

Feasibility Study of a Diffusion-Based Model for Cross-Modal Generation of Knee MRI from X-ray: Integrating External Radiographic Feature Information

Zhe Wang, Yung Hsin Chen, Aladine Chetouani, Fabian Bauer, Yuhua Ru, Fang Chen, Liping Zhang, Rachid Jennane*, Mohamed Jarraya*

Abstract—Knee osteoarthritis (KOA) is a prevalent musculoskeletal disorder, often diagnosed using X-rays due to its cost-effectiveness. While Magnetic Resonance Imaging (MRI) provides superior soft tissue visualization and serves as a valuable supplementary diagnostic tool, its high cost and limited accessibility significantly restrict its widespread use. To explore the feasibility of bridging this imaging gap, we conducted a feasibility study leveraging a diffusion-based model that uses an X-ray image as conditional input, alongside target depth and additional patient-specific feature information, to generate corresponding MRI sequences. Our findings demonstrate that the MRI volumes generated by our approach are not only visually closer to real MRI scans compared with other methods but also achieve the highest quantitative performance in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Furthermore, by increasing the number of inference steps to interpolate between slice depths, we enhance the continuity of the generated volume, achieving higher adjacent slice correlation coefficients. Through ablation studies, we further validate that integrating supplemental patient-specific information, beyond what X-rays alone can provide, enhances the accuracy and clinical relevance of the generated MRI, which underscores the potential of leveraging external patient-specific information to improve the performance of the MRI generation. This study is available at <https://zwang78.github.io/>.

Zhe Wang, Yung Hsin Chen and Mohamed Jarraya are with Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, 02114, USA (e-mail: zwang78@mgh.harvard.edu; ychen4@mgh.harvard.edu; mjarraya@mgh.harvard.edu).

Aladine Chetouani is with L2TI Laboratory, University Sorbonne Paris Nord, Villejuif, 93430, France (e-mail: aladine.chetouani@sorbonne-paris-nord.fr).

Fabian Bauer is with Division of Radiology, German Cancer Research Center, Heidelberg, 69120, Germany (e-mail: fabian.bauer@dkfz-heidelberg.de).

Yuhua Ru is with Jiangsu Institute of Hematology, The First Affiliated Hospital of Soochow University, Suzhou, 215006, China. (e-mail: ruyuhua@suda.edu.cn).

Fang Chen is with department of Medical School, Henan University of Chinese Medicine, Zhengzhou, 450046, China. (e-mail: chen-fangxy@hactcm.edu.cn).

Liping Zhang is with Athinoula A. Martinos Centre for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, 02114, USA. (e-mail: lzhang90@mgh.harvard.edu).

Rachid Jennane is with IDP Institute, UMR CNRS 7013, University of Orleans, Orleans, 45067, France (e-mail: rachid.jennane@univ-orleans.fr).

Index Terms—Knee osteoarthritis; X-ray; Magnetic Resonance Imaging; Diffusion-based

I. INTRODUCTION

KNEE OsteoArthritis (KOA) is a progressive disease marked by the degeneration and damage of articular cartilage, changes at the joint margins, and reactive hyperplasia of the subchondral bone [1]. Various factors, such as age, obesity, stress, and trauma, can contribute to its development [2]. Those affected often endure severe pain and limited mobility, which can greatly diminish their quality of life and elevate the risk of chronic conditions like cardiovascular disease [3]. Despite extensive research, the precise cause of KOA remains unknown, and there is currently no cure [4]. Traditionally, X-ray imaging has been the most common method for diagnosing KOA due to its accessibility and ability to reveal changes in bone structure, such as Joint Space Narrowing (JSN) and osteophyte formation [5]. To assess KOA on X-rays, five distinct grades (i.e., from KLG-0 to KLG-4), determined by the presence and severity of symptoms, are commonly used [6]. However, X-ray imaging has limitations, particularly in visualizing soft tissue structures and early cartilage changes. Magnetic Resonance Imaging (MRI), on the other hand, provides a more comprehensive assessment of the knee joint by offering detailed images of both bone and soft tissues, including cartilage, menisci, and synovium [7]. MRI can detect early signs of KOA that are not visible on X-rays, such as bone marrow lesions, synovitis, and early cartilage degeneration [8]. However, the widespread adoption of MRI in the clinical management of KOA is hindered by several challenges, most notably the high cost associated with MRI scans [9]. The expense of MRI technology, including the costs of the equipment, maintenance, and specialized personnel, makes it less accessible compared to traditional X-ray imaging [10]. This economic barrier is particularly significant in resource-limited settings and for patients without adequate health insurance coverage. In addition to the high financial cost, MRI scans require more time and patient cooperation, as they necessitate the patient to remain still for extended periods [11], which can be difficult for those experiencing severe pain or limited mobility due to KOA.

To address these limitations, researchers have explored various techniques to enhance MRI reconstruction and reduce its dependency on traditional methods. Fessler et al. [12] provides an in-depth review of iterative algorithms used for improving MRI image reconstruction, particularly in scenarios where traditional inverse Fourier transforms are inadequate. The author highlights the limitations of conventional methods, especially when dealing with non-Cartesian sampling, under-sampled data, or the presence of nonlinear magnetic fields, and then focuses on model-based approaches, which utilize iterative methods that incorporate physical models of MRI acquisition, allowing for better image quality by addressing inhomogeneity and under-sampling issues. The authors demonstrate that model-based methods can significantly enhance image reconstruction, providing superior results compared to traditional techniques, though often at the cost of increased computation. In [13], Han et al. proposed leveraging the k-space to improve the efficiency and accuracy of MRI reconstruction. The authors introduced a framework inspired by the low-rank Hankel matrix completion approach, which directly interpolates missing k-space data and enhances the representation of MRI data by learning the underlying signal structure, resulting in high-quality images even from undersampled data. The proposed k-space deep learning approach significantly reduces the computational complexity and memory requirements. In [14], Lustig et al. leveraged compressed sensing to significantly accelerate MRI scan times without compromising image quality. The authors explore the inherent sparsity of MR images, both in the image domain and in the transform domains. By randomly undersampling k-space and using nonlinear recovery schemes, the study shows that high-quality MR images can be reconstructed from incomplete data. This approach effectively reduces acquisition time and mitigates artefacts that arise from undersampling, which demonstrates that compressed sensing has the potential to substantially reduce scan times in clinical settings, making MRI more efficient and accessible.

