KNEEFORCE: LONGITUDINAL KNEE OSTEOARTHRITIS X-RAY PREDICTION VIA DIFFUSION FORCING WITH TEMPORAL RELATIONAL ALIGNMENT AND TRAJECTORY SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knee Osteoarthritis (KOA) is a serious chronic disease leading to disability, and anticipating its progression is essential for timely, personalized intervention. Existing prediction methods largely forecast non-image indicators such as future Kellgren-Lawrence (KL) grades or pain scores. However, these scalar outcomes provide only coarse, subjective assessments and lack the fine-grained visual details that clinicians need to guide treatment. Directly predicting future radiographs offers a more objective and interpretable alternative. Yet longitudinal forecasting remains challenging. Existing generative approaches struggle to maintain anatomical consistency over long horizons due to error accumulation and rigid inference. Moreover, they overlook the irregular sampling intervals inherent to clinical follow-up. To overcome these challenges, we propose **KneeForce**, a novel framework that adapts Diffusion Forcing for continuous disease trajectory modeling. Our method explicitly encodes the time interval to accommodate unevenly sampled inputs, and incorporates Temporal Relational Alignment, which leverages a DINO-based knee foundation model as a semantic expert to guide the generative process toward high-level disease patterns rather than mere texture synthesis. Additionally, a Trajectory Selection strategy is carefully designed to filter stochastic outputs, identifying the most clinically probable progression path. Extensive validation on the public OAI dataset and an external in-house dataset demonstrates that KneeForce consistently outperforms state-of-the-art medical and video synthesis methods in both radiographic fidelity and clinical grading accuracy. A blinded study by experienced clinicians further confirms the clinical reliability of our predictions, offering trustworthy visual evidence to support long-term prognosis.

1 INTRODUCTION

Knee osteoarthritis (KOA) is a serious chronic disease and a leading cause of disability, affecting over 600 million individuals globally (Cui et al., 2020). Initially presenting with knee pain and swelling, KOA progressively degrades the joint and ultimately leads to deformity and functional impairment. As no therapy can reverse established structural damage, total knee replacement remains the only option at end stage. Thus, early and personalized intervention is essential to slow progression and improve long-term outcomes (Mahmoudian et al., 2021).

To this end, existing research has extensively employed logistic regression and deep learning to forecast KOA progression (Riddle et al., 2016; Tiulpin et al., 2019a; Leung et al., 2020) in the form of **non-image clinical indicators**, most commonly future Kellgren-Lawrence (KL) grades (Kohn et al., 2016) or pain scores (Galvan-Tejada et al., 2014). While straightforward, such scalar outcomes provide only coarse, ordinal summaries of severity and are inherently subjective, as radiographic grading suffers from inter-reader variability. Crucially, a single grade cannot localize where and how the joint will change, and therefore fails to capture the fine-grained morphological evolution—such as joint space narrowing and osteophyte formation—that clinicians actually inspect when planning treatment. This motivates a fundamentally different paradigm: rather than regress-

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.

ing an abstract score, we can **directly predict the future radiograph**, providing an objective and readily interpretable visual reference for personalized early intervention. The rapid progress of generative models, from GANs (Singh & Raza, 2021; Armanious et al., 2020) to diffusion models (Ho et al., 2020; Kazerouni et al., 2023), has made this vision attainable: image-to-image methods such as Pix2Pix (Isola et al., 2017) and LDM (Rombach et al., 2022) can synthesize a future X-ray at a specified time-point from a baseline image, offering clinicians rich visual evidence for decision-making.

Nevertheless, KOA is a chronic disease monitored through serial follow-up rather than a single snapshot. Predicting an isolated time-point ignores the trajectory, while jointly modeling a sequence of radiographs captures the momentum of progression, better matching clinical practice. Early attempts along this line remain limited. The KOA-tailored method SLE (Han et al., 2022) performs nearest-neighbor latent extrapolation in a StyleGAN space but attains only limited accuracy. In the broader medical domain, methods such as SADM (Yoon et al., 2023) and ImageFlowNet (Liu et al., 2025a) predict brain-MRI progression by compressing historical scans or explicitly modeling latent trajectories, yet they transfer poorly to the subtle, localized changes of KOA. At the general-vision level, generic video generation models like SEINE (Chen et al., 2024b) can be repurposed for the task but optimize for perceptual smoothness while ignoring the diversity of patient-specific clinical outcomes. Consequently, existing frameworks still struggle to produce clinically reliable, long-term radiographs that faithfully reflect an individual's disease trajectory.

Recently, Diffusion Forcing (Chen et al., 2024a) has emerged as a promising sequence-generation framework that assigns each frame an independent noise level, enabling flexible, asynchronous denoising schedules at inference. This flexibility is well aligned with the inherently non-linear and uncertain nature of longitudinal medical prediction. However, directly applying Diffusion Forcing to KOA prediction is non-trivial due to three primary challenges. First, Diffusion Forcing is agnostic to the elapsed time between frames. It implicitly assumes uniformly spaced observations, whereas real-world follow-up radiographs are acquired at highly irregular intervals; without an explicit notion of elapsed time, the model cannot tell whether a transition spans one year or five, leading to mis-calibrated progression rates. Second, there is a lack of temporal feature consistency. Naive diffusion objectives defined on raw pixels tend to prioritize high-frequency textures over high-level dynamics (Choi et al., 2022), so the model easily reproduces plausible bone texture while failing to capture the global pathological evolution of the knee. Third, the stochasticity of generation is a hurdle for clinical use. The flexible schedule can yield multiple anatomically plausible trajectories, only some of which match the most clinically probable progression, so a mechanism is needed to filter candidates and select the clinically aligned one.

