolbox TIGRE (Biguri et al., 2016) to convert pixel values into model input via Beer–Lambert law(Beer, 1852), and then inversely convert the output back to X-ray projections for evaluation. As the three supervised learning methods were designed for nature images, we rectify them to conduct the X-ray view synthesis task.

Table 1 presents the comparative results, with reported values representing the mean across all test cases. It is observed that Sparse2View outperforms all baselines across three metrics. For example, with the 4 views and 8 views as inputs, our Sparse2View can attain the PSNR values of 36.72 and 38.64, respectively, substantially outperform the best baseline LVSM, which has the PSNR values of only 34.57 and 36.38, respectively. These results highlight Sparse2View's overall effectiveness in X-ray generation. We would like to remark that the 3DGS-based baselines assume an idealized pure-attenuation environment, which does not suit the real-world X-ray condition.

Figure 7 provides a qualitative comparison between Sparse2View and its baselines. As illustrated, with 8-input views, unsupervised baselines fail to recover detailed bone structures. In contrast, Sparse2View successfully suppresses blurring and eliminates linear and wave-pattern artifacts. It also accurately preserves fine anatomical details such as the spaces between adjacent tail vertebrae and the contrasting tibial cortices (high density) and marrow cavities (low density). Our model incorporates the X-ray source and camera positioning parameters, thereby enabling more accurate projections of pixel intensity and better discrimination of calcified bone tissues against surrounding

Table 1: Comparative results on the novel view synthesis task. We mark out best, second best, and third best method for all metrics.

Method	4-views PSNR↑	4-views SSIM↑	4-views LPIPS↓	8-views PSNR↑	8-views SSIM↑	8-views LPIPS↓
Traditional Methods						
FDK (Feldkamp et al., 1984)	9.71	0.301	0.601	11.05	0.395	0.566
SART (Andersen & Kak, 1984)	15.26	0.627	0.639	15.82	0.655	0.625
ASD_POCS (Sidky & Pan, 2008)	15.68	0.635	0.650	18.27	0.692	0.603
NeRF- and 3DGS-based Methods						
SAX-NeRF (Cai et al., 2024b)	24.08	0.788	0.472	30.01	0.841	0.410
X-Field (Wang et al., 2025)	23.90	0.793	0.539	28.51	0.841	0.494
R^2-Gaussian (Zha et al., 2024)	22.87	0.745	0.476	29.09	0.831	0.349
Supervised Learning Methods						
SEVA (Zhou et al., 2025)	34.12	0.911	0.299	35.51	0.927	0.292
Efficient-LVSM (Jia et al., 2026)	30.96	0.882	0.267	33.11	0.894	0.227
LVSM (Jin et al., 2025)	34.57	0.909	0.237	36.38	0.922	0.219
Ours	36.72	0.931	0.219	38.64	0.944	0.200

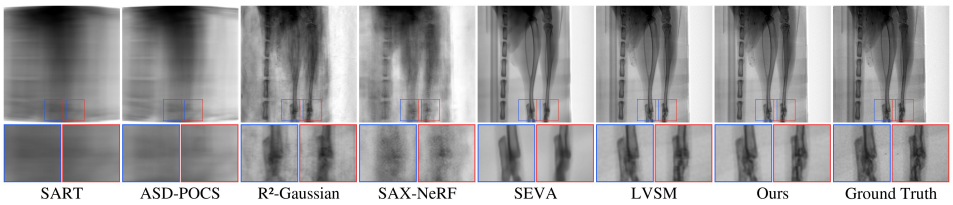

Figure 7: Qualitative comparison of novel view synthesis. Sparse2View produces clearer structural details under sparse-view inputs; please zoom in for a closer inspection.

Table 2: Ablation studies on model architecture.

Component	PSNR↑	SSIM↑	LPIPS↓
w/o HRoPE	37.30	0.935	0.210
w/o Anisotropic Processing	38.08	0.941	0.204
w/o Huber loss	38.45	0.942	0.202
Ours	**38.64**	**0.944**	**0.200**

soft tissues, which results in improved anatomical accuracy in bone shapes, especially near joint surfaces, compared to LVSM.