With the rapid advancements in deep learning technologies, the field of MRI reconstruction has seen significant growth [15] [16]. Recently, combining multi-modal and diffusion-based models has become a new trend for image generation. In [17], Zhan et al. presented a novel medical generative model, MedM2G, designed to unify multiple medical modalities (e.g., CT, MRI, X-ray) and tasks (e.g., text-to-image, image-to-text, modality translation) into a single, efficient framework. The authors address the challenge of aligning and generating medical multi-modal data, particularly in the absence of large, well-paired datasets. They propose a central alignment strategy to efficiently align different modalities using text as the central modality and introduce medical visual invariant preservation to retain key clinical knowledge from each modality. The model leverages a multi-flow cross-guided diffusion process that facilitates adaptive interaction among modalities, enhancing the generation of accurate and high-quality medical images. In [18], Jiang et al. introduced a novel diffusion-based approach to multi-modal MRI synthesis. Traditional diffusion models for MRI synthesis typically operate in the original image domain, leading to high memory consumption and less effective handling of anatomical structures. CoLa-Diff

addresses these challenges by operating in the latent space to reduce memory overhead while preserving anatomical details. The model introduces structural guidance using brain region masks and an auto-weight adaptation mechanism to balance multiple input modalities effectively during synthesis. The authors demonstrate that CoLa-Diff outperforms existing state-of-the-art models in multi-modal MRI synthesis tasks, showing superior results in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). In [19], Kim et al. propose an advanced method for translating MRI images between different modalities. The authors introduce an Adaptive Latent Diffusion Model (ALDM) that leverages a novel multiple-switchable spatially adaptive normalization (MS-SPADE) block to transform source latent representations into target-like latent. This approach enables high-quality image synthesis across multiple modalities without requiring patch cropping, preserving the global information of the images. By conditioning the model with target modalities and applying the ALDM in a compressed latent space, the proposed method significantly reduces computational complexity while achieving superior results in both one-to-one and one-to-many image translation tasks.

Inspired by the above work, we propose a novel approach based on the Conditional Latent Diffusion Model (CLDM) [20]. Specifically, during the training phase, each real MRI slice is first Gaussian noised and then progressively denoised using the corresponding X-ray as conditional input y , target depth d and radiographic features r of the specific patient as the guidance module to reconstruct the input MRI slice. In the inference phase, given a knee X-ray image, a target depth d and the corresponding radiographic features r , the model progressively denoises from Gaussian noise to generate the corresponding MRI slice at that depth. Then, the target depth is adjusted and the denoising process is repeated iteratively for each depth until all the required slices are synthesized. These individual slices are then stacked together to form a fully generated 3D MRI volume. It is noteworthy that the reconstructed MRI is not intended to serve as a clinical replacement for a real MRI. Nevertheless, it signifies a groundbreaking feasibility study in integrating patient-specific radiographic information with knee X-rays to generate corresponding MRI volumes.

The primary contributions of this study are summarized as follows:

- A novel CLDM-based framework is proposed for generating MRI volumes from X-ray images as the foundation for our feasibility study.
- Beyond the target depth, patient-specific radiographic features not present in the X-ray are integrated as guidance to improve the denoising process.
- Ablation studies demonstrate that incorporating additional patient-specific information enhances the accuracy and clinical relevance of the generated MRI.
- By increasing the number of inference steps (i.e., exceeding the depth of the original MRI volume), the model achieves more effective interpolation between MRI slices, allowing for the generation of intermediate anatomical structures.

II. METHODOLOGY

A. Classical conditional latent diffusion model

Before delving into the proposed approach, we provide a brief overview of the classic conditional latent diffusion model, which serves as the foundation for our methodology. As shown in Fig. 1, a conditional latent diffusion model is a type of generative model used for producing data (such as images, audio, or text) based on certain conditions or inputs. This model combines the strengths of diffusion models and latent variable models, allowing it to generate complex data distributions while being guided by specific conditions. Different modules are to be introduced, respectively.

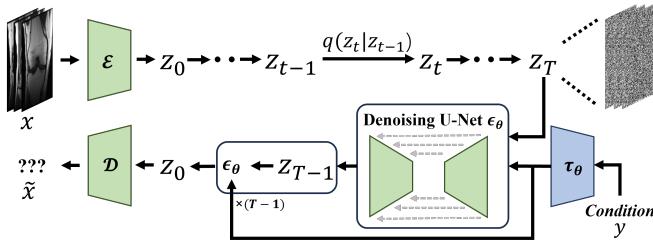


Fig. 1: The structure of the classical conditional latent diffusion model.

1) Encoding process: The latent variable model is used to map the input image x to a latent space using an encoder:

$$z_0 = \mathcal{E}(x) \quad (1)$$

where \mathcal{E} is an encoder which maps an image x to a latent representation z_0 .

2) Diffusion process: A diffusion process describes the evolution of a random variable over time. Through the forward diffusion process, for the original latent representation $z_0 \sim q(z_0)$, each step of the diffusion process, which comprises a total of T steps, involves adding Gaussian noise ϵ to the data obtained from the previous step z_{t-1} as follows:

$$q(z_t | z_{t-1}) = \mathcal{N}(z_t; \sqrt{1 - \beta_t} z_{t-1}, \beta_t \mathbf{I}) \quad (2)$$

where β_t represents the variance used for each step in the range of $[0, 1]$, and \mathbf{I} represents the identity matrix. The entire diffusion process constitutes a Markov chain [21]:

$$q(z_T | z_0) = \prod_{t=1}^T q(z_t | z_{t-1}) \quad (3)$$

Here, we directly sample z_t for any step t based on the original data z_0 : $z_t \sim q(z_t | z_0)$ defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Through the reparameterization technique [22], we have:

$$\begin{aligned} z_t &= \sqrt{\alpha_t} z_{t-1} + \sqrt{1 - \alpha_t} \epsilon_{t-1} \\ &= \sqrt{\alpha_t} (\sqrt{\alpha_{t-1}} z_{t-2} + \sqrt{1 - \alpha_{t-1}} \epsilon_{t-2}) + \sqrt{1 - \alpha_t} \epsilon_{t-1} \\ &= \sqrt{\alpha_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \forall \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \quad (4)$$

3) Denoising process: After the diffusion process, denoising is performed by sampling from the updated noise distribution. Estimating distribution $q(x_{t-1} | x_t)$ requires the utilization of the entire training set. Typically, a neural network, such as a U-Net [23], is employed to estimate these distributions. Here, the reverse process is also defined as a Markov chain composed of a sequence of Gaussian distributions parameterized by neural network parameters:

$$p_\theta(z_{t-1} | z_t, y) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, y), \sigma_\theta^2(t)) \quad (5)$$

where $p_\theta(z_{t-1} | z_t, y)$ represents parameterized Gaussian distributions, the mean $\mu_\theta(z_t, t, y)$ and variance $\sigma_\theta^2(t)$ at each time step with the condition y are determined by trained learning networks.

4) Training process: The encoder-decoder and diffusion modules are trained separately. For the encoder-decoder, the reconstruction loss \mathcal{L}_{rec} is as:

$$\mathcal{L}_{rec} = \|(x, \mathcal{D}(\mathcal{E}(x)))\|_2^2 \quad (6)$$

The parameters θ of the diffusion module aim to minimize the distance between the true noise ϵ and the predicted noise ϵ_θ . The diffusion loss \mathcal{L}_{diff} quantifies this difference as the expected value of the squared norm between the actual noise and the noise estimated by the neural network:

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0 \sim \mathcal{E}(x), y, t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2 \quad (7)$$

where \mathbb{E} represents the expectation over all data points under the joint distribution of the conditioning information y , time steps t , and Gaussian noise ϵ .