To address these challenges, we propose **KneeForce**, a novel framework tailored for longitudinal KOA prediction. First, we introduce time-interval encoding that injects the elapsed time between consecutive visits into the generative process, allowing the model to align its predicted progression with the true, unevenly sampled follow-up schedule. Second, we introduce Temporal Relational Alignment (TRA) during training: using OrthoFoundation (Yu et al., 2026), a DINO (Siméoni et al., 2025) model pre-trained on a large corpus of knee radiographs, as a domain expert, we align the inter-frame similarity matrices of the diffusion model with those of the expert, distilling high-level structural-evolution priors rather than mere texture. Third, we propose a Trajectory Selection strategy at inference: it generates multiple candidate trajectories under varying schedules and employs a lightweight ranking model to select the most clinically probable one, mitigating the inherent stochasticity of generation.

We evaluate KneeForce on both the public OAI (Nevitt et al., 2006) dataset and an in-house dataset, covering an 8-year follow-up period. Extensive experiments show that KneeForce consistently outperforms state-of-the-art medical baselines and general video diffusion methods in both image fidelity and clinical grading accuracy. In addition, a blinded reader study conducted by experienced clinicians is designed to assess the real-world clinical plausibility of the generated trajectories, where KneeForce attains the best mean expert rank and is preferred in 80.0% of cases. Our main contributions are summarized as follows:

- We introduce **KneeForce**, adapting Diffusion Forcing to longitudinal KOA radiographic prediction and equipping it with an explicit time-interval encoding that accommodates the irregular sampling of clinical follow-up, enabling reliable long-term prognosis.

- To shift the modeling focus from mere texture synthesis to high-level disease progression patterns, we propose **Temporal Relational Alignment (TRA)**, which distills semantic guidance from a domain-specific foundation model.
- At inference, a **Trajectory Selection** mechanism is designed to mitigate stochastic uncertainty and ensure clinically reliable predictions, validated with quantitative metrics and a blinded clinical reader study.

2 Related Work

Medical Longitudinal Image Prediction. Early approaches synthesize a future scan from a baseline image at a specified time-point using image translation or diffusion models (Isola et al., 2017; Rombach et al., 2022; Kazerouni et al., 2023). More recent methods model disease progression across visits: SADM (Yoon et al., 2023) conditions sequence-aware diffusion on historical scans, ImageFlowNet (Liu et al., 2025a) learns latent disease trajectories, and BrLP (Puglisi et al., 2024) and TADM (Litrico et al., 2024) model individual-level or temporally coherent brain-image progression. Δ-LFM (Chen et al., 2026) uses flow matching to learn patient-specific latent changes and generate future MRIs at arbitrary target times. For knee osteoarthritis (KOA), SLE (Han et al., 2022) extrapolates radiographs in a StyleGAN latent space; other work uses cycle-consistent translation (Prezja et al., 2025) or identity-consistent diffusion (Wu et al., 2024). KneeForce instead jointly generates a long-horizon radiographic sequence from observed visits, accounts for irregular follow-up times, and selects among plausible progression trajectories.

Generative Models for Image and Video Synthesis. GANs (Singh & Raza, 2021; Armanious et al., 2020) and diffusion models (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022) provide the foundations for image synthesis. Video diffusion (Ho et al., 2022) and image-to-video models such as SEINE (Chen et al., 2024b) extend generation to temporally coherent sequences. Their objectives, however, do not specifically emphasize the subtle, patient-dependent structural changes that determine clinical progression in KOA.

Autoregressive Diffusion and Sequence Generation. Autoregressive generation can accumulate errors over long rollouts, whereas parallel diffusion typically denoises all frames under a shared schedule. Diffusion Forcing (Chen et al., 2024a) assigns each frame an independent noise level, enabling asynchronous denoising and flexible rollouts. Subsequent work further explores asynchronous schedules and long-horizon synthesis (Sun et al., 2025; Huang et al., 2025; Liu et al., 2025b). We adapt this flexibility to longitudinal imaging, where observations have unequal time intervals and future disease courses are uncertain.

Representation Guidance for Diffusion Models. Pixel-level diffusion objectives can favor texture over high-level structure (Choi et al., 2022). Representation alignment addresses this limitation: REPA (Yu et al., 2025), for example, guides diffusion features using a pretrained visual encoder such as DINO (Siméoni et al., 2025). KneeForce draws instead on a knee-radiograph foundation model (Yu et al., 2026) and aligns inter-frame feature relationships to guide disease progression across the generated sequence.

3 Method

3.1 Problem Formulation

We formulate the longitudinal prediction of KOA as a sequence prediction task. For a given patient's knee, we denote the sequence of X-ray images at T sequential time-points as $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$, which represents the disease progression trajectory. Given a set of s initial context frames $\{x_1, \ldots, x_s\}$, the objective of longitudinal prediction is to generate the subsequent trajectory $\{x_{s+1}, \ldots, x_T\}$.

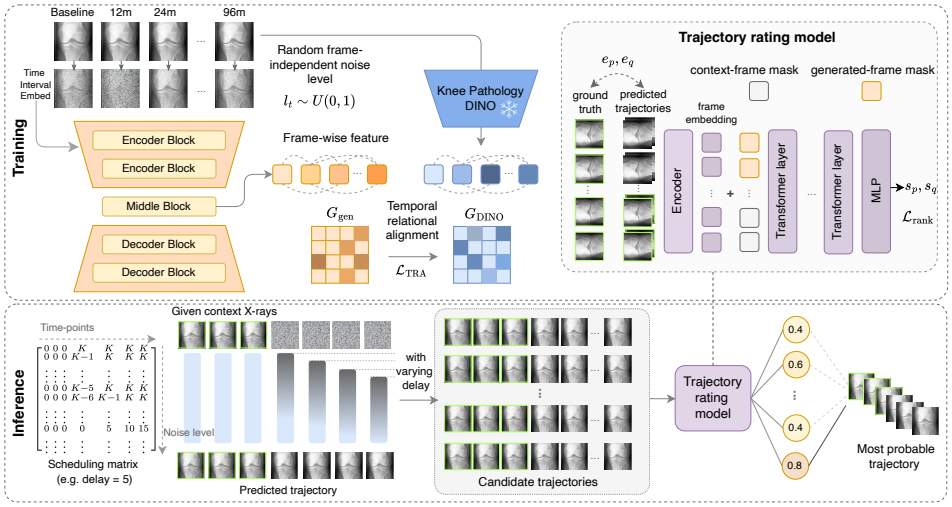

Figure 1: Overview of KneeForce. (Top) Training: We employ a Diffusion Forcing backbone augmented with a time-interval encoding that conditions each frame on its absolute acquisition time (in months since baseline), enabling the model to account for the irregular follow-up intervals in longitudinal KOA imaging. We further introduce Temporal Relational Alignment (TRA), which distills disease progression patterns from a frozen knee-DINO expert by aligning inter-frame similarity matrices. (Bottom) Inference: We adopt a Trajectory Selection strategy. Diverse candidate trajectories are generated via varying diffusion delays controlled by the scheduling matrix, and the trajectory rating model selects the most clinically plausible outcome.

