COUNTERFACTUAL 3D PELVIC FRACTURE CT SYNTHESIS VIA ANATOMY-GUIDED GEOMETRIC PRIORS AND REGION-AWARE LATENT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pelvic fractures are severe traumatic orthopedic injuries that exhibit an extreme long-tail data distribution, severely limiting the generalization of deep learning segmentation models on rare but clinically critical subtypes, such as Denis type-III sacral fractures. Unlike pathological anomalies that merely manifest as local texture changes, fractures involve complex geometric discontinuities of cortical bone and rigid displacements of detached fragments. Synthesizing such injuries requires anatomically valid fracture masks, but relying solely on real, scarce fracture annotations severely limits augmentation diversity. We address this by reformulating fracture synthesis as a local counterfactual editing task: generating a fracture on a healthy pelvic CT while strictly preserving the original soft tissue background. Our framework introduces a Geometric Fracture Simulation Module (GFSM) to construct explicit clinical topology priors directly from healthy bone masks, effectively decoupling the generation process from annotation scarcity. A Mask-Conditioned Latent Diffusion Model (MC-LDM) then synthesizes high-fidelity fracture CTs based on these geometric priors, optimized by a region-aware composite loss to preserve fine structural details. To ensure seamless structural integration, we employ a novel ghost-erasure and 3D Gaussian feathering pipeline. Evaluated on PENGWIN2024 and CTPelvic1k datasets, our method achieves state-of-the-art generation quality. Crucially, augmenting downstream 3D segmentation with our synthetic data improves the Dice score on Denis type-III fractures from 0.0291 to 0.6342 and reduces HD95 from 94.7068 mm to 26.3188 mm, directly demonstrating the clinical value of geometry-grounded synthesis in the long-tail trauma regime. Code and data: https://anonymous.4open.science/r/Counterfactual-Fracture-Synthesis-6DD3/.

1 INTRODUCTION

Pelvic fractures, typically resulting from high-energy trauma, account for 1–3% of all fractures and carry a mortality rate of up to 10% (Zheng, 2022; Mennen et al., 2023). Accurate preoperative planning increasingly relies on deep learning models trained to segment fracture regions and classify injury severity (Lindsey et al., 2018; Moolenaar et al., 2022). However, building such models faces a fundamental data bottleneck: real fracture CT data is scarce, expensive to annotate (Upadhyay & Bhandari, 2024; Chlap et al., 2021), and exhibits an extreme *long-tail distribution* (Buda et al., 2018; Zhang et al., 2023b). For instance, while common fracture subtypes may accumulate tens of cases, rare but critical subtypes—such as Denis type-III sacral fractures—may have as few as two training samples in public datasets (Sang et al., 2025). Consequently, this scarcity directly suppresses the segmentation performance on injuries most critical to surgical decision-making.

Data augmentation via generative synthesis offers a principled route to remedy this scarcity (Frid-Adar et al., 2018; Wang et al., 2021). Diffusion-based medical image synthesis has advanced substantially (Ho et al., 2020; Rombach et al., 2022; Kazerouni et al., 2023), and recent works have demonstrated high-fidelity generation of 3D CT volumes (Khader et al., 2023; Hamamci et al., 2024; Xu et al., 2024). In particular, Khader et al. (2023) show that synthetic 3D medical images generated by diffusion models can effectively augment downstream segmentation under data scarcity. These advances have motivated a line of work on *anomaly synthesis*: generating images of individuals under

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.

a pathological state—such as brain tumors, edema, or stroke lesions—while preserving personal identity and background, often framed as counterfactual image editing (Wolleb et al., 2022; Fontanella et al., 2025; Gu et al., 2023; Kumar et al., 2025; Alaya et al., 2024). In this pathological synthesis paradigm, generating an anomalous image requires specifying a target region and synthesizing plausible *texture* within it; the overall anatomy and organ morphology remain structurally intact.

While these methods excel at modeling pathological changes, pelvic fractures present a distinct challenge. They are traumatic structural injuries involving geometric discontinuities, rigid-body displacements, and complex surface irregularities, rather than localized alterations in tissue composition. Consequently, realistic fracture synthesis relies heavily on the *fracture mask itself* to encode precise geometric topology, rather than merely delineating a region for re-texturing. Furthermore, because real fracture masks are even scarcer than the CT images themselves, conditioning a diffusion model solely on the small pool of available masks yields limited augmentation diversity, restricting the benefits of generation for rare subtypes.