B. Proposed approach

As presented in Fig. 2, our approach consists of three main modules: an auto-encoder module, a CLDM module, and a guidance module. x_d is a slice of the original knee MRI sequence x at the depth position d with the size of $H \times W \times S$. \mathcal{E}_1 and \mathcal{D} are encoder and decoder of the AutoencoderKL [22]. Unlike traditional autoencoders that map input data to a deterministic latent space, the AutoencoderKL introduces a probabilistic approach by encoding the input into a Gaussian distribution over the latent space. During training, the Kullback-Leibler (KL) divergence is used as a regularization term to ensure that the learned latent space closely follows a standard normal distribution, promoting smoother and more continuous transitions between latent representations. Specifically, the encoder operates at a resolution of $256 \times 256 \times 3$ and begins with a convolutional layer configured with 128 channels and then proceeds through a series of stages, each with 2 residual blocks. The channel dimensions are multiplied according to the sequence 1, 2, 4, and 4. The decoder mirrors the structure of the encoder. The architecture is designed with no dropout, focusing on retaining all features without stochastic regularization during training. Simultaneously, the noised latent code z_T is calculated following Eq. 4:

$$z_T = \sqrt{\bar{\alpha}_t} \mathcal{E}_1(x_d) + \sqrt{1 - \bar{\alpha}_T} \epsilon \quad (8)$$

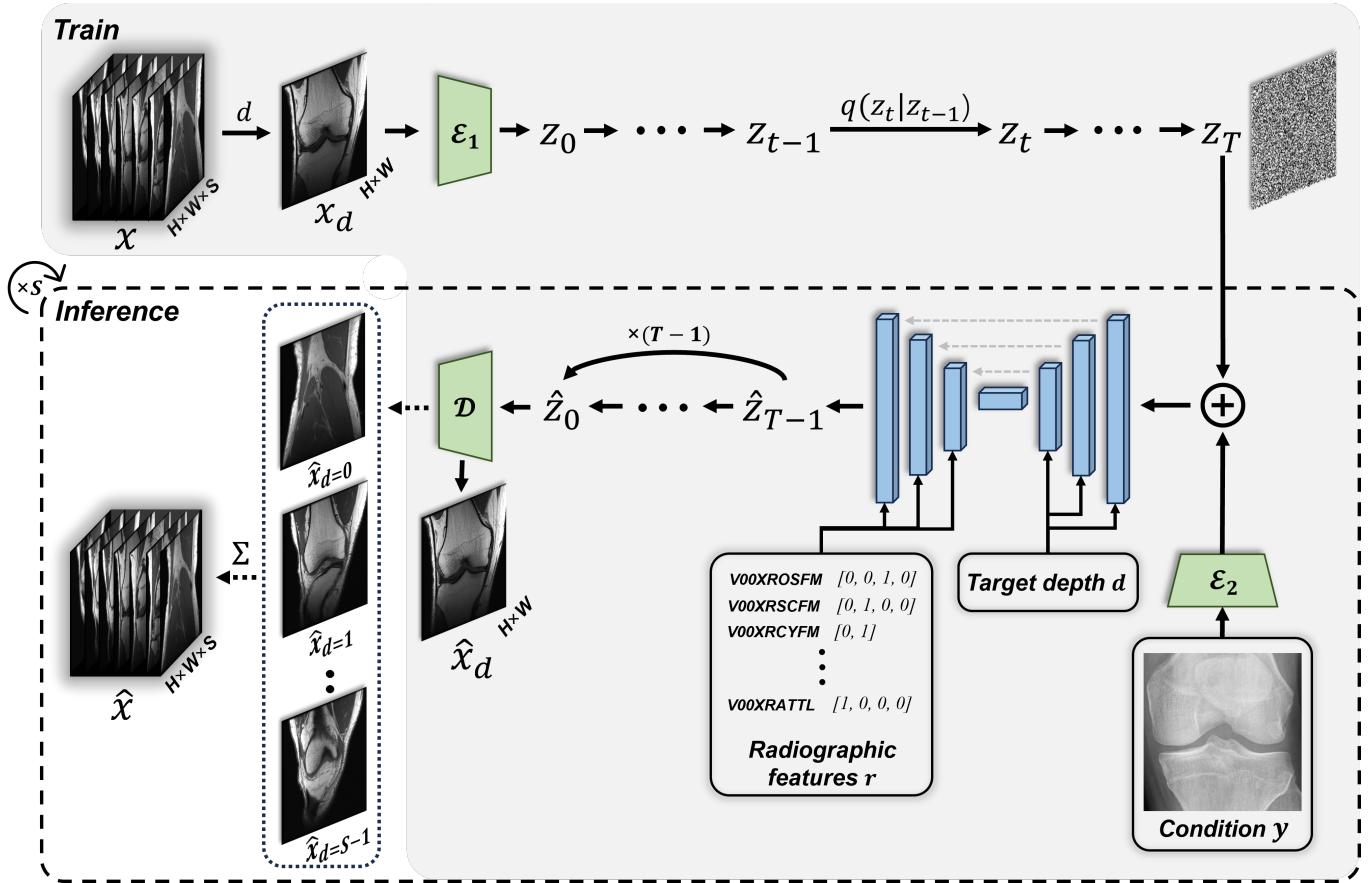


Fig. 2: The flowchart of the proposed approach, starts by extracting a slice at position d from a real MRI sequence x , denoted as x_d . This slice is encoded by the encoder \mathcal{E}_1 to produce a latent representation z_0 , which is progressively noised following the distribution $q(z_t|z_{t-1})$, ultimately generating the noised latent representation z_T . The corresponding X-ray image y is then used as a conditional input, encoded by \mathcal{E}_2 , and concatenated with z_T to form the initial input for the downsampling network. During the denoising process, the target depth d and the patient-specific radiographic feature information r are jointly embedded at various stages of the downsampling and upsampling phases, collectively providing denoising guidance. This denoising process is repeated for $T - 1$ iterations to progressively refine the latent representation, resulting in the denoised latent \hat{z}_0 , which is decoded by \mathcal{D} to reconstruct the MRI slice \hat{x}_d at depth d . During inference, the process begins with z_T and is repeated S times, corresponding to the number of channels S in the original MRI sequence. Finally, the full generated MRI sequence \hat{x} is obtained by stacking all the generated MRI slices in sequential order (i.e., $\hat{x} = \{\hat{x}_d\}_{d=0}^S$).

Conditional image y is a corresponding knee joint X-ray image, which is then encoded by the encoder \mathcal{E}_2 and concatenated with the z_T as the input of the U-Net, serves as the beginning of the reverse process. The employed 2D U-Net architecture adopts a multi-scale hierarchical structure comprising a time embedding module, an encoder module, a middle module, and a decoder module. This U-Net model takes 8-channel input and produces 4-channel output, built upon a base channel configuration of 320. The encoder module begins with a 2D convolutional layer featuring a kernel size of 3 and padding of 1, followed by a series of residual blocks with progressively increasing feature dimensions. Each residual block includes multiple convolutional layers, GroupNorm [24] for normalization, the Swish activation function [25] for non-linearity, and optional dropout for regularization. Two residual blocks are employed per downsampling stage, with attention mechanisms activated at resolutions of 4, 2, and 1. The channel count increases across network layers following a multiplier

sequence of 1, 2, 4, and 4. The decoder mirrors this structure, utilizing the same residual blocks to progressively upsample feature maps, reducing the number of channels in reverse order while merging them with corresponding encoder feature maps through skip connections. Additionally, the model incorporates 8 attention heads and employs a spatial transformer with a depth of 1, conditioned on a context dimension of 768. Next, the target depth d is embedded into the downsampling stage to allow the model to condition the latent representation with spatial hierarchy early in the process, which ensures that depth-related features. In the upsampling stage, the patient-specific radiographic feature information r (refer to Table I) is embedded to guide the synthesis of fine clinical features based on the spatial hierarchy established during downsampling. The depth d is normalized to the range $[0, 1]$, and the patient-specific radiographic feature information r is encoded using a one-hot encoding scheme.

