TIDAL: A Temporal Causal Diffusion Framework for Visualizing Knee Osteoarthritis Treatment Outcomes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Generating realistic patient-specific counterfactual images of treatment outcomes from longitudinal medical imaging is a challenging task, often complicated by confounding and selection bias in observational datasets. To address this challenge, we propose TIDAL (**T**emporal **I**PW **D**iffusion **A**dversarial **L**earning), a novel longitudinal causal diffusion framework that integrates causal inference techniques directly into diffusion model training. TIDAL utilizes a Stable Diffusion backbone conditioned on patient history and incorporates two key causal adaptations: (1) Temporal Inverse Propensity Weighting (IPW) that reweights the diffusion loss based on treatment propensity scores; and (2) Domain Adversarial Training that encourages treatment-invariant representations. We demonstrate TIDAL's effectiveness by simulating knee osteoarthritis (OA) progression with longitudinal X-Rays from the Osteoarthritis Initiative (OAI). Performance is assessed using image fidelity metrics and causally-relevant Individual Treatment Effect (ITE) metrics for OA features like Kellgren-Lawrence grade. Our experiments show that TIDAL outperforms baseline approaches, achieving 21.52% reduction in image generation error and 18.43% improvement in causal validity, demonstrating significant improvements for longitudinal medical counterfactual generation.

1 Introduction

Visualizing patient-specific future health outcomes under hypothetical interventions holds transformative potential for personalized medicine [7, 20]. However, generating faithful counterfactuals from observational medical data faces significant challenges due to confounding bias: factors influencing both treatment assignment and outcomes can lead to spurious correlations and misleading predictions [5, 23]. We focus on knee osteoarthritis (OA), a chronic joint disease affecting 10-37% of people over 60 [27, 3]. Using the Osteoarthritis Initiative (OAI) dataset [17], we propose TIDAL (Temporal IPW Diffusion Adversarial Learning), a framework integrating causal inference techniques into diffusion model training. TIDAL is the first longitudinal causal diffusion framework for patient-specific treatment outcome visualization applied to the OAI dataset.

Contributions: (1) We proposed TIDAL Framework, combining temporal propensity weighting with adversarial training; (2) We incorporated TIDAL with diffusion generative model by reweighting diffusion loss based on treatment propensity scores and introducing Domain Adversarial Training that encourages treatment-invariant representations; (3) we conducted Comprehensive evaluation, demonstrating 21.52% reduction in image generation error and 18.43% improvement in causal validity over baseline approaches.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

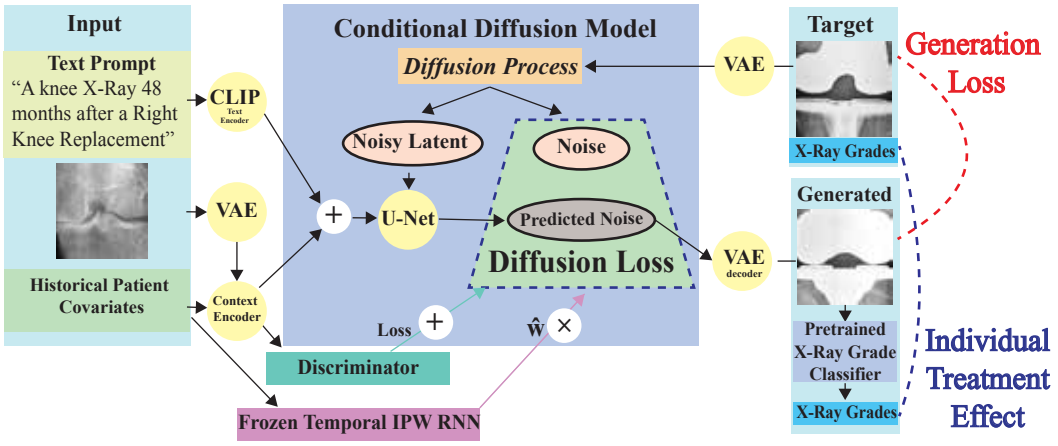

Figure 1: Overview of the TIDAL framework. Inputs (baseline X-Ray, patient covariates, text prompt) condition a U-Net for diffusion-based counterfactual generation. TIDAL applies causal adjustment via Temporal IPW RNN weighting and adversarial training with a treatment discriminator. Both components use historical patient covariates to mitigate confounding bias.

2 Related Work

Counterfactual Outcome Prediction: Traditional causal inference methods focusing on tabular data use techniques like Inverse Propensity Weighting (IPW) [22, 5] or representation learning [8]. Recent deep learning adaptations include sequence models [1, 16] and domain adversarial training [16]. While some works have explored counterfactual image generation using diffusion models [26, 11, 32, 30], integrating temporal causal inference into longitudinal medical imaging remains challenging. Some recent works [21, 19, 31] have advanced counterfactual medical image synthesis. However, existing approaches typically rely on conditional diffusion models [32, 30] without incorporating causality, leading to confounding bias. Our model combines diffusion generation with causal inference to directly address this challenge.

Diffusion Models for Causal Inference: We choose diffusion models for their superior training stability and sample quality [6, 24]. Previous causal diffusion works either insufficiently account for causality [32, 30] or train from scratch [11, 26]. Instead, we fine-tune pre-trained models with causal inference-motivated losses.