To overcome this bottleneck, we reformulate fracture synthesis as a *local counterfactual editing* task grounded in geometric simulation (Pearl, 2009; Wolleb et al., 2022; Fontanella et al., 2025). Specifically, given a healthy pelvic CT X_H and its bone mask M_H, we seek to synthesize the fracture while holding the surrounding soft tissue fixed. Our key innovation is decoupling mask generation from image synthesis: instead of relying on scarce real annotations, we construct synthetic fracture topology priors via anatomy-guided geometric simulation applied directly to healthy bone masks. To realize this pipeline, we design a complete generation closed-loop featuring three core components. First, a Geometric Fracture Simulation Module (GFSM) constructs diverse fracture topology priors on demand without requiring real fracture masks. Second, a Mask-Conditioned Latent Diffusion Model (MC-LDM) synthesizes high-fidelity CTs in latent space, supervised by a region-aware composite loss to preserve fine structural details. Finally, a Ghost-Erasure and 3D Gaussian Feathering pipeline seamlessly composites the synthesized fracture into the preserved soft tissue background. Experiments show that the framework can simulate fracture topology and improve downstream segmentation on rare categories. The largest gain is for Denis type-III small fragments, where augmentation increases Dice by about 21.8 times.

The main contributions of this work are:

- We propose a geometric-prior counterfactual synthesis framework for simulating fracture topology beyond texture-focused anomaly synthesis.
- We combine a Geometric Fracture Simulation Module, a mask-conditioned latent diffusion model, region-aware losses, ghost removal, and 3D Gaussian feathering for local synthesis with background preservation.
- Experiments show improved downstream segmentation on rare categories, including a 21.8$\times$ relative Dice gain for Denis type-III small fragments; this evaluates task utility rather than clinical validity.

2 Related Work

Medical CT image synthesis. Generative models (Goodfellow et al., 2014; Kingma & Welling, 2014; Ho et al., 2020) offer an effective solution to medical image data scarcity. Early GAN-based approaches like CycleGAN (Zhu et al., 2017) enabled unpaired translation but suffered from training instability and limited structural fidelity. Diffusion models (Ho et al., 2020; Rombach et al., 2022) have since surpassed GANs in mode coverage and stability (Dhariwal & Nichol, 2021). Methods like MedDDPM (Dorjsembe et al., 2024) demonstrate superior 3D volumetric synthesis; building on this, GenerateCT (Hamamci et al., 2024) and 3D MedDiffusion (Wang et al., 2025) further expand controllability and resolution. Text-conditioned variants such as MediSyn (Cho et al., 2025) and GuideGen (Dai et al., 2025) additionally show semantic prompts can guide whole-body and pathological CT synthesis (Yu et al., 2024a). Despite these advances, existing work exhibits a clear bias toward soft-tissue organs, heavily relying on chest (Han et al., 2023) and abdominal (Wang et al., 2024; Zhuang et al., 2024) datasets, leaving bone structure generation underexplored. This imbalance poses unique challenges for pelvic trauma tasks, where pathological samples are inherently scarcer (Iyer et al., 2023). Unlike mainstream organ-level studies, we focus on 3D pelvic fracture synthesis, proposing a generative framework robust to severe data constraints.