During the denoising process, the latent variable z_{t-1} is

TABLE I: Description of radiographic features

Variable	Description
V00XROSF M	Osteophytes (OARSI 0-3), femur medial compartment
V00XRSCFM	Sclerosis (OARSI 0-3), femur medial compartment
V00XRCYFM	Cysts (OARSI 0-1), femur medial compartment
V00XRJSM	Joint space narrowing (OARSI 0-3), medial compartment
V00XRCHM	Chondrocalcinosis (Grades 0-1), medial compartment
V00XROSTM	Osteophytes (OARSI 0-3), tibia medial compartment
V00XRSCSTM	Sclerosis (OARSI 0-3), tibia medial compartment
V00XRCYTM	Cysts (OARSI 0-1), tibia medial compartment
V00XRATTM	Attrition (OARSI 0-1), tibia medial compartment
V00XRKLL	Kellgren and Lawrence (0-4) for osteoarthritis severity
V00XROSFL	Osteophytes (OARSI 0-3), femur lateral compartment
V00XRSCFL	Sclerosis (OARSI 0-3), femur lateral compartment
V00XRCYFL	Cysts (OARSI 0-1), femur lateral compartment
V00XRJSLL	Joint space narrowing (OARSI 0-3), lateral compartment
V00XRCHL	Chondrocalcinosis (Grades 0-1), lateral compartment
V00XROSTL	Osteophytes (OARSI 0-3), tibia lateral compartment
V00XRSCTL	Sclerosis (OARSI 0-3), tibia lateral compartment
V00XRCYTL	Cysts (OARSI 0-1), tibia lateral compartment
V00XRATTL	Attrition (OARSI 0-1), tibia lateral compartment

* OARSI: Osteoarthritis Research Society International Grading System

sampled using the Denoising Diffusion Implicit Model (DDIM) [26], which enables deterministic or stochastic generation depending on the chosen schedule. DDIM refines the sampling process by approximating the reverse diffusion trajectory more efficiently than traditional diffusion models. Specifically, it leverages a non-Markovian process that skips some diffusion steps, allowing for faster inference without compromising the quality of the generated images. The sampling of z_{t-1} is computed as follows:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta \quad (9)$$

As presented before, the diffusion model is implemented through a U-Net \mathcal{U} , which is trained based on Eq. 7. In this study, the goal function is computed as follows:

$$\min_\theta \mathbb{E}_{z \sim \epsilon_1(x_d), t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} [\| \epsilon - \epsilon_\theta[z_t, t, (z_T, y, d, r)] \|_2^2] \quad (10)$$

Finally, each reconstructed slice at the depth d , denoted as \hat{x}_d , is generated through a decoder \mathcal{D} . During the inference phase, this process is repeated S times, corresponding to the depth (i.e., channel) of the original MRI volume. The final generated MRI volume \hat{x} is formed by stacking all generated slices, represented as $\hat{x} = \{\hat{x}_d\}_d^S$. For more clarity, Algorithm 1 and Algorithm 2 describe the training and inference processes of the proposed approach, respectively.

III. EXPERIMENTAL DATA AND SETTINGS

This section provides an overview of the experimental data, the preprocessing steps undertaken, and the specifics of the experimental procedures.

Algorithm 1: Training process of the approach

```

Input:  $x, t, y, r, S, \mathcal{E}_2$ , and all initial parameters
Output:  $\hat{x}_d, \mathcal{E}_1, \mathcal{D}, \mathcal{U}$  and learned parameters
while not converged do
    for  $d$  in  $S$  do
        Get  $x_d \leftarrow x[d]$ 
        Get  $\hat{x}_d \leftarrow \mathcal{D}(\mathcal{E}_1(x_d))$ 
        Compute  $\mathcal{L}_{rec}(x_d, \hat{x}_d)$  ▷ Eq. 6
        Update Parameters of  $\mathcal{E}_1$  and  $\mathcal{D}$ 
    end
end
Freeze Parameters of  $\mathcal{E}_1$  and  $\mathcal{D}$ 
while not converged do
    for  $d$  in  $S$  do
        Get  $x_d \leftarrow x[d]$ 
        Get  $\epsilon, z_T \leftarrow noise(\mathcal{E}_1(x_d), t)$  ▷ Eq. 4
        Get  $\epsilon_\theta, \hat{z}_d \leftarrow \mathcal{U}(z_T, \mathcal{E}_2(y), t, d, r)$  ▷ Eq. 10
        Compute  $\mathcal{L}_{diff}(\epsilon, \epsilon_\theta)$ 
    end
end

```

Algorithm 2: Inference process of the approach

```

Input:  $t, r, y, S, \mathcal{E}_2, \mathcal{D}$  and  $\mathcal{U}$ 
Output:  $\hat{x}_d$  and  $\hat{x}$ 
Get  $\epsilon \leftarrow \mathcal{N}(0, \mathbf{I})$ 
for  $d$  in  $S$  do
    Get  $\hat{z}_d \leftarrow \mathcal{U}(\epsilon, \mathcal{E}_2(y), t, d, r)$ 
    Get  $\hat{x}_d \leftarrow \mathcal{D}(\hat{z}_d)$ 
end
Get  $\hat{x} \leftarrow \{\hat{x}_d\}_d^S$ 

```

A. Employed database

The OsteoArthritis Initiative (OAI) [27] represents a significant and rich source of data for researchers investigating KOA and related conditions. By analyzing data from 4,796 participants aged between 45 and 79 years over 96 months, the initiative provides a comprehensive longitudinal dataset. Each participant underwent nine follow-up examinations, allowing for detailed tracking of the disease's progression and the identification of potential risk factors associated with KOA development or progression.

B. Data preprocessing

1) **X-ray:** As shown in Fig. 3, the knee joint (Fig. 3b) was identified from the plain radiograph (Fig. 3a) through the YOLOv2 learning model [28], with images sized of $299 \times 299 \times 1$, which were resized to the size of $256 \times 256 \times 1$, and image intensity distribution was normalized to $[-1, 1]$.