3.2 Diffusion Forcing Backbone

Diffusion Forcing (DF) (Chen et al., 2024a) is a sequence modeling paradigm. During training, it treats noise as a generalized masking mechanism, independently corrupting each frame with a noise level $k_t \sim U(0,1)$ (where 1 is pure noise and 0 is clean). The model recursively updates a latent state $z_t = f_\theta(z_{t-1}, x_t^{k_t})$ that encapsulates the history $x_{1:t}$, from which the denoised distribution $p_\theta(x_t|z_t)$ is predicted.

During inference, this per-frame noise decoupling enables flexible and asynchronous sampling governed by a *scheduling matrix*. As illustrated in the Inference pipeline of Fig. 1, rather than relying on strict autoregressive completion, we introduce a constant delay step d between adjacent frames. Specifically, frame $t+1$ begins its denoising process only after frame t has completed d denoising steps, creating a staggered schedule. The choice of d directly controls the generation dynamics: a larger d strongly conditions subsequent frames on highly-denoised preceding frames, reducing uncertainty but limiting diversity; conversely, a smaller d allows for a more stochastic progression. By varying the delay d, we can dynamically sample a diverse set of candidate trajectories for subsequent evaluation.

Time-Interval Encoding. The vanilla DF backbone indexes frames only by their ordinal position t, implicitly assuming that consecutive observations are uniformly spaced. This assumption is violated in longitudinal KOA imaging, where follow-up radiographs are acquired at irregular calendar intervals. To remedy this, we augment DF with an explicit *time-interval encoding* that conditions each frame on its acquisition time. Concretely, we associate every frame x_t with an absolute time-point $\tau_t \in \mathbb{Z}_{\geq 0}$, measured in months relative to the baseline scan ($\tau_1 = 0$). For a standard OAI trajectory, this yields $\{\tau_t\} = \{0, 12, 24, 36, 48, 72, 96\}$, which naturally encodes the non-uniform gaps between visits (e.g., a 12-month step from 36m to 48m versus a 24-month step from 48m to 72m). We then turn each scalar time-point τ_t into a continuous embedding e_t in two steps: we first lift the scalar to a multi-frequency feature vector via a fixed sinusoidal mapping (Vaswani et al., 2017), so that even small differences in elapsed time are represented at multiple scales rather than as a single raw magnitude. This feature is then passed through a learnable MLP that projects it to

the model's embedding space,

$$e_t = \mathrm{MLP}\big(\mathrm{PE}(\tau_t)\big).$$

The resulting e_t is injected into the recurrent update so that the latent state becomes $z_t = f_\theta(z_{t-1}, x_t^{k_t}, e_t)$. This makes the denoising of each frame explicitly aware of *how much time has elapsed* since its predecessor, allowing the model to modulate the magnitude of predicted anatomical change according to the true, unevenly sampled follow-up schedule. Because the encoding operates on absolute time-points rather than frame indices, it also enables prediction at arbitrary target intervals at inference, including intervals unseen during training.

3.3 Temporal Relational Alignment

The overall training objective of DF is optimized using a standard denoising loss $\mathcal{L}_{\text{denoise}}$, which reconstructs the target frame from its noisy observation (Chen et al., 2024a). However, diffusion models trained on raw pixels often prioritize high-frequency radiographic textures (Choi et al., 2022). For KOA progression modeling, capturing global clinical biomarkers (e.g., osteophyte formation) and the correct structural evolution over time (e.g., joint space narrowing) is far more critical. To instill disease-specific semantic priors and shift the generative focus from mere texture synthesis to high-level temporal dynamics, we introduce a Temporal Relational Alignment (TRA) loss. The total training objective is thus formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{denoise}} + \lambda \mathcal{L}_{\text{TRA}},$$

where $\mathcal{L}_{\text{TRA}}$ encourages the diffusion model to learn structural priors of temporal evolution from a domain-specific expert. We adopt OrthoFoundation (Yu et al., 2026) as this expert–a DINO (Siméoni et al., 2025) model pre-trained on an extensive corpus of knee radiographs, equipping it with rich knee pathology knowledge.

However, directly aligning the diffusion features to a static DINO expert is overly restrictive and may hinder generative flexibility. To address this, we propose to align the relative similarities between frames across the sequence. Specifically, we capture this through a Gram matrix G, where each element of this matrix is calculated via the cosine similarity between frame-wise features $r_{1:T}$, given by $G_{i,j} = \frac{r_i \cdot r_j}{\|r_i\| \|r_j\|}$. For an input sequence $\mathcal{X}$, we extract intermediate representations $\mathcal{R}^{\text{gen}} = \{r_1^{\text{gen}}, \ldots, r_T^{\text{gen}}\}$ from the middle block of the diffusion U-Net and corresponding semantic embeddings $\mathcal{R}^{\text{DINO}} = \{r_1^{\text{DINO}}, \ldots, r_T^{\text{DINO}}\}$ from the frozen DINO expert, yielding respective matrices G^{gen} and $G^{\text{DINO}} \in \mathbb{R}^{T \times T}$. Finally, $\mathcal{L}_{\text{TRA}}$ is formulated as the squared distance between two matrices:

$$\mathcal{L}_{\text{TRA}} = ||G^{\text{gen}} - G^{\text{DINO}}||_F^2.$$

By aligning the relative similarity matrix rather than raw features, we effectively transfer the expert's clinical progression knowledge into the diffusion model.