3 TIDAL: Temporal IPW Diffusion Adversarial Learning

TIDAL generates patient-specific future medical images X_{t_l} conditioning on baseline images X_{t_e} and treatments $\mathbf{A}_{\text{int}}$ during interval $(t_e, t_l]$, while mitigating confounding bias. The framework leverages patient history $H_{t_e}^{\text{long}}$ with two causal adaptations: Temporal IPW and Domain Adversarial Training.

3.1 Temporal Conditional Diffusion Model

Diffusion models [6, 28] progressively add noise to data in a forward process, then learn to reverse this process to generate samples close to the distribution of data. We build on Stable Diffusion [24] with: (1) Frozen VAE for latent encoding/decoding, (2) Trainable U-Net for denoising, (3) Frozen CLIP text encoder, (4) DDIM scheduler.

U-Net Conditioning: The U-Net is conditioned on text prompts "A knee X-Ray Δt months after {treatment list}" and spatio-temporal context c_{ctx} from a Context Encoder RNN that processes baseline image features, patient longitudinal history $H_{t_e}^{\text{long}}$, follow-up duration Δt, and knee side S. The final conditioning vector is $c_{\text{U-net}} = c_{\text{text}} + c_{\text{ctx}}$. The diffusion loss is:

$$\mathcal{L}_{\text{diffusion}}(\theta) = \mathbb{E}_{z_{t_l}, \epsilon, t, c_{\text{U-net}}} \|\epsilon - \epsilon_\theta(z_t, t, c_{\text{U-net}})\|_2^2 \tag{1}$$

where z_{t_l} is the target latent, $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, t is the diffusion timestep, and z_t is the noisy latent.

3.2 Inverse Propensity Weighted Diffusion Model

Modeling only the conditional distribution of outcome given treatments in the observational OAI dataset can result in selection bias and confounding bias due to non-random treatment assignment [25]. Inverse Propensity Weighting (IPW) addresses confounding bias by reweighting samples to create a pseudo-randomized population. Unlike DiffPO [15] which uses static covariates, our temporal IPW employs LSTM-based sequence modeling for evolving treatment propensities based on patient history $H_{t_e}^{\text{long}}$. We develop a propensity score model g_{ϕ_p} to estimate the probability of receiving treatments $\mathbf{A}_{\text{int}}$ during interval $(t_e, t_l]$:

$$\hat{\pi}_k = g_{\phi_p,k}(H_{t_e}^{\text{long}}, \Delta t, S) \approx P(A_{\text{int},k} = 1 \mid H_{t_e}^{\text{long}}, \Delta t, S) \tag{2}$$

The IPW weight for sample i is the inverse of the joint treatment probability:

$$w_i = \frac{1}{\hat{P}(\mathbf{A}_{\text{int}} = \mathbf{a}_{i,\text{int}} \mid H_{i,t_e}^{\text{long}}, \Delta t_i, S_i)} \tag{3}$$

The IPW-adjusted diffusion loss reweights samples:

$$\mathcal{L}_{\text{IPW-Diffusion}} = \sum_{i=1}^{N} w_i \cdot \ell_{\text{diffusion},i} \tag{4}$$

where $\ell_{\text{diffusion},i}$ is the per-sample diffusion loss.

3.3 Domain Adversarial Training

IPW training can introduce unstable training due to exploding weights when treatment probabilities approach zero. Domain adversarial training provides an alternative approach by encouraging treatment-invariant representations [14, 29].

We train a treatment discriminator $\mathbf{D}$ (MLP parameterized by ϕ) to predict interval treatments $\mathbf{A}_{\text{int}}$ from generator context c_{ctx}, while generator $\mathbf{G}$ aims to fool $\mathbf{D}$. The discriminator predicts over $K = 13$ classes (12 specific treatments plus "no treatment"). The discriminator loss uses Binary Cross-Entropy:

$$\mathcal{L}_D(\phi) = \mathbb{E}_{c_{\text{ctx}}, \mathbf{A}_{\text{int}}} \left[\sum_{k=1}^{13} \text{BCE}(D(c_{\text{ctx}}; \phi)_k, A_{\text{int},k}) \right] \tag{5}$$

The generator loss combines diffusion and adversarial terms:

$$\mathcal{L}_G(\theta) = \mathcal{L}_{\text{diffusion}}(\theta) + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}(\theta) \tag{6}$$

where $\mathcal{L}_{\text{adv}}(\theta)$ encourages uniform discriminator output distributions (maximize prediction entropy; see Appendix G.3). Parameters are updated iteratively: fix θ, update ϕ; then fix ϕ, update θ.

3.4 TIDAL: Combined IPW and Adversarial Training

TIDAL combines both mechanisms with the overall loss:

$$\mathcal{L}_{\text{TIDAL}}(\theta, \phi) = \mathcal{L}_{\text{IPW-Diffusion}}(\theta) + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}(\theta) + \mathcal{L}_D(\phi) \tag{7}$$

Training alternates between propensity model pre-training and joint adversarial updates. We alternate $\min_\phi \mathcal{L}_D$ and $\min_\theta(\mathcal{L}_{\text{IPW-Diffusion}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}})$; the propensity model is pre-trained and frozen during these updates.