2) **MRI:** After normalization and standardization, each MRI slice has a size of $256 \times 256 \times 1$. To meet the requirements of the pre-trained diffusion model with a channel size of 3, the grayscale images were duplicated three times and then concatenated along the channel dimension. Moreover, the distribution of slice counts was analyzed across three pulse sequences of Intermediate-Weighted (IW), T1-weighted (T1),

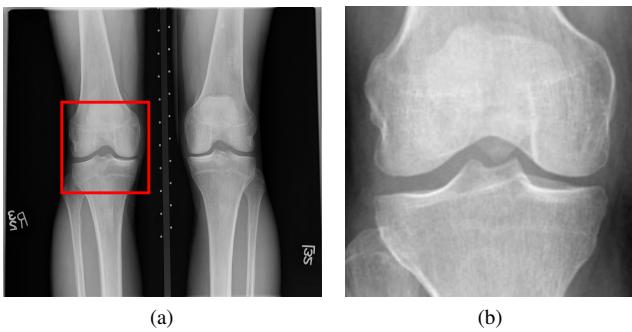


Fig. 3: An identified knee joint is detected and highlighted with a red box in a standard knee plain radiograph (a). The cropped Region Of Interest (ROI) containing the knee joint is shown separately (b).

and Magnetization Prepared Rapid Acquisition Gradient Echo (MPR). As shown in Fig. 4a, the slice count for the T1 sequence remains almost consistently fixed at 80, the MPR sequence displays a broader distribution, ranging from 60 to 80 slices, while the IW sequence has the lowest slice counts, between 30 and 40, which suggests that T1 slices are more uniform and contain richer information for model learning. To ensure consistency in our study, we focused exclusively on T1-weighted MRI scans with 80 slices as the original dataset. Fig. 4b illustrates the image intensity distribution within each T1 MRI sequence. As can be seen, grey values rise sharply at first, peaking around the 0.15 proportional slice position. After this peak, the grey values drop steeply by the 0.4 proportional slice position and then gradually taper off, eventually returning to levels similar to those of the initial slices. Based on this and subjective assessment, we extracted slices between the 16% to 78% proportional positions (i.e., from the 13th slice to the 63rd one) from each MRI sequence to ensure consistency and avoid low information slices at the beginning and end of the sequence.

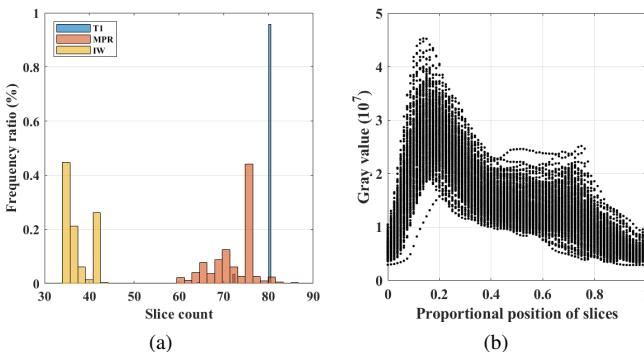


Fig. 4: Distribution of the number of slices for each type of MRI sequence (a). Grey value distribution at each proportional position of slices within the T1-weighted MRI sequences (b).

3) Radiographic features: The encoding process for the radiographic features involves transforming each feature into a One-Hot vector representation based on its grades. For missing

values, all positions in the One-Hot vector are set to zero to indicate the absence of valid information for that feature, which ensures that the model can distinguish missing values from valid inputs without introducing artificial bias. The resulting One-Hot vectors for all radiographic features are concatenated to form a unified representation, which is subsequently used as part of the guidance input during the denoising process in the diffusion model.

After the above processing, the slice count was fixed at 50 for the T1-weighted MRI scans as the final experimental dataset. The resulting 4,262 X-ray-MRI pairs with respective radiographic feature information were then randomly divided into training, validation and testing sets with a ratio of 7:1:2.

C. Experimental details

The diffusion module utilized pre-trained weights from Zero-123 [29]. The pre-trained encoder \mathcal{E}_2 is sourced from [30]. During the training, a base learning rate of 1e-06 was applied, with a lambda linear scheduler incorporating a warm-up period every 100 steps. The batch size was set to 64, with T configured to 2,000 steps, and an Exponential Moving Average (EMA) [31] of 0.99 was used for the diffusion sampling process. The scale factor of the latent space was set as 0.2. The AdamW optimizer [32] was utilized for training over 5,000 epochs. We implemented our approach using PyTorch v1.8.1 [33] on Nvidia A100 graphics cards with 80 GB of memory.

IV. RESULTS AND DISCUSSION

A. Comparisons with State-of-the-Art Methods

The comparison is divided into two main parts: Qualitative visualization analysis and quantitative metric-based analysis.

1) Qualitative visualization analysis: Table II provides a comprehensive visual comparison of corresponding MRI slices generated by various methods, using X-ray images as input. To facilitate a clear and focused analysis, we selected three representative stages of KOA: G-0, indicating healthy knees; G-2, reflecting early-stage KOA; and G-4, corresponding to late-stage KOA. The compared methods include classic LDM [34], Cola-diff [18], X-diff [35], and our proposed approach. Each row displays the X-ray input on the far left, followed by the real MRI slice and the MRI slices generated by the respective methods. To provide a more intuitive illustration of each method's performance, difference maps generated using a Sobel operator filter are shown beneath each slice, highlighting pixel-wise deviations from the ground truth MRI by detecting edge differences. The visual results consistently show that our model produces generated MRI slices that are closer to the ground truth, with significantly fewer and less pronounced deviations compared to other methods. Specifically, LDM suffers from the lack of constraints provided by the input X-ray. Without this conditional guidance, the model struggles to capture the detailed structural information of the knee joint, leading to great deviations from the ground truth MRI. On the other hand, compared with LDM, although Cola-Diff and X-Diff generate bone structures that are closer to the real MRI in terms of overall shape, they both lack a significant amount of fine details. This limitation prevents

TABLE II: Visualization of a representative generated MRI slice

	X-ray input	MRI Ground Truth	LDM	Cola-diff	X-diff	Ours
G-0						
		Error map				
G-2						
		Error map				
G-4						
		Error map				

* To make it more intuitive and convenient, we only display one slice from each MRI sequence that best represents the complete knee joint.

them from accurately capturing and reproducing the KOA symptoms. Moreover, it is noteworthy that it is particularly evident as the KOA grade of the input X-ray increases. The progressive narrowing of the knee joint space and the formation of osteophytes are learned and generated more accurately by our model, which demonstrates that our method not only ensures the generation of high-quality MRI but also maintains clinical relevance. However, the model's predictions in areas outside the skeletal regions are less accurate due to the absence of critical information in X-ray images such as muscle and fat

distribution, which will be discussed in Section IV-D.4.