5.3 Ablation Study

We next examine the contribution of three key design components, i.e., HRoPE, Anisotropic Processing, and Huber loss under 8-view inputs setting. Table 2 presents the results of our ablation studies. We observe that all component contribute remarkable performance improvement. That is, when replacing the self-attention with our HRoPE-MSA, the model improves 1.3 PSNR, demonstrating that the introduced relative positional information enhances generation quality. When incorporating anisotropic processing blocks in place of the standard transformer blocks, the model yields further gains. We attribute this to the 2D inductive bias from the anisotropic design, which facilitates the learning of fine-grained features. Integrating a Huber loss provides an additional boost. Cumulatively, they together lead to the final PSNR of 38.64 dB and SSIM of 0.944.

6 Conclusion

This paper introduces Sparse2View, a novel transformer-based framework for rotatory X-ray novel view synthesis. Extensive experimental results demonstrate its effectiveness in achieving high-fidelity rotatory X-ray novel view synthesis under extremely sparse input settings and enabling 3D CT reconstruction with substantially reduced radiation dose. Sparse2View also shows exceptional capability of generalization to unseen anatomical sites and human organs. The platform demonstrates its potentials in clinical applications, *e.g.*, radiology, disease screening, treatment monitoring, and surgical planning. In addition, we have curated a large, high-quality, human expert-annotated, real-world X-ray projection dataset, a valuable asset for future research in medical settings.

AI USE STATEMENT

We used generative AI tools only for language polishing after we had written the paper. We did not use them for idea generation, literature review, method design, code, experiments, data analysis, or figures. All AI-assisted text was reviewed by the authors, who take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anders H Andersen and Avinash C Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic imaging*, 6(1):81–94, 1984.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5855–5864, 2021.

Beer. Bestimmung der absorption des rothen lichts in farbigen flüssigkeiten. *Annalen der Physik*, 162(5):78–88, 1852. doi: https://doi.org/10.1002/andp.18521620505. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18521620505.

Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani. Tigre: a matlab-gpu toolbox for cbct image reconstruction. *Biomedical Physics & Engineering Express*, 2(5):055010, 2016.

Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative gaussian splatting for efficient x-ray novel view synthesis. In *European Conference on Computer Vision*, pp. 283–299. Springer, 2024a.

Yuanhao Cai, Jiahao Wang, Alan Yuille, Zongwei Zhou, and Angtian Wang. Structure-aware sparse-view x-ray 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11174–11183, 2024b.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pp. 333–350. Springer, 2022.

James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G Willcocks. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *2022 44th annual international conference of the IEEE engineering in medicine & Biology society (EMBC)*, pp. 3843–3848. IEEE, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on communications*, 43(12):2959–2965, 2002.

Lee A Feldkamp, Lloyd C Davis, and James W Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984.

William T Freeman, Edward H Adelson, et al. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Xiaosong Jia, Yihang Sun, Junqi You, Zichen Zou, Junchi Yan, Zuxuan Wu, Yu-Gang Jiang, et al. Efficient-lvsm: Faster, cheaper, and better large view synthesis model via decoupled co-refinement attention. In *International Conference on Learning Representations*, volume 2026, pp. 40029–40046, 2026.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QQBPWtvtcn.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025.

Yingtai Li, Xueming Fu, Shang Zhao, Ruiyang Jin, and S Kevin Zhou. Sparse-view ct reconstruction with 3d gaussian volumetric representation. *arXiv preprint arXiv:2312.15676*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

Julius Plucker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, (155):725–791, 1865.

Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17):4777–4807, 2008.

Rebecca Smith-Bindman, Philip W Chu, Hana Azman Firdaus, Carly Stewart, Matthew Malekhedayat, Susan Alber, Wesley E Bolch, Malini Mahendra, Amy Berrington de González, and Diana L Miglioretti. Projected lifetime cancer risks from current computed tomography imaging. *JAMA internal medicine*, 185(6):710–719, 2025.