4 Experiments

We use the Osteoarthritis Initiative (OAI) dataset [17], creating longitudinal image pairs (X_{t_e}, X_{t_l}) from chronologically ordered X-Ray scans spanning up to 144 months. Our preprocessing pipeline extracts patient longitudinal histories $H_{t_e}^{\text{long}}$ including clinical assessments, prior treatments, and demographic covariates. Data is split at patient level: 80% training (51,726 pairs), 10% validation (6,684 pairs), 10% test (6,331 pairs). See Appendix F for detailed statistics.

Table 1: Test Performance of Treatment Outcome Modeling with different causal inference adaptations, including percentage decrease in error compared to Baseline.

Model	Predicted Noise MSE	Generated Image MSE	SSIM
	Value (% vs Base)	Value (% vs Base)	Value (% vs Base)
Baseline	0.1361 -	0.0079 -	0.77 -
+ IPW Training	0.1359 (0.15% ↓)	0.0075 (5.06% ↓)	0.80 (3.90% ↑)
+ Adversarial Training	0.1301 (4.41% ↓)	0.0067 (15.19% ↓)	0.81 (5.19% ↑)
TIDAL	**0.1294** (**4.92%** ↓)	**0.0062** (**21.52%** ↓)	**0.83** (**7.79%** ↑)

Table 2: Causal Performance of Treatment Outcome Modeling with different causal inference adaptations, including percentage decrease in ITE error compared to Baseline.

Model	ITE - KL Grade		ITE - JSN Medial Grade	
	Value	(% Decrease vs Baseline)	Value	(% Decrease vs Baseline)
Baseline	0.8152	-	0.2996	-
+ IPW Training	0.7785	(4.50% ↓)	0.2754	(8.08% ↓)
+ Adversarial Training	0.7712	(5.40% ↓)	0.2511	(16.19% ↓)
TIDAL	**0.7689**	(**5.68%** ↓)	**0.2444**	(**18.43%** ↓)

Implementation Details: We use Stable Diffusion v1.5 with frozen VAE and CLIP encoders. The Context Encoder uses 2-layer LSTMs with 128 hidden dimensions. Propensity models employ 128-dimensional bidirectional LSTMs. The treatment discriminator uses a 3-layer MLP with ReLU activations. Training uses AdamW optimizer with learning rate 1e-5, batch size 64 on 2×L40S GPUs, and $\lambda_{\text{adv}} = 0.05$. We assess image fidelity using predicted noise MSE, generated image MSE, and SSIM. For causal validity, we measure Individual Treatment Effect (ITE) error on clinically relevant OA features: Kellgren-Lawrence grade and JSN Medial grade [10]. ITE compares how much an X-Ray grade changes between the generated image and the baseline, versus how much it changes between the target image and the baseline. It measures the absolute error between these two predicted X-Ray grade deltas. Calibration details appear in Appendix C. ITE error quantifies how well the model captures true treatment effects.

4.1 Results

TIDAL achieves 21.52% reduction in image generation error and 18.43% improvement in causal validity (Tables 1, 2). Both IPW and adversarial training contribute to these improvements, with their combination providing the strongest performance across all metrics. Adversarial training shows stronger improvements than IPW across all image quality metrics. The 15.19% MSE reduction with adversarial training suggests better structural preservation in generated images. SSIM improvements (7.79% for TIDAL) indicate better perceptual quality maintenance. Both approaches significantly reduce ITE error, with adversarial training excelling particularly for JSN Medial grade (16.19% improvement). This suggests the domain adversarial component effectively learns treatment-invariant representations crucial for causal inference. The combined TIDAL approach achieves the best performance, demonstrating complementary benefits of both causal mechanisms. The improvements in Kellgren-Lawrence grade prediction (5.68% ITE reduction) are clinically meaningful as this grade is the primary diagnostic metric for OA severity. Better causal validity ensures more reliable counterfactual predictions for clinical decision support.

5 Conclusion

We presented TIDAL, a longitudinal causal diffusion framework that generates patient-specific counterfactual medical images while addressing confounding bias. TIDAL achieves significant improvements in both image fidelity and causal validity through temporal IPW and adversarial training, establishing new state-of-the-art for longitudinal medical counterfactual generation. While our evaluation uses standard image metrics, future work should explore clinically relevant evaluation and prospective validation to assess real-world clinical utility.

References

[1] Jeroen Berrevoets, Alicia Curth, Ioana Bica, Eoin F. McKinney, and Mihaela van der Schaar. Disentangled counterfactual recurrent networks for treatment effect inference over time. *ArXiv*, abs/2112.03811, 2021.

[2] Espen Andreas Brembo, Heidi Kapstad, Tom Eide, Lukas Månsson, Sandra Van Dulmen, and Hilde Eide. Patient information and emotional needs across the hip osteoarthritis continuum: a qualitative study. *BMC health services research*, 16:1–15, 2016.

[3] Robert H Brophy and Yale A Fillingham. Aaos clinical practice guideline summary: management of osteoarthritis of the knee (nonarthroplasty). *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, 30(9):e721–e729, 2022.

[4] Pingjun Chen, Linlin Gao, Xiaoshuang Shi, Kyle Allen, and Lin Yang. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 75:84–92, 2019.

[5] Miguel A Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

[6] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[7] Xuelin Huang and Jing Ning. Analysis of multi-stage treatments for recurrent diseases. *Statistics in Medicine*, 31(24):2805–2821, 2012.