3.4 Trajectory Selection

Trajectory Generation. The stochasticity in Diffusion Forcing's flexible sequence modeling presents a challenge for precise clinical forecasting. For a given sample, manipulating the temporal dependency through different sampling schedules (varying the delay parameter d) yields multiple candidate trajectories. Thus, we can acquire a candidate set $\mathcal{X}_{\text{cand}} = \{\mathcal{X}^{(j)}\}_{j=1}^{|D|}$, where $|D|$ is the number of delay configurations. While each trajectory represents a distinct, anatomically plausible hypothesis, they may diverge from the patient's actual clinical outcome. To bridge this gap, we introduce a rating-selection mechanism to filter these candidates and identify the most clinically probable progression path.

Training the Trajectory Rating Model. To evaluate the global consistency and anatomical plausibility of a predicted trajectory, we design a sequence scoring model f_ϕ. As illustrated in Fig. 1, this rating model is trained independently of the diffusion model to rank the quality of generated candidate trajectories. Specifically, f_ϕ consists of three primary components: a ResNet-based (He et al., 2016) encoder to extract spatial embeddings for each frame, a temporal transformer to capture temporal coherence (employing learnable embeddings to differentiate given context from generated frames), and an MLP regression head that outputs a plausibility score $s_j = f_\phi(\mathcal{X}^{(j)})$.

Rather than regressing an absolute quality score, we employ a pairwise ranking to measure relative quality (Liu et al., 2017). During training, we sample pairs of generated candidate trajectories $(\mathcal{X}^{(p)}, \mathcal{X}^{(q)}) \in \mathcal{X}_{\text{cand}}$. Let $e_p = \mathcal{D}(\mathcal{X}^{(p)}, \mathcal{X}^{gt})$ and $e_q = \mathcal{D}(\mathcal{X}^{(q)}, \mathcal{X}^{gt})$ denote their respective discrepancies from the ground truth sequence $\mathcal{X}^{gt}$, where $\mathcal{D}(\cdot, \cdot)$ represents a distance metric. We optimize f_ϕ using a margin ranking loss:

$$\mathcal{L}_{\text{rank}}(p, q) = \max(0, -\text{sgn}(e_q - e_p) \cdot (s_p - s_q) + \delta),$$

where δ is a predefined margin. This objective penalizes the model if a trajectory closer to the ground truth does not receive a sufficiently higher score, thereby training the network to reliably identify the most clinically probable progression.

Inference and Selection. During the inference stage, the diffusion model generates $\mathcal{X}_{\text{cand}}$ under varying delay configurations. The trained rating model f_ϕ then acts as a evaluator, scoring all candidates for a given patient. The final prediction is selected as the trajectory with the highest confidence, that is $\mathcal{X}^* = \arg\max_{\mathcal{X}^{(j)} \in \mathcal{X}_{\text{cand}}} f_\phi(\mathcal{X}^{(j)})$. Thus, the selected trajectory ensures both temporal consistency and clinical reliability.

Table 1: Quantitative comparison on the OAI dataset. The models utilize the first 3 frames (baseline, 12m, 24m) as context to predict the subsequent 4 time-points.

Month	**Metric**	**SLE** Han et al. (2022)	**SADM** Yoon et al. (2023)	**SEINE** Chen et al. (2024b)	**ImageFlowNet** Liu et al. (2025a)	**KneeForce** (Ours)
36m	PSNR ↑	22.29	26.12	28.03	<u>28.84</u>	**29.46**
	LPIPS ↓	<u>0.325</u>	0.419	0.330	0.358	**0.307**
	SSIM ↑	0.525	0.676	0.803	<u>0.805</u>	**0.812**
	KL acc (%) ↑	49.68	50.70	71.46	<u>71.99</u>	**72.45**
48m	PSNR ↑	22.25	25.77	27.76	<u>28.36</u>	**29.37**
	LPIPS ↓	0.334	0.422	<u>0.333</u>	0.372	**0.316**
	SSIM ↑	0.517	0.667	0.791	<u>0.803</u>	**0.811**
	KL acc (%) ↑	49.04	48.41	65.61	<u>68.54</u>	**72.87**
72m	PSNR ↑	21.55	25.09	26.82	<u>27.98</u>	**28.57**
	LPIPS ↓	0.346	0.429	<u>0.345</u>	0.389	**0.330**
	SSIM ↑	0.493	0.641	<u>0.763</u>	0.757	**0.783**
	KL acc (%) ↑	48.66	48.66	62.42	<u>63.82</u>	**68.92**
96m	PSNR ↑	21.37	24.37	26.45	<u>27.12</u>	**28.12**
	LPIPS ↓	<u>0.353</u>	0.436	0.369	0.400	**0.340**
	SSIM ↑	0.486	0.622	0.759	<u>0.761</u>	**0.781**
	KL acc (%) ↑	47.39	46.75	55.41	<u>59.87</u>	**62.13**

4 Experiments and Results

4.1 Datasets and Implementation

We evaluated KneeForce on the public Osteoarthritis Initiative (OAI) dataset (Nevitt et al., 2006). Knee X-rays were acquired at baseline and subsequent 12, 24, 36, 48, 72, and 96-month follow-ups for each patient. By leveraging KNEEL (Tiulpin et al., 2019b) to extract anatomical landmarks, we applied cropping with a 120 mm side length for each knee, resulting in 3,925 unilateral knees with complete 7-time-point trajectories. The knee samples were randomly split into a training set (3,140) and a test set (785) at the patient level, ensuring no patient overlap between the two sets. Moreover, there is no overlap between the test set and the data used to pre-train the OrthoFoundation expert (Yu et al., 2026), ensuring a strictly fair evaluation.

Furthermore, to assess generalization, we collected an in-house private dataset for external validation, which includes 57 cases with a total of 57 baseline and 116 follow-up X-rays acquired at irregular sampling intervals.