2) Quantitative metric-based analysis: PSNR is a measure used to assess the quality of reconstructed or compressed images compared to their original versions. On the other hand, SSIM is a widely used metric for evaluating the visual similarity between two images considering human visual perception by measuring changes in luminance, contrast, and structural information. However, it does not account for the specific Regions Of Interest (ROI) that are critical in medical imaging. In knee MRI volumes, the most relevant diagnostic information

often lies within cartilage, joints, and other areas affected by disease. Therefore, in addition to SSIM, we also adopt a specially designed indicator, the so-called Region-specific Structural Similarity Index (RSSIM), to represent the clinical evaluation of bone morphology. To do so, we utilize the Canny operator with a sigma of 3 to filter and extract bone edges, followed by the application of SSIM. Specifically, we focus on the lower third of the femoral edge to the upper third of the tibial edge to capture ROIs most likely to contain areas of potential pathologies, such as JSN and osteophytes. We compared the performance of the proposed approach with the previously selected methods. Fig. 5 illustrates the statistical distribution of these metrics. As can be seen in Fig. 5a, our proposed method has the highest median PSNR, indicating better average image quality, with a median slightly above 29.2 dB. Cola-diff and X-diff also demonstrate a high median around 28.9 dB, with LDM being the lowest, around 28.6 dB. This suggests that, based on the PSNR metric, the image quality performance differences among the four methods are relatively minor. On the other hand, Fig. 5b reveals that while our method achieves the highest median SSIM value, approximately 0.28, the differences between our method, Cola-diff, and X-diff are not substantial, mirroring the trend seen with PSNR. However, our method exhibits a significant advantage under the RSSIM metric. With a median RSSIM value close to 0.5, our approach outperforms the closest competitor, X-diff, which only reaches around 0.42, representing a roughly 20% improvement, which underscores that our approach's ability to maintain image quality while providing a stronger focus on clinically relevant regions in MRI sequences.

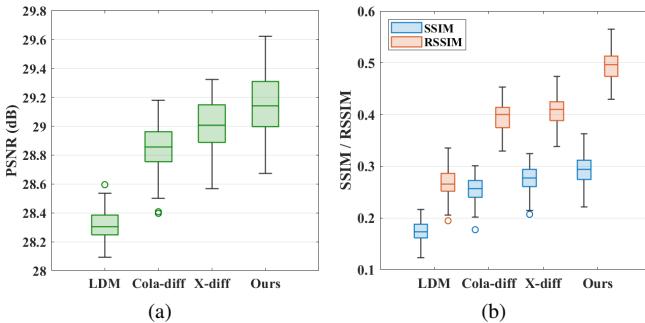


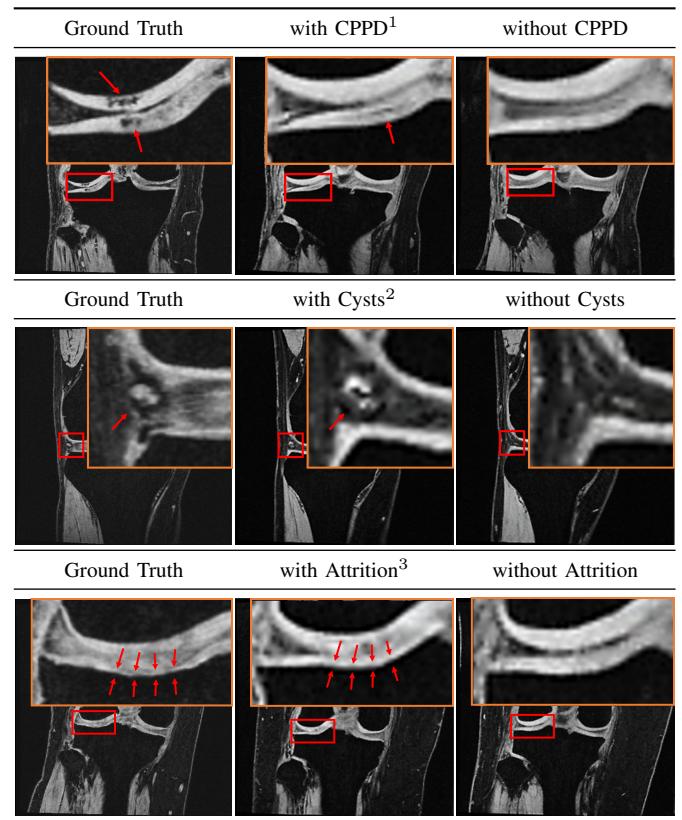
Fig. 5: The box plots show the different performance of the evaluated approaches using PSNR (a) and SSIM/RSSIM (b).

B. Ablation study for feature information

To assess the contribution of the integrated patient-specific radiographic feature information r within the guidance module, we conducted a detailed ablation study, analyzing the impact of specific variables within r . This ablation study was performed using the same dataset configuration to ensure consistency and comparability across experiments. Here, we particularly focused on the roles of the variables of chondrocalcinosi (CPPD), attrition, and cysts in Table I, given their clinical relevance and the challenges associated with interpreting these features from X-ray images alone. For instance, CPPD is associated with

the deposition of calcium pyrophosphate crystals in the joint tissues, which manifests as subtle calcifications, which often overlap with normal joint structures, leading to ambiguity in detection using Xrays. Attrition, on the other hand, reflects the loss of articular cartilage and underlying bone, a hallmark of joint degeneration. Its appearance on X-rays can be subtle. Similarly, cystic changes represent fluid-filled cavities or bone erosions that occur in advanced joint diseases, these changes can appear as radiolucent areas on X-rays, and their true extent and clinical implications are difficult to discern without complementary information. By isolating these variables, we aimed to determine the extent to which they enhance the model's ability to synthesize MRI outputs that accurately reflect the underlying pathology.

TABLE III: Ablation studies for feature variable integration



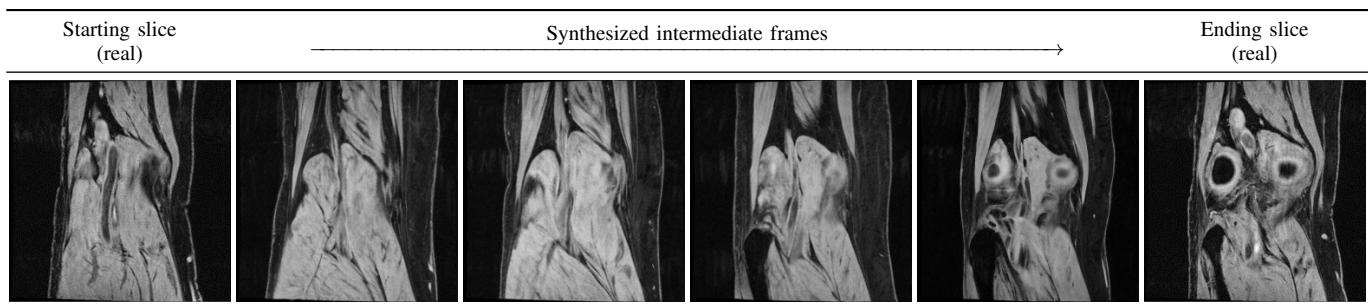
¹ V00XRCHL=1: CPPD present in lateral compartment

² V00XRCYTL=1: Cysts present in the tibia lateral compartment

³ V00XRATTL=1: Attrition present in tibia lateral compartment

As shown in Table III, before integrating the CPPD variable (V00XRCHL), the model generated MRI sequences that entirely lacked any representation of CPPD, as it was unable to incorporate this feature without explicit guidance. However, after integrating CPPD as a guidance variable, the model began to display CPPD in a roughly accurate region (i.e., the lateral compartment). For the attrition, integrating the corresponding variable (V00XRATTL) enabled the model to better capture signs of cartilage loss and bone alterations in the tibia lateral compartment. Similarly, for the cysts, the variable (V00XRCYTL) allowed the model to successfully synthesize fluid-filled regions in the tibia lateral compartment.