[8] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029. PMLR, 2016.

[9] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024.

[10] 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.

[11] Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. *ArXiv*, abs/2404.17735, 2024.

[12] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. EfficientFormer: Vision transformers at MobileNet speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[14] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8036–8046, 2022.

[15] Yuchen Ma, Valentyn Melnychuk, Jonas Schweisthal, and Stefan Feuerriegel. DiffPO: A causal diffusion model for learning distributions of potential outcomes. *ArXiv*, abs/2410.08924, 2024.

[16] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. *ArXiv*, abs/2204.07258, 2022.

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

[18] L Pacheco-Brousseau, M Charette, S Poitras, and D Stacey. Effectiveness of patient decision aids for total hip and knee arthroplasty decision-making: a systematic review. *Osteoarthritis and Cartilage*, 29(10):1399–1411, 2021.

[19] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Counterfactual contrastive learning: robust representations via causal image synthesis. *arXiv preprint arXiv:2403.09605*, 2024.

[20] Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, and Mihaela van der Schaar. SyncTwin: Treatment effect estimation with longitudinal outcomes. *Advances in Neural Information Processing Systems*, 34:3178–3190, 2021.

[21] Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *International Conference on Machine Learning*, 2023.

[22] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[23] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[24] 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)*, pages 10684–10695, June 2022.

[25] Donald B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322 – 331, 2005.

[26] Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In *CLEaR*, 2022.

[27] Leena Sharma. Osteoarthritis of the knee. *New England Journal of Medicine*, 384(1):51–59, 2021.

[28] Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020.

[29] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[30] Zhe Wang, Aladine Chetouani, Rachid Jennane, Yuhua Ru, Wasim Issa, and Mohamed Jarraya. Temporal evolution of knee osteoarthritis: A diffusion-based morphing model for x-ray medical image synthesis. *ArXiv*, abs/2408.00891, 2024.

[31] Tian Xia, Athanasios Chartsias, and Ben Glocker. Mitigating attribute amplification in counterfactual image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[32] Yousef Yeganeh, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Bjorn Ommer, Nassir Navab, Azade Farshad, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis. *ArXiv*, abs/2412.20651, 2024.

A Additional Experimental Setups and Results

A.1 Common Model Architecture and Training Setup

All TIDAL variants (Baseline, IPW-enhanced, Adversarially-trained, and combined) share a core generative architecture based on conditional latent diffusion, fine-tuned from Stable Diffusion v1-5 [24]. All models are implemented in PyTorch, utilizing the PyTorch Lightning framework for training and the Hugging Face Diffusers library for diffusion model components. Training is performed using AdamW optimizers with 16-bit Automatic Mixed Precision (AMP). Shared hyperparameters include a learning rate of 1e-5 for the generator components (U-Net and conditioning MLPs) and a batch

size of 64 spread across 2 NVIDIA L40S GPUs. All model variants take up 45,000 MB on each of the two GPUs and take 1.5 days to finish 100 training epochs. The LSTMs used in the Temporal IPW model and Context Encoder had 2 layers with a hidden dimension of 128, they both also used a Dense layer of size 8 for the time delta and 4 for the knee side. Adversarial weight was set to 0.05. All experiments are seeded for reproducibility.

A.2 Qualitative Generated Image Evaluation

These images were generated by TIDAL with domain adversarial training. During inference, the Stable Diffusion backbone utilized a strength of 0.75, guidance scale of 7.5, and 50 inference steps.

Figure 2: Example X-Ray generated from TIDAL framework

Figure 3: Example X-Ray generated from TIDAL correctly predicting joint space narrowing on right side.

Figure 4: Example X-Ray generated from TIDAL incorrectly predicting joint space narrowing on right side.

A.3 Adversarial Weight Ablation

Table 3: Impact of Adversarial Weight (λ_{adv}) on Validation Loss. The reported Validation Loss is the lowest value achieved during training on one validation set for each corresponding adversarial weight.

Adversarial Weight (λ_{adv})	Validation Loss
1.0	0.1391
0.5	0.1337
0.1	0.1345
0.05	**0.1325**
0.01	0.1333

B Diffusion Pair Dataset Details

B.1 Image Processing and Knee Localization.

The OAI provides bilateral X-Ray images at various timepoints. To focus on individual knee data, we first process these bilateral scans. A YOLOv11-based object detection model [9], pre-trained on a dedicated knee X-Ray dataset for localization [30, 4], was employed to detect and crop the left and right knees from each bilateral image, see Figure 5. This step ensures that our models receive standardized single-knee views. All cropped images are resized to 224×224 pixels, converted to tensors scaling pixel values to $[0, 1]$, and then normalized to $[-1, 1]$ (mean 0.5, std 0.5) for input to the diffusion models.

B.2 Extracted Features.