The diffusion process was configured with 1,000 timesteps, utilizing DDIM (Song et al., 2021) sampling with 100 steps. We trained the model from scratch for 100 epochs with a batch size of 3, learning rate of 8.0×10^{-5} and weight decay of 1.0×10^{-4}. The distance metric $\mathcal{D}$ is implemented as the pointwise ℓ_2 distance across all frames in the trajectory. The weight λ for the TRA loss was set to 10. The parameter d for Trajectory Selection is varying from 10 to 60 with a step size of 10.

4.2 Main Results

We compared KneeForce with SOTA methods. These include the KOA longitudinal prediction method SLE (Han et al., 2022), medical imaging approaches such as SADM (Yoon et al., 2023), ImageFlowNet (Liu et al., 2025a) and Δ-LFM (Chen et al., 2026), and the general video generation method SEINE (Chen et al., 2024b). Since Δ-LFM is formulated as a single-image longitudinal prediction method, we include it only in the context=1 setting to ensure a fair comparison. The context=1 setting also provides a more challenging evaluation of long-term forecasting from minimal historical information, while allowing direct comparison with single-image prediction baselines. Together with the context=3 results, these experiments evaluate KneeForce under both limited- and richer-context scenarios.

4.2.1 OAI Dataset

Table 1 presents the results when the model is provided with the first three time-points (baseline, 12m, and 24m) and tasked with predicting the subsequent four time-points. We evaluated image quality (PSNR, SSIM (Wang et al., 2004)) and perceptual quality (LPIPS (Zhang et al., 2018)) by comparing the generated X-rays with the ground truth. Additionally, we assessed KL accuracy using a ResNet-based (He et al., 2016) KL classifier to label both generated and ground truth images. The results demonstrate that KneeForce surpasses competing approaches in both image quality and clinical accuracy across long-term predictions.

To further investigate the model's capacity to capture long-term progression with limited information, we reduced the input context to only the baseline image, requiring the model to predict six follow-up time-points. As shown in Fig. 2, KneeForce consistently outperforms other methods including single-point method (Rombach et al., 2022). Notably, the performance naturally declines when predicting more distant time-points, indicating the challenge of long-term progression forecasting.

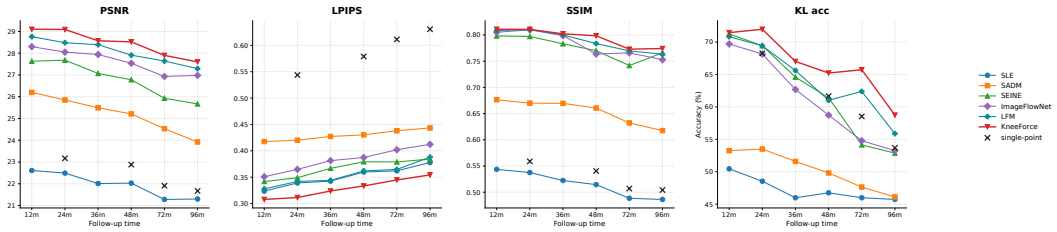

Figure 2: Long-term prediction performance on OAI given only the baseline image. "$\times$" markers denote single-point model (Rombach et al., 2022) trained specifically for each fixed time interval (e.g., baseline $\rightarrow$ 48m), serving as a reference for non-sequential prediction performance.

4.2.2 External Validation on In-house Dataset

To assess the generalization ability of the compared methods, we conducted external validation on the in-house dataset. All models trained on the OAI training set were directly tested on in-house data. Due to the incomplete follow-up visits, we used only the baseline X-rays as input to predict all subsequent time-points. Metrics were calculated on all available sample-ground-truth pairs, and the average results are reported.

As shown in Tab. 2, all methods exhibited a performance drop compared to OAI (Fig. 2) due to the distribution gap, including differences in racial demographics and misalignments in patient positioning across different time-points. Although KneeForce yields a lower PSNR than ImageFlowNet, this

is probably because ImageFlowNet tends to generate overly smoothed predictions that favor PSNR. Meanwhile, KneeForce maintains advantage in KL accuracy and LPIPS, which are less susceptible to misalignment and better reflect clinical validity.

Table 2: External validation on the in-house dataset. Given only the baseline image as input, average performance across all subsequent time-points are reported.

Metric	SADM	SEINE	ImageFlowNet	KneeForce
PSNR ↑	22.12	24.73	**25.67**	25.46
LPIPS ↓	0.497	0.404	0.415	**0.399**
SSIM ↑	0.539	0.685	0.706	**0.722**
KL acc (%) ↑	46.55	47.41	49.14	**52.59**

4.3 Ablation Study and Qualitative Analysis

To verify the effectiveness of our proposed Temporal Relational Alignment (TRA) and Trajectory Selection, we conducted an ablation study on OAI. As summarized in Tab. 3, we observed a progressive improvement in performance when incrementally adding each component to the Diffusion Forcing baseline. We note that explicitly encoding the time interval yields only marginal gains on OAI, as its follow-up visits are acquired at fixed, aligned time-points (36m, 48m, 72m, 96m) where temporal spacing carries little discriminative information; nevertheless, this conditioning is essential for real-world clinical settings, where follow-up visits are irregularly spaced. Building on this, TRA enhances the temporal consistency of features, while the Selection strategy further ensures high-fidelity generation by filtering out clinically unreliable trajectories. The full model consistently achieves the best results across all time-points, demonstrating the complementarity of both modules.

Visual comparisons (Fig. 3) highlight the practical advantages of KneeForce over comparison methods. While ImageFlowNet and SADM fail to preserve basic anatomical boundaries, our method produces realistic textures and accurate morphology. Moreover, KneeForce tracks the progression of joint space narrowing more precisely than the DF baseline, especially at longer horizons. This demonstrates the effectiveness of our proposed TRA and trajectory selection in maintaining long-term structural fidelity.

Table 3: Ablation Study. We report the average performance across all future time-points (36m, 48m, 72m, 96m) to validate the effectiveness of each component.