H V Sorensen, D Jones, Michael Heideman, and C Burrus. Real-valued fast fourier transform algorithms. *IEEE Transactions on acoustics, speech, and signal processing*, 35(6):849–863, 2003.

Feiran Wang et al. X-field: A physically informed representation for 3d x-ray reconstruction. In *Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight.

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

Philip J Withers, Charles Bouman, Simone Carmignato, Veerle Cnudde, David Grimaldi, Charlotte K Hagen, Eric Maire, Marena Manley, Anton Du Plessis, and Stuart R Stock. X-ray computed tomography. *Nature Reviews Methods Primers*, 1(1):18, 2021.

Shaokai Wu, Yapan Guo, Yanbiao Ji, Jing Tong, Yuxiang Lu, Mei Li, Suizhi Huang, Yue Ding, and Hongtao Lu. More: Multi-organ medical image reconstruction dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12890–12896, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758233. URL https://doi.org/10.1145/3746027.3758233.

Shaokai Wu, Yuxiang Lu, Yapan Guo, Wei Ji, Suizhi Huang, Fengyu Yang, Shalayiding Sirejiding, Qichen He, Jing Tong, Yanbiao Ji, et al. Discretized gaussian representation for tomographic reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25073–25082, 2025b.

Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5438–5448, 2022.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19447–19456, 2024.

Guangming Zang, Ramzi Idoughi, Rui Li, Peter Wonka, and Wolfgang Heidrich. Intratomo: self-supervised learning-based tomography via sinogram synthesis and prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1960–1970, 2021.

Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: neural attenuation fields for sparse-view cbct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 442–452. Springer, 2022.

Ruyi Zha, Tao Jun Lin, Yuanhao Cai, Jiwen Cao, Yanhao Zhang, and Hongdong Li. R^2-gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414. IEEE, 2025.

Table 3: Specifications of the MiceBoneX.

Property	Description
Scale	
Number of Cases	1,614
Total Images	419,640
Acquisition	
Projections	260 (0.8° inc.)
Resolution	
Original	4032 × 2688
Processed	256 × 256
Metadata	
`fxfycxcy`	Camera intrinsics
`w2c`	Camera extrinsics

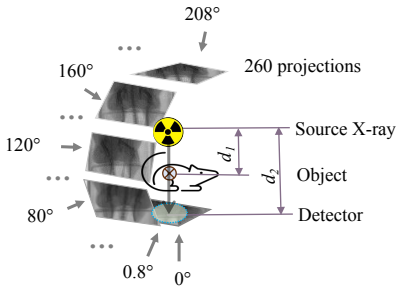

Figure 8: microCT scanning process.

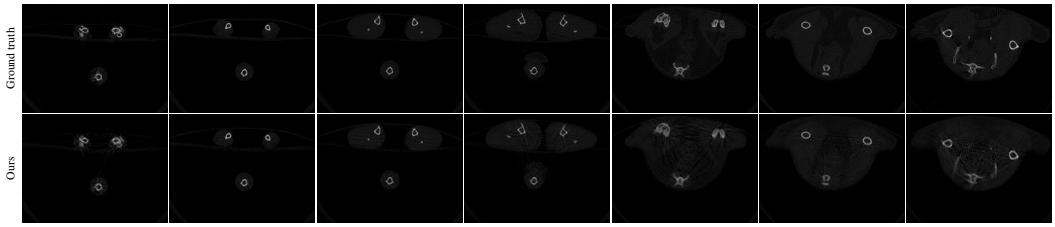

Figure 9: Comparison of reconstructed CT slices at varying axial locations.

A Appendix

A.1 Specifications of the MiceBoneX

The specifications of our MiceBoneX dataset are listed in Table 3. Figure 8 presents a sample case where sequential projections are generated.

A.2 Axial CT Slice Comparisons

We further evaluate how well the CT slices derived from the model agree with ground truth (Figure 9). Due to limited space, we cannot show all of the CT slices (approx. 2,600). Seven evenly spaced slices were selected to demonstrate the high agreement of the model predictions with ground truth from the distal (top row) to the proximal axial locations (bottom row). This spatially consistent performance is critical in clinical settings.