- **Interval Treatment Information ($\mathbf{A_{int}}$):** For a pre-defined list of K treatments (e.g., specific injections, NSAID usage, arthroscopy, knee replacement; $K = 13$ in our setup covering left and right knee treatments such as Arthroscopy, Knee Replacement, Meniscectomy, Steroid Injection, Hip Replacement, and Hyaluronic Injection, plus "no treatment". This results in a multi-hot vector indicating treatments received during the interval.
- **Radiographic Grades ($H^{\mathbf{tab}}$):** Standardized radiological assessments, including Kellgren-Lawrence (KL) grade, and Joint Space Narrowing (JSN) for medial and lateral compartments, are extracted for both the left and right knees at both $month\_earlier$ and $month\_later$.
- **Clinical Information:** Time-varying clinical data such as Body Mass Index (BMI) and patient age are recorded.
- **Static Demographics:** Patient-level demographic information like sex, ethnicity, and race are included once per patient.
- **Longitudinal History ($H^{\mathbf{long}}_{t_e}$):** For models utilizing temporal context (IPW and the RNN-based adversarial discriminator), we construct sequences of historical covariates and treatments up to $month\_earlier$.
- **Knee Side (S)** and **Follow-up Duration ($\Delta t = month\_later - month\_earlier$)** are also recorded for each pair.

C Pretrained X-Ray Grade Model Details

To evaluate the causal validity of our generated counterfactual X-Ray images, particularly for assessing Individual Treatment Effects (ITE) on specific radiographic features, we pre-trained separate classifier models for key osteoarthritis (OA) indicators. We specifically trained models for (KL) Grade and JSN Medial Grade used in our main paper's ITE evaluations.

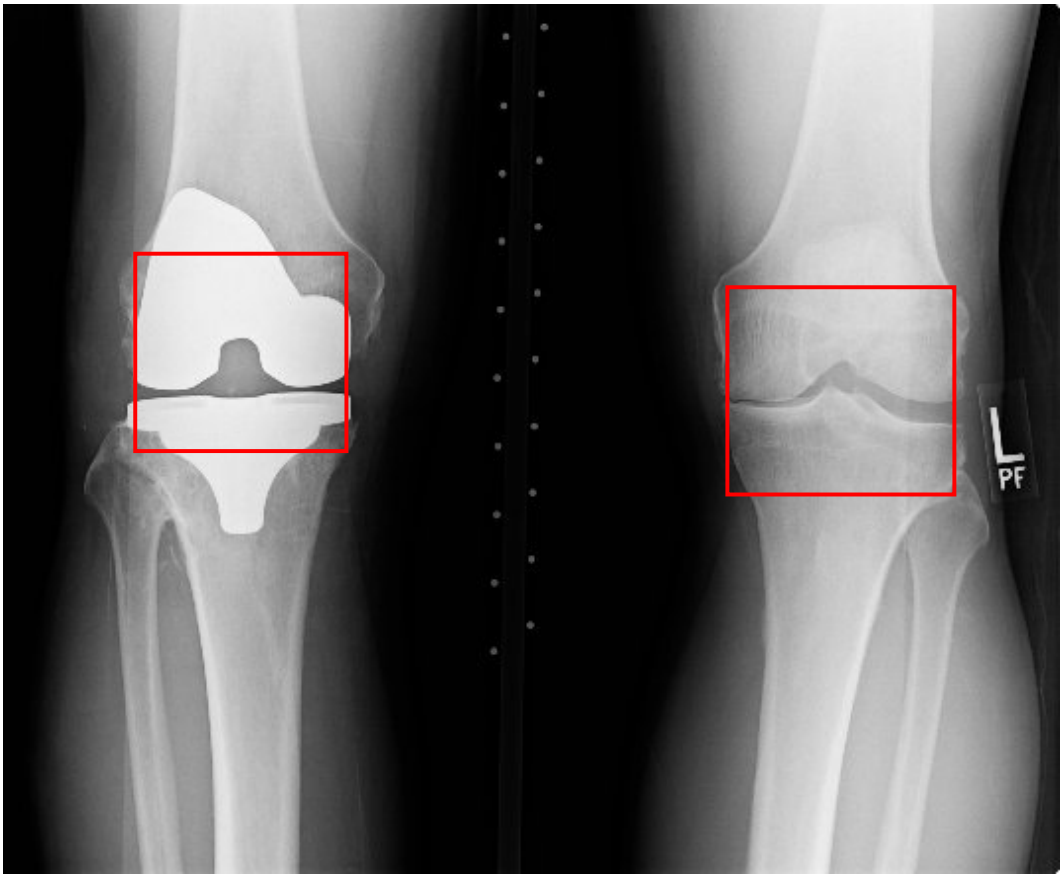

Figure 5: Image showing YOLO detecting bounding boxes for each knee from a Bilateral X-Ray from the OAI dataset.

C.1 Dataset and Preprocessing

The feature classifiers were trained using cropped single-knee X-Ray images derived from the Osteoarthritis Initiative (OAI) dataset, consistent with the images used for training our main diffusion models. The dataset splits used the same unique patient splits from the Diffusion Model dataset. The specific X-Ray grade (e.g., KL Grade ranging from 0-4, JSN Medial from 0-3) served as the target label for each respective model.

Input images were resized to 224×224 pixels. For training, we applied data augmentation techniques including random horizontal flips, random rotations (up to 10 degrees), color jitter (brightness, contrast, saturation by a factor of 0.2), and random affine transformations (translations up to 10%). All images (for training, validation, and testing) were then converted to tensors and normalized using ImageNet statistics (mean: $[0.485, 0.456, 0.406]$, std: $[0.229, 0.224, 0.225]$).