Method	TRA	Selection	PSNR ↑	SSIM ↑	LPIPS ↓	KL acc (%) ↑
Diffusion Forcing	✗	✗	28.29	0.774	0.334	66.39
+ time interval	✗	✗	28.33	0.772	0.330	66.45
+ TRA	✓	✗	28.50	0.786	0.329	67.83
+ Trajectory Selection (**KneeForce**)	✓	✓	**28.88**	**0.797**	**0.323**	**69.09**

4.4 Clinical Blind Evaluation

To validate the clinical relevance of our predictions beyond automated metrics, we conducted a blinded reader study with two sports medicine specialists. We randomly selected 50 cases from the OAI test set. For each case, the follow-up X-rays generated by four methods—KneeForce (Ours), SEINE (Chen et al., 2024b), SADM (Yoon et al., 2023), and ImageFlowNet (Liu et al., 2025a)—were shuffled and anonymized into four options (A, B, C, and D). Given the baseline and the ground-truth follow-up X-rays, each specialist was asked to rank the four options from 1 to 4 (1 being the best, 4 the worst) according to how faithfully the predicted image reproduced the clinical KOA progression features of the ground truth(e.g., joint space narrowing and osteophyte formation). The rankings from both readers across all 50 cases were then aggregated.

As summarized in Tab. 4, KneeForce achieved the best mean rank of 1.21 ± 0.43, substantially outperforming SEINE (1.84), SADM (3.25), and ImageFlowNet (3.70). Notably, KneeForce was ranked first in 80.0% of the cases, indicating a strong and consistent clinical preference for our

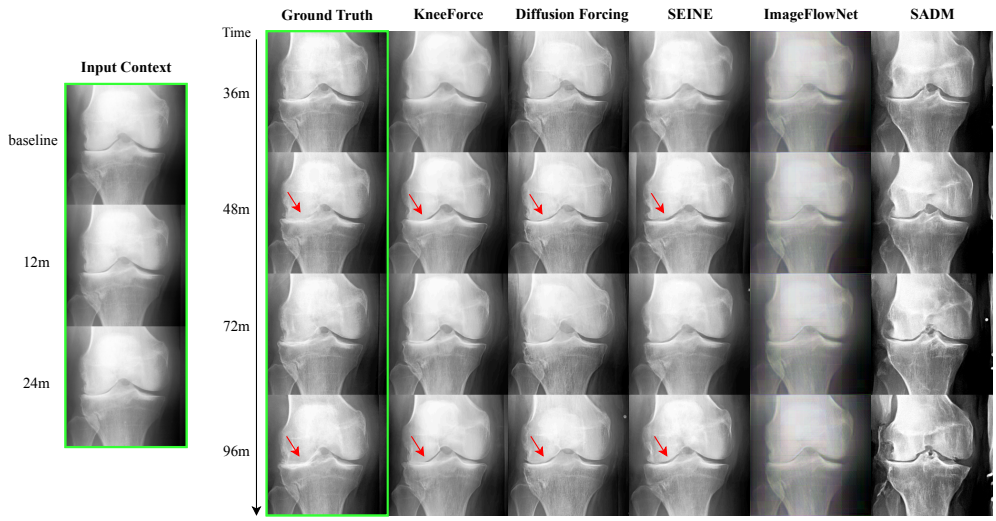

Figure 3: Visualization of the generated X-ray sequence. Red arrows indicate that KneeForce yields a joint space narrowing trend more consistent with Ground Truth.

Table 4: Clinical blind evaluation on 50 cases (context = 5), where each model predicts the subsequent 2 follow-ups. For each case, we first average the ranks assigned by the two readers to obtain a per-case score; the reported values are then the mean and standard deviation of these per-case scores over the 50 cases (lower is better).

Method	**Mean Rank** $\downarrow$
SADM (Yoon et al., 2023)	3.25 ± 0.54
ImageFlowNet (Liu et al., 2025a)	3.70 ± 0.46
SEINE (Chen et al., 2024b)	1.84 ± 0.47
KneeForce (Ours)	$\mathbf{1.21 \pm 0.43}$

predictions. These results demonstrate that KneeForce not only excels on quantitative metrics but also aligns closely with expert clinical judgment, underscoring its potential for real-world clinical application.

5 Conclusion

This paper proposes KneeForce, a novel framework for longitudinal KOA radiograph prediction. By integrating Diffusion Forcing with Temporal Relational Alignment and Trajectory Selection, our method accurately models KOA disease trajectories. Unlike prior approaches that regress abstract scalar scores or predict isolated time-points, KneeForce jointly models the continuous progression trajectory under irregular follow-up intervals, producing spatially explicit and clinically interpretable radiographs. Extensive evaluations on public and in-house datasets demonstrate SOTA performance, offering a visual tool for long-term clinical prognosis. A blinded reader study by experienced clinicians further confirms that our predictions align closely with expert judgment, underscoring the framework's readiness for real-world deployment.

AI Use Statement

Generative AI tools were used to assist with literature search, identification and summarization of relevant prior work, and manuscript editing for readability and organization. We did not use generative AI to develop the research methodology, design experiments or interpret the experimental

results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript, references, and empirical claims.

REFERENCES

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, et al. MedGAN: Medical image translation using GANs. *Computerized medical imaging and graphics*, 79:101684, 2020.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:24081–24125, 2024a.

Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. In *International Conference on Learning Representations*, 2026.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, et al. SEINE: Short-to-long video diffusion model for generative transition and prediction. In *International Conference on Learning Representations (ICLR)*, 2024b.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11472–11481, 2022.

Aiyong Cui, Huizi Li, Dawei Wang, Junlong Zhong, Yufeng Chen, and Huading Lu. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. *EClinical Medicine*, 29, 2020.

Jorge I Galvan-Tejada, Víctor Treviño, José M Celaya-Padilla, and Jose G Tamez-Pena. Knee osteoarthritis pain prediction from x-ray imaging: Data from osteoarthritis initiative. In *2014 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pp. 194–199. IEEE, 2014.

Tianyu Han, Jakob Nikolas Kather, Federico Pedersoli, Markus Zimmermann, Sebastian Keil, et al. Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. *Nature Machine Intelligence*, 4(11):1029–1039, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, et al. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023.