TABLE IV: Visualization of the synthesized intermediate frames between adjacent real MRI slices



It is noteworthy that, while the model struggled to fully capture the precise severity and spatial details of certain features, the ablation studies demonstrate that integrating patient-specific radiographic information beyond what X-rays alone provide improves the anatomical realism and clinical relevance of the generated MRI sequences, allows the model to generate MRI outputs that are, to some extent, closer to real MRI scans, which underscores the immense potential of incorporating additional information (e.g., Body Mass Index (BMI), neural distribution, etc) as guidance into such models.

C. Impact of the number of inference steps

For the inference steps s , in addition to adhering to the original depth setting (i.e., $s = S = 50$), we also experimented with varying inference steps (i.e., 30, 100, 200, 300, 400, and 500), to explore the impact of increasing the number of inference steps on the quality and accuracy of the generated slices.

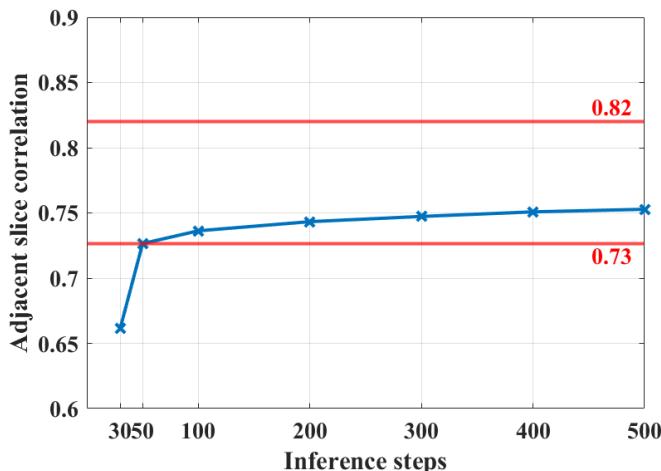


Fig. 6: The adjacent slice correlation coefficients under different inference steps.

Fig. 6 shows the adjacent slice correlation coefficients under different inference steps. As can be seen, the red horizontal lines at 0.73 and 0.82 represent the mean adjacent slice correlation coefficients of the generated MRI matching the volume of the original MRI (i.e., $s = S = 50$) and that of the original MRI itself, respectively, with the latter serving as the upper benchmark for comparison. At 30 inference steps, the adjacent slice correlation starts at approximately 0.66, reflecting the

impact of downsampling, which introduces inconsistencies and reduces the alignment between neighbouring slices. As the number of inference steps increases to 50, where the depth aligns with that of the original MRI volume, the correlation coefficient jumps sharply to around 0.73, suggesting a significant improvement in slice coherence. At 100 steps, the correlation continues to increase modestly. Beyond 100 steps, the improvement becomes more gradual as the number of inference steps increases, with diminishing returns, eventually stabilizing around 0.76 by 500 inference steps. Table IV illustrates the synthesized intermediate frames between the original adjacent images (i.e., starting and ending images) using 500 inference steps. A distinct representative stage is evaluated: The transition from predominantly muscle tissue to the initial appearance of bone. In this stage, the synthesized intermediate frames display relatively smooth and continuous transitions between bone and soft tissue structures, which underscores the robustness and effectiveness of these interpolation frames.

Furthermore, we conducted a visualization experiment for interpolation at $s = 500$, focusing on a representative stage where the transition occurs from predominantly muscle tissue to the initial appearance of bone. This stage is particularly important as it reflects the model's ability to generate anatomically coherent structures in regions with complex tissue variations. Table IV presents the synthesized intermediate frames between the original adjacent images (i.e., starting and ending images). As observed, these intermediate frames exhibit smooth and continuous transitions between bone and soft tissue structures, effectively bridging the morphological differences. It is noteworthy that the model successfully preserves fine anatomical details and structural consistency throughout the interpolation process, demonstrating its potential to generate high-quality, clinically meaningful intermediate frames.

The above results demonstrate that increasing the number of inference steps s enhances interpolation, showcasing its potential for improving temporal resolution. By generating additional intermediate frames from a smaller set of real frames, interpolation effectively offers more continuous data without lengthening the actual scan duration. On the other hand, despite the improvements observed, the correlation coefficient from the interpolated MRI fails to reach the 0.82 benchmark of the original MRI volume, even with the maximum 500 inference steps, which indicates that, while the interpolation process can enhance the consistency between adjacent slices, inherent limitations remain. These limitations arise from the inherent

constraints of using X-rays as the sole imaging source of information, which will be discussed in Section IV-D.4.

D. Discussion

In this paper, we conduct a feasibility study of a diffusion-based model for cross-modal generation of knee MRI from X-ray. Specifically, the model leverages conditional inputs from X-rays, combined with target depth and patient-specific feature information, to guide the synthesis process for the MRI slice generation. Experimental results demonstrate that, compared to SOTA methods, our approach generates pseudo-MRI sequences that are closer to real MRI scans in terms of image quality and clinical relevance. Furthermore, ablation studies on the guidance module validate that incorporating supplementary information beyond what is available from X-rays alone improves the accuracy of the generated MRI, which highlights the immense potential for integrating additional independent information in the future to further enhance MRI generation performance. Lastly, by increasing the number of inference steps, the model exhibits effective interpolation, resulting in smoother transitions between slices.

Several points are to be discussed.

1) How it works?: In our approach, we move away from focusing on the distinct imaging principles and techniques of X-ray and MRI. Instead, we only concentrate on the mapping relationship between these two imaging modalities, much like an extremely experienced senior radiologist who, when viewing an X-ray image, could to some extent mentally infer what the corresponding MRI scans might reveal, which is our proposed AI model's aim. To this end, for the model's training, vast amounts of high-quality paired X-ray and MRI have significantly propelled this.

2) Methodological contribution: While our approach is grounded in the established framework of conditional latent diffusion models, its methodological novelty lies not in a fundamental architectural reinvention, but in its unique adaptation to the complex challenge of cross-modal image generation from heterogeneous clinical data. Specifically, our key contribution is the integration of patient-specific radiographic feature information, data not visually present in the source X-ray, as an explicit guidance module within the denoising process. As validated by our ablation studies (Section IV-B), this tailored guidance is critical for synthesizing clinically relevant pathologies, such as chondrocalcinosis, attrition, and cysts, which are otherwise absent in the generated output. Therefore, this work should be viewed as a pioneering feasibility study that introduces a novel conditioning paradigm for this specific clinical problem, demonstrating how disparate data sources can be synergistically combined to bridge the information gap between common imaging modalities.

3) Why use a 2D diffusion model instead of a 3D one?: We opted to use a 2D diffusion model, generating individual MRI slices at different target depths during the inference phase, which are then stacked to form a 3D MRI sequence. By using this method, the model can effectively learn and capture the spatial relationships between the individual slices, allowing it to understand inter-slice dependencies. Additionally, employing

a 2D model reduces computational complexity, making both the training and inference processes more efficient compared to handling full 3D data. This approach also enhances the interpretability of the entire process, as it enables a more granular analysis of how each slice contributes to the final 3D reconstruction.