C.2 Model Architecture and Training

For each X-Ray feature, we fine-tuned a pre-trained EfficientFormerV2-L model [12]. The original classifier head of the model was replaced with a new linear layer randomly initialized to output C logits, where C is the number of classes for the specific radiographic feature (e.g., $C = 5$ for KL Grade 0-4, $C = 4$ for JSN Medial Grade 0-3).

The models were trained using a cross-entropy loss function. We employed the AdamW optimizer [13] with an initial learning rate of 1×10^{-5}. Training was conducted for 30 epochs, and the model state corresponding to the best validation macro-averaged AUC (Area Under the Receiver Operating Characteristic Curve) was saved. The batch size was set to 64.

C.3 Performance on Test Set

The performance of the pre-trained classifiers for KL Grade and JSN Medial Grade on the held-out test set is summarized in Table 4. These models are subsequently used in a frozen state to evaluate the ITE of the generated counterfactual images from our main diffusion pipelines.

Table 4: Test Set Performance of Pre-trained X-Ray Grade Classifiers.

Feature	Test Loss	Accuracy	Macro AUC	Num Classes
KL Grade	0.7918	0.6724	0.8867	5
JSN Medial Grade	1.4385	0.8160	0.9330	4

KL Grade Per-Class Test Accuracy: {0: 0.858, 1: 0.125, 2: 0.660, 3: 0.843, 4: 0.774}
KL Grade Class Prevalence (Test Set): {0: 0.392, 1: 0.175, 2: 0.262, 3: 0.131, 4: 0.039}

JSN Medial Grade Per-Class Test Accuracy: {0: 0.902, 1: 0.557, 2: 0.763, 3: 0.822}
JSN Medial Grade Class Prevalence (Test Set): {0: 0.669, 1: 0.204, 2: 0.096, 3: 0.029}

D Propensity Model Pretraining

We pretrained two distinct propensity models to predict treatment probabilities: a temporal model that incorporates sequential patient history and a non-temporal baseline model. Both models employ RNN architectures but differ significantly in their input representations and temporal modeling capabilities.

D.1 Temporal IPW vs. DiffPO Comparison

Our temporal IPW addresses fundamental limitations of DiffPO [15] in longitudinal medical settings:

Key Differences: (1) **Sequential vs. Static Modeling**: DiffPO uses time-agnostic propensity models with fixed covariates, while our temporal IPW employs LSTM-based sequence modeling to capture evolving treatment propensities based on longitudinal patient history $H_{t_e}^{\text{long}}$. (2) **Interval vs. Point Treatment Modeling**: DiffPO predicts single-point treatment assignments, whereas our model estimates probabilities for multi-treatment sets administered during specific time intervals $(t_e, t_l]$, reflecting real-world clinical practice. (3) **Temporal Context Integration**: Unlike DiffPO's static approach, our propensity model incorporates follow-up duration Δt and contextual factors (knee side S) that influence treatment timing decisions, enabling more accurate propensity estimation in longitudinal settings.

D.2 Model Architectures

Temporal Propensity Model: The temporal model processes sequential patient histories using an LSTM-based encoder (2 layers, 128 hidden dimensions). We compared LSTM against Transformer architectures, finding that LSTM achieved superior validation performance (AUC: 0.714 vs 0.682 for Transformer). The model takes as input:

- Sequential covariate vectors (medical history over time)
- Sequential treatment vectors (previous treatments)
- Temporal features including normalized time intervals (Δt) between observations
- Side information (left/right knee distinction)

The LSTM processes concatenated sequence features, followed by specialized MLPs for temporal (Δt) and side features. The final prediction head combines the sequence encoding with processed features to output treatment probabilities for $K = 13$ classes.

Detailed Architecture: The LSTM-based propensity model employs the following detailed architecture: The final hidden state from the LSTM h_{hist} summarizes the patient's entire history. Features for Δt and S are processed by separate small MLPs to yield $h_{\Delta t}$ and h_S. The concatenated representation $[h_{\text{hist}}; h_{\Delta t}; h_S]$ is passed through a final feed-forward network with sigmoid activation to output the K-dimensional probability vector $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_K)$.

Training Details: The propensity model g_{ϕ_p} is pre-trained separately by minimizing binary cross-entropy loss between predictions $\hat{\boldsymbol{\pi}}$ and true multi-hot interval treatment labels $\mathbf{A}_{\text{int}}$. To address class imbalance, the loss uses positive class weights derived from inverse treatment frequencies in the training data.

Non-temporal Propensity Model: The baseline model uses a simpler fusion approach, combining image features from an EfficientFormer backbone with tabular features (X-Ray grades, clinical information, and demographics). This model lacks temporal sequence processing and instead operates on static feature representations at individual time points.

D.3 Training Performance Comparison

The temporal model demonstrated superior performance across all key metrics:

Temporal Model Results:

- Final validation AUC: 0.714
- Final validation accuracy: 68.8%
- Macro recall (positives): 94.1%
- Training converged in 40 epochs with early stopping

Non-temporal Model Results:

- Final validation AUC: 0.706
- Final validation accuracy: 62.6%
- Macro recall (positives): 65.4%
- Training completed 50 full epochs

D.4 Key Findings

The temporal model's superior performance can be attributed to several factors:

1. **Sequential Information Utilization:** The temporal model leverages the full patient history sequence, capturing temporal dependencies and treatment progression patterns that the static model cannot access.