Mark D Kohn, Adam A Sassoon, and Navin D Fernando. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research®*, 474: 1886–1893, 2016.

Kevin Leung, Bofei Zhang, Jimin Tan, Yiqiu Shen, Krzysztof J Geras, et al. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. *Radiology*, 296(3):584–593, 2020.

Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, and Sebastiano Battiato. Tadm: Temporally-aware diffusion model for neurodegenerative progression on brain mri. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

Chen Liu, Ke Xu, Liangbo L Shen, Guillaume Huguet, Zilong Wang, et al. Imageflownet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025b.

Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1040–1049, 2017.

Armaghan Mahmoudian, L Stefan Lohmander, Ali Mobasheri, Martin Englund, and Frank P Luyten. Early-stage symptomatic osteoarthritis of the knee—time for action. *Nature Reviews Rheumatology*, 17(10):621–632, 2021.

Michael Nevitt, David Felson, and Gayle Lester. The osteoarthritis initiative. *Protocol for the cohort study*, 1:2, 2006.

Fabi Prezja, Leevi Annala, Sampsa Kiiskinen, Suvi Lahtinen, Timo Ojala, and Paavo Nieminen. Generating synthetic past and future states of knee osteoarthritis radiographs using cycle-consistent generative adversarial neural networks. *Computers in Biology and Medicine*, 187: 109785, 2025.

Lemuel Puglisi, Daniel C Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

Daniel L Riddle, Paul W Stratford, and Robert A Perera. The incident tibiofemoral osteoarthritis with rapid progression phenotype: development and validation of a prognostic prediction rule. *Osteoarthritis and cartilage*, 24(12):2100–2107, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Nripendra Kumar Singh and Khalid Raza. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare*, pp. 77–96, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Aleksei Tiulpin, Stefan Klein, Sita MA Bierma-Zeinstra, Jérôme Thevenot, Esa Rahtu, Joyce van Meurs, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports*, 9(1):20038, 2019a.

Aleksei Tiulpin, Iaroslav Melekhov, and Simo Saarakkala. KNEEL: Knee anatomical landmark localization using hourglass networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

Wenhua Wu, Kun Hu, Wenxi Yue, Wei Li, Milena Simic, Changyang Li, Wei Xiang, and Zhiyong Wang. Identity-consistent diffusion network for grading knee osteoarthritis progression in radiographic imaging. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2024.

Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. SADM: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, pp. 388–400. Springer, 2023.

Kang Yu, Dingyu Wang, Zimu Yuan, Nan Zhou, Jiajun Liu, et al. Orthofoundation: A large-scale multimodal vision foundation model for generalizable musculoskeletal diagnosis and prognosis, 2026. `https://github.com/ytrsk/OrthoFoundation-MSK-5M`.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, et al. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

A SUPPLEMENTARY MATERIAL

A.1 EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

All experiments were conducted on 2 NVIDIA A100 (40 GB) GPUs. The diffusion model was optimized with the AdamW optimizer, and all images were processed at a resolution of 256×256 across every experiment.

The trajectory rating model f_ϕ adopts an ImageNet-pretrained ResNet-50 as its frame-level feature extractor, projecting each frame to a 1024-dimensional embedding. Learnable positional encodings and two learnable mask embeddings—distinguishing context frames from generated frames—are added before an 8-layer, 8-head Transformer encoder aggregates the per-frame features into a sequence-level representation via mean pooling. A two-layer MLP with a hidden dimension of 512 then regresses the final plausibility score. The model is trained with a margin ranking loss (margin $\delta = 0.1$) using the Adam optimizer with a learning rate of 1.0×10^{-5} under a linear decay schedule, for 400 epochs with a batch size of 24.

Unless otherwise specified, all sampling results reported in the main paper are averaged over 3 independent runs to account for the stochasticity of the generative process.

The complete set of hyper-parameters is summarized in Tab. 5. For detailed implementation please refer to the source code.

Table 5: Hyper-parameter settings of KneeForce. Values already reported in the main text are included for completeness.

Component / Hyper-parameter	**Value**
Diffusion model	
Optimizer	AdamW
Learning rate	8.0×10^{-5}
Weight decay	1.0×10^{-4}
Batch size	3
Training epochs	100
Image resolution	256×256
Diffusion timesteps	1,000
Sampler / steps	DDIM / 100
TRA loss weight λ	10
Delay d range (step)	10–60 (10)
Trajectory rating model	
Backbone	ResNet-50
Frame feature dim	1,024
Transformer layers / heads	8 / 8
MLP hidden dim	512
Ranking margin δ	0.1
Optimizer	Adam
Learning rate	1.0×10^{-5} (linear decay)
Batch size	24
Training epochs	400

A.2 SUBJECTIVE EVALUATION INTERFACE

To ensure an unbiased blinded reader study, we designed a standardized evaluation interface, illustrated in Fig. 5. For each case, the ground-truth longitudinal sequence is displayed in the top row, with the target follow-up time-points to be assessed highlighted by a red box. Below it, the follow-up X-rays predicted by the four compared methods—KneeForce, SEINE, SADM, and ImageFlowNet—are randomly shuffled and anonymized as options A, B, C, and D, so that neither the method identity nor a fixed ordering can bias the readers. Each specialist ranks the four options from 1 (best) to 4 (worst) according to how faithfully the predicted images reproduce the clini-

cal KOA progression features consistent with the ground truth. The option-to-method mapping is randomized independently per case and kept hidden from the readers throughout.

A.3 Limitations and Failure Cases

Despite its strong performance, KneeForce has two main limitations.

Sensitivity to the training data distribution. As a generative model, KneeForce is inherently influenced by the distribution of its training data. Since mild cases dominate the training corpus, the model tends to regress toward the more frequently observed mild progression patterns. Consequently, for severe-progression cases, whose rate and degree of degeneration deviate markedly from the training distribution, the model is prone to under-predicting the progression, yielding follow-up X-rays that appear milder than the ground truth and thereby degrading prediction accuracy. A representative failure case is shown in Fig. 4.