A.3 Generalization to Other Anatomical Regions.

Despite that our MiceBoneX dataset contains only the hindlimb scans around the knee joints, we next examine our model's zero-shot generalization capability to other anatomical regions, Using 8-view inputs from three cases of these unseen anatomical sites, our model still attains a mean PSNR of 35.07 dB and a mean SSIM of 0.908. Figure 10 presents the qualitative results of our zero-shot experiments in comparison with the ground truth X-ray projections. The images in the first two rows show the ankle joints (consisting of distal tibiae and feet), pelvis, and tail as the femurs bent towards the upper body. The 3rd and 4th rows show the ankle joints and tail vertebrates with the lower legs fully extended. Remarkably, our model, even though trained solely on the knee joints X-ray projections, can capture the subtle internal structure of the unseen anatomical regions like pelvis, hip, and feet, rendering its high fidelity synthesis ability. These results demonstrate the strong generalization ability of our model: once trained, it can be readily applied to other anatomical regions without any additional training.

A.4 More Qualitative Results on Fractured bones

For the 21 fractured-object scans, we already present 8 test cases in the main paper. Here, we provide the qualitative results for the remaining test cases, including 13 cases for 8-view zero-shot

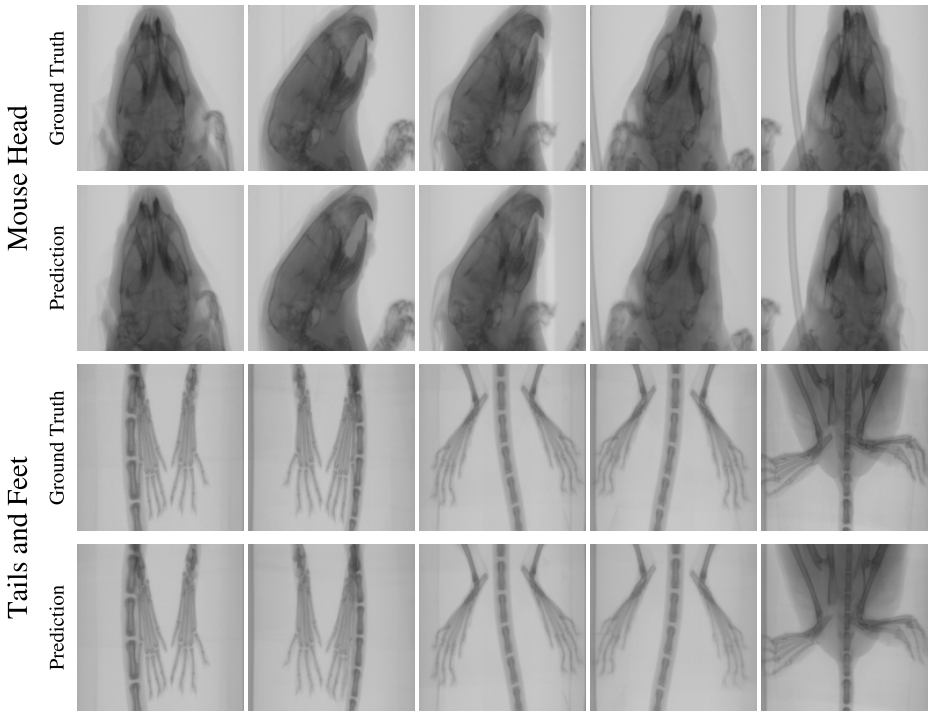

Figure 10: Qualitative results of zero-shot novel X-ray view synthesis of unseen anatomical regions (heads and tails/feet).

setting, 2 cases for 8-view fine-tuning setting (with 11 used for fine-tuning), and 13 cases for 16-view zero-shot setting. Figure 11 present all qualitative results. As shown, our model can accurately reconstruct the fractured structures across all test cases with high fidelity.