2. **Temporal Feature Engineering:** The explicit modeling of time intervals (Δt) between observations, with normalization (mean=35.17, std=22.99), allows the model to understand the temporal spacing of medical events.

3. **Enhanced Recall Performance:** The temporal model achieved significantly higher macro recall (positives) (94.1% vs 65.4%), indicating better identification of patients who actually received treatments.

4. **Class Imbalance Handling:** Both models employed positive weight rebalancing to address the severe class imbalance (90.7% "No Treatment" cases in temporal model), but the temporal model's sequential processing provided better discrimination.

The temporal model's architecture effectively captures the dynamic nature of treatment decisions in longitudinal healthcare data, demonstrating the importance of sequential modeling for propensity score estimation in medical applications.

E Longitudinal Data Pair Creation

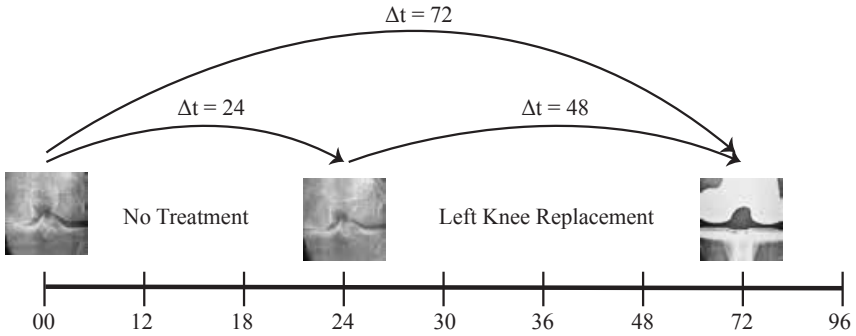

Figure 6: Illustration of longitudinal data pair creation from patient timelines. For each patient, all chronologically ordered pairs of X-Ray scans (e.g., month 00 to 24, 00 to 72, 24 to 72) are formed. The interval treatments (e.g., "No Treatment", "Left Knee Replacement") and the duration Δt between scans are recorded for each pair.

F Dataset Statistics

Table 5: Summary Statistics of Dataset Splits. Treatment occurrences show the number of image pairs where the treatment was recorded in the interval, with the percentage of total pairs for that split in parentheses. "No Treatment" is inferred for pairs where none of the specified treatments occurred.

Characteristic	**Train**	**Validation**	**Test**
Image Pairs	51,726	6,684	6,331
Unique Subjects	3,604	450	451
Treatment Occurrences (Count (%))			
L. Arthroscopy	933 (1.80)	132 (1.97)	153 (2.42)
R. Arthroscopy	963 (1.86)	131 (1.96)	95 (1.50)
L. Meniscectomy	702 (1.36)	130 (1.94)	138 (2.18)
R. Meniscectomy	775 (1.50)	76 (1.14)	85 (1.34)
L. Hyaluronic Inj.	815 (1.58)	105 (1.57)	111 (1.75)
R. Hyaluronic Inj.	793 (1.53)	85 (1.27)	114 (1.80)
L. Steroid Inj.	1902 (3.68)	253 (3.79)	247 (3.90)
R. Steroid Inj.	1805 (3.49)	182 (2.72)	302 (4.77)
L. Knee Replacement	679 (1.31)	108 (1.62)	93 (1.47)
R. Knee Replacement	704 (1.36)	51 (0.76)	108 (1.71)
L. Hip Replacement	411 (0.79)	34 (0.51)	26 (0.41)
R. Hip Replacement	417 (0.81)	93 (1.39)	48 (0.76)
No Treatment	45,312 (87.60)	5,845 (87.45)	5,384 (85.04)

G Detailed Method Descriptions

G.1 U-Net Conditioning Strategy Details

To generate a target latent z_{t_l} (corresponding to X_{t_l}), the U-Net is conditioned on a combination of textual information and a rich, spatio-temporal context vector:

1. *Textual Prompts (c_{text}):* Dynamically generated prompts of the form "A knee X-Ray Δt months after {treatment list}", where {treatment list} enumerates all treatments within the time interval. These are tokenized and encoded by the CLIP text encoder.
2. *Spatio-Temporal Rich Context (c_{ctx}):* This comprehensive conditioning vector is derived by a dedicated Context Encoder RNN (E_{ctx}). This encoder processes:
   - The baseline image condition c_{img}, which is the output of a linear layer (f_{img}) applied to the VAE-encoded latent representation of the baseline X-Ray X_{t_e}.
   - The patient's longitudinal history $H_{t_e}^{\text{long}}$, comprising sequences of historical covariates and treatments up to t_e. These sequences are processed by an LSTM within E_{ctx} to capture temporal dependencies, yielding h_{hist}.
   - The normalized follow-up duration Δt and the knee side S, each processed by separate small MLPs to get $h_{\Delta t}$ and h_S.

   E_{ctx} concatenates these features, $[h_{\text{hist}}; c_{\text{img}}; h_{\Delta t}; h_S]$, to form a single vector and is subsequently projected by a linear layer (f_{proj}) to match the dimensionality of the text embeddings, resulting in c_{ctx}.