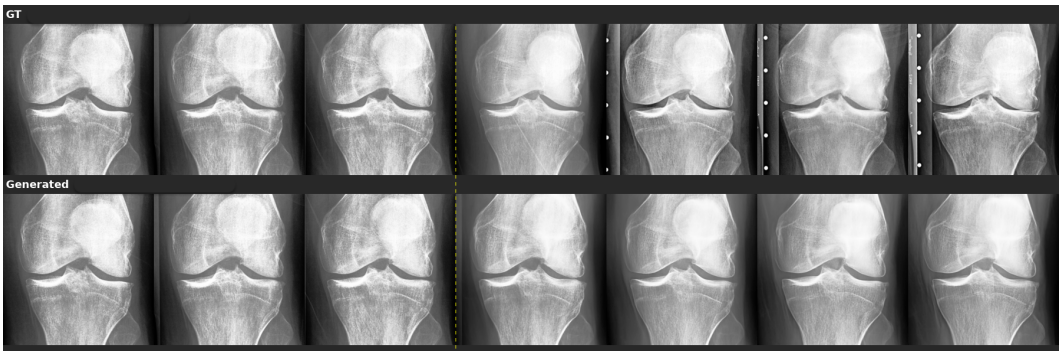

Figure 4: A representative failure case of a severe-progression subject. The model reverts toward the dominant mild progression patterns and under-predicts the disease progression, producing a follow-up trajectory that appears milder than the ground truth.

Automatic-only supervision of trajectory selection. The trajectory selection module is trained to rank candidates using an metric against the ground-truth sequence. As a result, its selection criterion is fully driven by image quality metric, and it cannot yet trade off between pure image fidelity and human clinical preference. Bridging this gap—for instance, by incorporating expert feedback into the ranking objective—is left for future work.

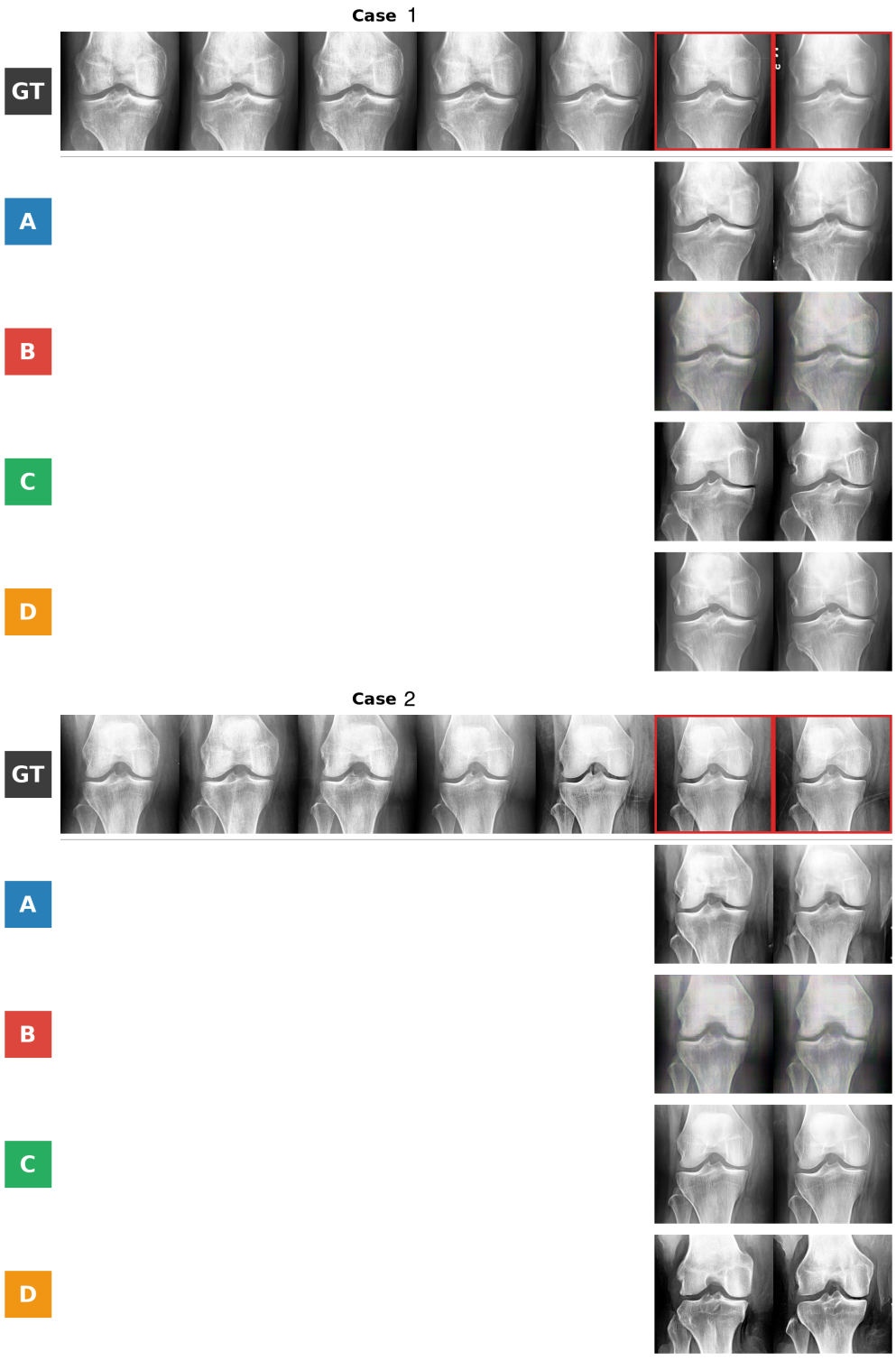

Figure 5: Two example cases from the subjective evaluation interface. The top (GT) row shows the ground-truth longitudinal sequence, with the follow-up time-points to be evaluated marked by the red box. The predictions of the four methods are shuffled and anonymized as options A–D, presented to the readers without any method identity or ordering cue.