A.5 CDFs of PSNR and SSIM for 8v zero-shot, 8v finetune, 16v zero-shot result on fracture data

The CDF of fracture mice of PSNR, SSIM, LPIPS for 8v zero-shot, 8v finetune, 16v zero-shot results on fracture data are shown in Figure 12. With the 8-view input zero-shot experiments, 56.3% of projections achieve a PSNR above 32 and 65.1% attain an SSIM above 0.90. With the 8-view input and fine-tuning, 77.86% of projections achieve a PSNR above 32 and 72.78% achieve an SSIM above 0.90. With the 16-view input zero-shot experiments, 75.25% projections achieve a PSNR above 32 and 76.89% achieve an SSIM above 0.90.

A.6 Novel X-ray View Synthesis on Human Data

We present additional synthesis results on publicly available human organ data in Figure 13 under a zero-shot setting. Here, we take 8 views as input for synthesizing the full set of X-ray projections. The results further demonstrate the robust transferability of our Sparse2View framework. For future practical application, we anticipate that incorporating sufficient human-centric data for model fine-tuning would further enhance the performance and unlock the full potential of the proposed framework in clinical applications.

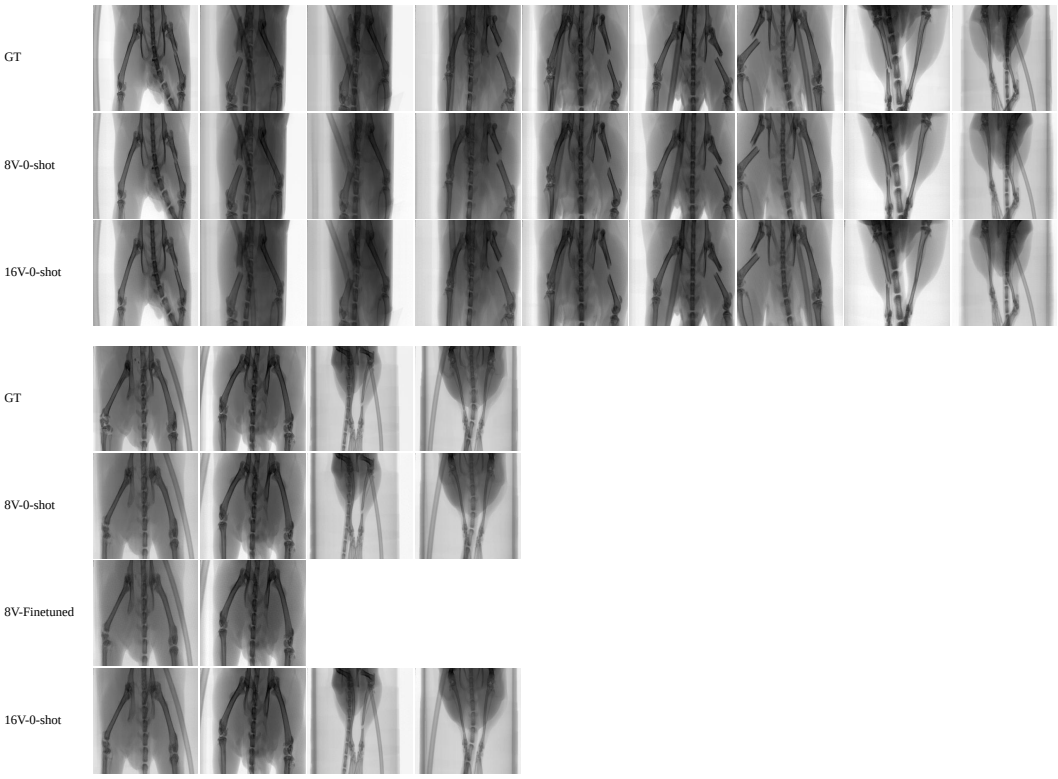

Figure 11: Qualitative results of zero-shot 8v, 16v and finetuned 8v model on novel X-ray view synthesis of remaining fracture set.