The final conditioning vector $c_{\text{U-net}}$ fed to the U-Net's cross-attention layers is then the sum of the text embeddings and this projected rich context: $c_{\text{U-net}} = c_{\text{text}} + c_{\text{ctx}}$.

G.2 IPW Background and Extension Details

Standard IPW Background: Inverse Propensity Weighting (IPW) is a causal inference technique that addresses confounding bias by reweighting samples to create a pseudo-randomized population. The propensity score $\pi(x) = P(A = 1|X = x)$ represents the probability of receiving treatment

given covariates X. By weighting each sample by $1/\pi(x)$ for treated units and $1/(1-\pi(x))$ for control units, IPW balances the covariate distributions between treatment groups, simulating a randomized experiment.

Temporal IPW Extension: Unlike DiffPO which uses static covariates, our temporal IPW employs LSTM-based sequence modeling for evolving treatment propensities based on patient history $H_{t_e}^{\text{long}}$. Key differences include: (1) sequential vs. static modeling, (2) interval vs. point treatment prediction, and (3) temporal context integration.

We develop a propensity score model g_{ϕ_p} to estimate the probability of receiving treatments $\mathbf{A}_{\text{int}}$ during interval $(t_e, t_l]$, conditioned on patient history $H_{t_e}^{\text{long}}$, knee side S, and interval duration Δt. The model provides estimated marginal probabilities $\hat{\pi}_{i,k}$ for each treatment, and the joint probability is:

$$\hat{P}(\mathbf{A}_{\text{int}} = \mathbf{a}_{i,\text{int}} \mid H_{i,t_e}^{\text{long}}, \Delta t_i, S_i) = \prod_{k=1}^{K} \left(\hat{\pi}_{i,k}^{a_{i,\text{int},k}} \times (1 - \hat{\pi}_{i,k})^{(1-a_{i,\text{int},k})} \right) \tag{8}$$

G.3 Adversarial Training Details

The objective is to encourage the generator to learn representations of the baseline patient state (c_{ctx}) that are invariant to the actual treatment $\mathbf{A}_{\text{int}}$ received during the subsequent interval, conditioned on the patient's prior history $H_{t_e}^{\text{long}}$.

Diffusion Image Generator (G): The core conditional diffusion model (with trainable parameters θ) generates realistic future X-Ray images X_{t_l} and the rich context vector c_{ctx}.

Treatment Discriminator (D): An auxiliary MLP (parameterized by ϕ) predicts interval treatments $\mathbf{A}_{\text{int}}$ using the generator's context vector c_{ctx} as input.

The generator aims to fool $\mathbf{D}$ by making c_{ctx} uninformative about $\mathbf{A}_{\text{int}}$ via adversarial loss:

$$\mathcal{L}_{\text{adv}}(\theta) = -\mathbb{E}_{c_{\text{ctx}}} \left[H(D(c_{\text{ctx}}(\theta); \hat{\phi})) \right] \tag{9}$$

Training Procedure: Parameters θ (generator) and ϕ (discriminator) are updated iteratively by alternating between minimizing $\mathcal{L}_D(\phi)$ and $\mathcal{L}_G(\theta)$. This encourages treatment-invariant representations while maintaining image quality.

G.4 Key Technical Innovations

TIDAL introduces three critical advances: (1) *Longitudinal Sequence Modeling*: Unlike DiffPO's static covariates, we employ LSTM-based temporal modeling of patient histories, capturing dynamic treatment decisions over time. (2) *Multi-Treatment Interval Modeling*: We predict treatment combinations over intervals rather than single point treatments, better reflecting clinical reality. (3) *Dual Causal Mechanism Integration*: We uniquely combine temporal IPW with adversarial training, providing complementary causal adjustments that neither method achieves alone.

This unified approach allows TIDAL to benefit from both explicit propensity-based reweighting and implicit treatment-invariant representation learning, resulting in superior causal performance.

H Limitations and Broader Impacts

H.1 Limitations

Our work has several limitations. While we used standard image fidelity metrics, we acknowledge their limitations in fully capturing clinically significant changes in longitudinal medical images; future work should explore more clinically relevant image-based evaluation metrics. Another limitation is that our method is described for counterfactual generation but is evaluated on factual outcomes. While synthetic counterfactual medical datasets exist, to our knowledge, none take into account longitudinal patient information, a critical component for medical utility and treatment-decision making.

H.2 Broader Impacts

Our research carries significant broader impacts regarding the crucial need for informed patient decision-making in osteoarthritis management. As highlighted by studies showing that patients often lack a clear understanding of potential treatment outcomes [2, 18], leading to suboptimal choices, tools that improve patient comprehension are vital. By enabling visualization of patient-specific future outcomes under different treatment scenarios, our framework has the potential to significantly enhance clinical decision support and facilitate shared decision-making. This visual aid can empower patients, fostering more informed and appropriate treatment pathways. A prospective clinical trial is needed to rigorously assess its clinical utility and impact on patient decision-making.

However, potential negative impacts require careful consideration. Risks include the generation of misleading or unrealistic images that could lead to incorrect clinical interpretations if not used responsibly. Fairness is crucial, as performance disparities across diverse patient subgroups could exacerbate healthcare disparities. Privacy concerns regarding sensitive medical data necessitate secure handling and deployment. Finally, the potential for misuse, such as generating fraudulent images, highlights the need for robust safeguards.