4) Strengths and limitations: This study has several notable strengths. Firstly, it introduces an innovative method for generating MRI from one single X-ray image, attempting to bridge the gap between these two imaging modalities using advanced AI technology, which represents a novel attempt in cases where MRI may not be readily available. Our experiments demonstrate that incorporating supplementary information (i.e., patient-specific radiographic feature information in this study) beyond the X-ray images themselves enhances the quality of the generated MRI, which underscores the potential of integrating more diverse information sources in the future to produce more accurate generated MRI volumes. Second, the study employs a 2D diffusion model, which not only reduces the computational resources required but also allows for more detailed depth control over the generated images, leading to a more interpretable and transparent reconstruction process. Finally, the study's focus on clinically relevant metrics, including both traditional image quality assessments and pathology-specific evaluations, ensures that the generated MRIs are not only visually similar to ground truth but also clinically relevant. While not a replacement for diagnostic MRI, this technology holds potential for future applications such as aiding in patient triage in resource-limited settings or generating augmented data for training other AI models. This study has several limitations that warrant consideration. Our model are highly dependent on the availability of large, high-quality paired datasets. The scarcity of such datasets and handcrafted features, particularly in certain clinical scenarios or for rare diseases, may limit the generalizability of our approach. Moreover, the computational resources required for training, despite being reduced by using a 2D model, high-performance hardware is still necessary, which presents a challenge for broader implementation in clinical settings.

5) Future work: For future work, a more explicit mathematical formalization for 3D continuity could be a valuable direction for future work involving 3D-aware architectures or specific continuity-promoting regularization terms, while balancing computational feasibility. Furthermore, more additional information (e.g., Body Mass Index (BMI), age, gender, etc) could be integrated to further enhance the generation performance. Self-supervised learning, uncertainty quantification, developing uncertainty-aware evaluation and cross-database validation could also be of interest.

V. STATEMENTS

This manuscript was prepared using data from the OAI. The views expressed in it are those of the authors and do not necessarily reflect the opinions of the OAI investigators, the National Institutes of Health (NIH), or the private funding partners.

VI. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Ralph Schlaeger Research Fellowship under award number 246448 from Massachusetts General Hospital (MGH), Harvard Medical School (HMS), and the French National Agency of Research (ANR) under project number ANR-20-CE45-0013-01.

- [1] R. F. Loeser, S. R. Goldring, C. R. Scanzello, and M. B. Goldring, "Osteoarthritis: a disease of the joint as an organ," *Arthritis and rheumatism*, vol. 64, no. 6, p. 1697, 2012.
- [2] A. Litwic, M. H. Edwards, E. M. Dennison, and C. Cooper, "Epidemiology and burden of osteoarthritis," *British Medical Bulletin*, vol. 105, no. 1, pp. 185–199, 01 2013. [Online]. Available: <https://doi.org/10.1093/bmb/lds038>
- [3] G. Singh, J. D. Miller, F. H. Lee, D. Pettitt, and M. W. Russell, "Prevalence of cardiovascular disease risk factors among us adults with self-reported osteoarthritis: data from the third national health and nutrition examination survey," *Population*, vol. 7, p. 17, 2002.
- [4] J. C. Mora, R. Przkora, and Y. Cruz-Almeida, "Knee osteoarthritis: pathophysiology and current treatment modalities," *Journal of pain research*, vol. 11, p. 2189, 2018.
- [5] P. A. Sukerkar and Z. Doyle, "Imaging of osteoarthritis of the knee," *Radiologic Clinics*, vol. 60, no. 4, pp. 605–616, 2022.
- [6] J. H. Kellgren and J. S. Lawrence, "Radiological Assessment of Osteo-Arthritis," *Annals of the Rheumatic Diseases*, vol. 16, no. 4, p. 494, 1957.
- [7] F. Eckstein, F. Cicuttini, J.-P. Raynauld, J. C. Waterton, and C. Peterfy, "Magnetic resonance imaging (mri) of articular cartilage in knee osteoarthritis (oa): morphological assessment," *Osteoarthritis and cartilage*, vol. 14, pp. 46–75, 2006.
- [8] C. L. Piccolo, C. A. Mallio, F. Vaccarino, R. F. Grasso, and B. B. Zobel, "Imaging of knee osteoarthritis: a review of multimodal diagnostic approach," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 11, p. 7582, 2023.
- [9] F. Reyes-Santias, C. García-García, B. Aibar-Guzmán, A. García-Campos, O. Cordova-Arevalo, M. Mendoza-Pintos, S. Cinza-Sanjurjo, M. Portela-Romero, P. Mazón-Ramos, and J. R. Gonzalez-Juanatey, "Cost analysis of magnetic resonance imaging and computed tomography in cardiology: A case study of a university hospital complex in the euro region," in *Healthcare*, vol. 11, no. 14. MDPI, 2023, p. 2084.
- [10] R. A. Bell, "Economics of mri technology," *Journal of Magnetic Resonance Imaging*, vol. 6, no. 1, pp. 10–25, 1996.
- [11] B. McLean and D. Thompson, "MRI and the critical care patient: clinical, operational, and financial challenges," *Critical Care Research and Practice*, vol. 2023, no. 1, p. 2772181, 2023.
- [12] J. A. Fessler, "Model-based image reconstruction for mri," *IEEE signal processing magazine*, vol. 27, no. 4, pp. 81–89, 2010.
- [13] Y. Han, L. Sunwoo, and J. C. Ye, " k -space deep learning for accelerated mri," *IEEE transactions on medical imaging*, vol. 39, no. 2, pp. 377–386, 2019.
- [14] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.
- [15] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo *et al.*, "Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1310–1321, 2017.
- [16] K. Bahrami, F. Shi, X. Zong, H. W. Shin, H. An, and D. Shen, "Reconstruction of 7t-like images from 3t mri," *IEEE transactions on medical imaging*, vol. 35, no. 9, pp. 2085–2097, 2016.
- [17] C. Zhan, Y. Lin, G. Wang, H. Wang, and J. Wu, "Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11502–11512.
- [18] L. Jiang, Y. Mao, X. Chen, X. Wang, and C. Li, "Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis," *arXiv preprint arXiv:2303.14081*, 2023.
- [19] J. Kim and H. Park, "Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7604–7613.
- [20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [21] J. R. Norris, *Markov chains*. Cambridge university press, 1998, no. 2.
- [22] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 2015, pp. 234–241.
- [24] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [25] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.
- [26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
- [27] G. Lester, "The Osteoarthritis Initiative: A NIH Public–Private Partnership," *HSS Journal: The Musculoskeletal Journal of Hospital for Special Surgery*, vol. 8, no. 1, pp. 62–63, 2011.
- [28] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.
- [29] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 9298–9309.
- [30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
- [31] J. S. Hunter, "The exponentially weighted moving average," *Journal of quality technology*, vol. 18, no. 4, pp. 203–210, 1986.
- [32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: <https://arxiv.org/abs/1711.05101>
- [33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019.
- [34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [35] E. Bourigault, A. Hamdi, and A. Jamaludin, "X-diffusion: Generating detailed 3d mri volumes from a single image using cross-sectional diffusion models," *arXiv preprint arXiv:2404.19604*, 2024.