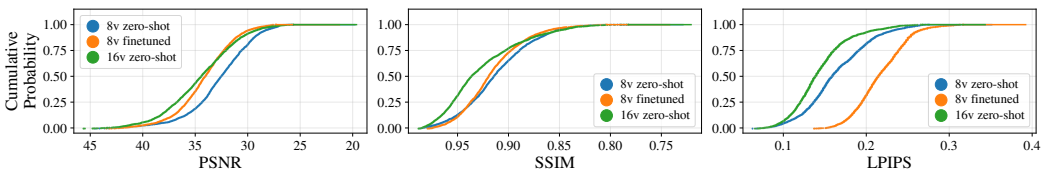

Figure 12: Cumulative Distribution Functions (CDFs) for per-view rendering quality metrics on fracture mice dataset from 8-view 0-shot input; 8-view fine-tuned model input; 16-view 0-shot input.

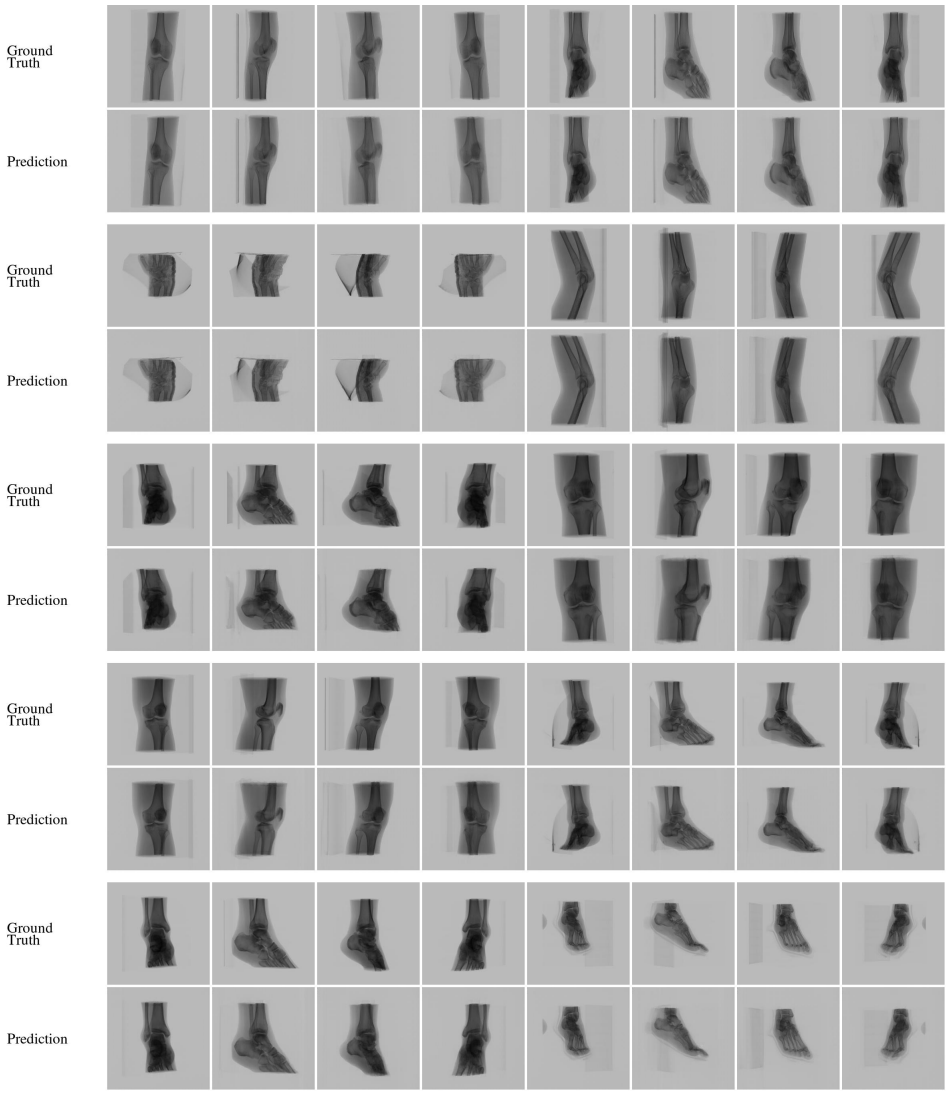

Figure 13: Zero-shot novel view synthesis on human X-ray samples.