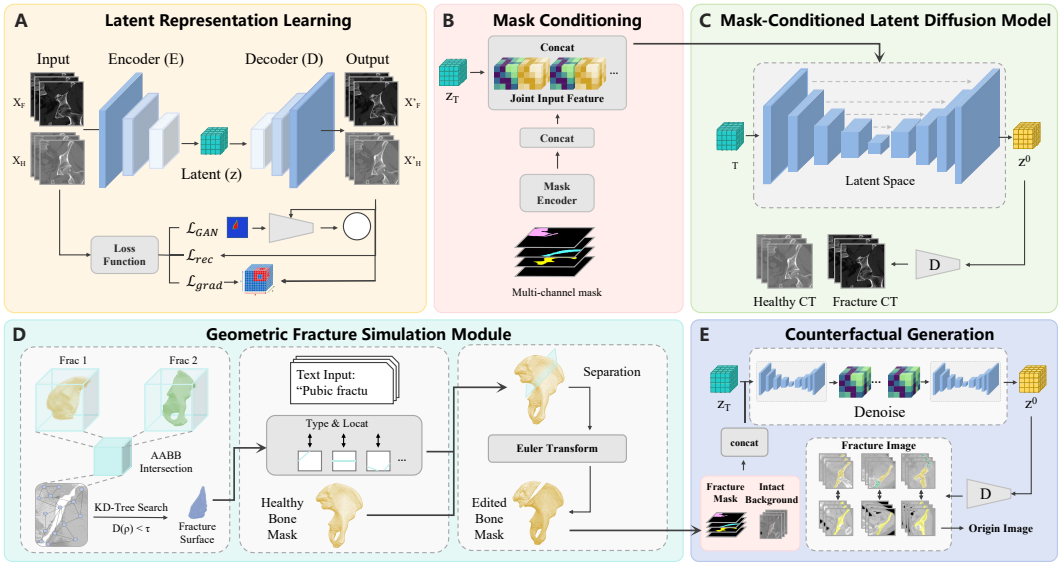

Figure 1: Overview of the proposed anatomy-guided 3D counterfactual fracture synthesis framework. (A) Region-aware reconstruction module (3D VAE-GAN); (B) Mask condition encoding; (C) Mask-conditioned latent diffusion model (MC-LDM); (D) Geometric Fracture Simulation Module (GFSM); (E) Counterfactual generation pipeline.

Counterfactual generation in medical imaging. Rooted in causal inference (Pearl, 2009), counterfactual generation in medical imaging initially focused on anomaly detection, synthesizing "pseudo-healthy" counterparts to localize lesions (Baur et al., 2021; Fontanella et al., 2025; Bi et al., 2025). Recently, the focus shifted to anomaly synthesis for data augmentation, targeting long-tailed pathological data distributions. Nie et al. (2025) propose a causal decoupling framework that exchanges pathological features while preserving anatomy. MedDiff-FM (Yu et al., 2024b) provides a foundation model for lesion inpainting via text or mask prompts, supporting cross-organ tumor synthesis. Controllable generation is further advanced by general-purpose frameworks like ControlNet (Zhang et al., 2023a) and improved diffusion schemes (Nichol & Dhariwal, 2021). However, a fundamental limitation persists: these methods rely on *texture-based manipulation* targeting soft-tissue composition changes. Such strategies inherently fail to model the structural nature of fractures—bone discontinuity, rigid fragment displacement, and cortical angulation—which constitute *traumatic topological mutations* rather than texture deviations. Our work introduces explicit geometric priors to address this gap, enabling anatomically plausible fracture topology synthesis.

3 Method

3.1 Overview

As illustrated in Figure 1, the framework comprises three core components: a **3D VAE-GAN** (Kingma & Welling, 2014; Esser et al., 2021) for latent representation learning, a **Geometric Fracture Simulation Module (GFSM)** that constructs counterfactual fracture topology priors from healthy anatomical masks via anatomy-guided geometric rules, and a **Mask-Conditioned Latent Diffusion Model (MC-LDM)** that synthesizes high-fidelity fracture CTs conditioned on GFSM-generated masks.

The pipeline proceeds as follows: GFSM first produces a counterfactual fracture mask; a ghost-erasure step cleans the source CT; MC-LDM then denoises conditional on this mask; and 3D Gaussian feathering fuses the result with the preserved background.

3.2 Region-Aware Reconstruction Loss

Traditional voxel-wise reconstruction losses frequently fail to preserve high-frequency details critical to fracture imaging, such as fine fracture lines and trabecular textures. To mitigate this, we formulate

a region-aware composite optimization objective for the 3D VAE-GAN. The total loss function is defined as:

$$\mathcal{L}_{AE} = \mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{GAN}\mathcal{L}_{GAN}^{ROI} + \lambda_{grad}\mathcal{L}_{grad} \tag{1}$$

where $\mathcal{L}_{rec}$ and $\mathcal{L}_{KL}$ ensure basic voxel consistency and regularization of the latent distribution, respectively. The region-aware adversarial loss $\mathcal{L}_{GAN}^{ROI}$ directs the discriminator to focus on the complex topology of bone trauma via spatial weighting. To counteract the "mean-regression" smoothing effect inherent in pixel-level losses, we introduce an explicit 3D gradient loss $\mathcal{L}_{grad}$. This term constrains the first-order derivative domain, forcing the generated image to maintain texture gradients consistent with the ground truth across all three orthogonal directions. Consequently, this high-frequency constraint sharply defines the geometric boundaries of fracture surfaces and trabecular edges.

The ROI mask used by $\mathcal{L}_{GAN}^{ROI}$ is obtained fully automatically from the fragment label maps, with no manual annotation: we extract the interfacial voxels between adjacent fragments (i.e., the fracture surface) to form a binary fracture-surface mask $M_{fs} \in \{0,1\}^{D\times H\times W}$, and define a spatial weighting map $W = 1 + (\alpha - 1)M_{fs}$ (with $\alpha \geq 1$) that up-weights the fracture-surface region relative to the rest of the volume. The discriminator is a 3D PatchGAN (three stride-2 convolutional blocks plus one stride-1 block, each with 3D batch normalization and LeakyReLU(0.2)) producing a single-channel, low-resolution logit map; W is downsampled to that resolution by nearest-neighbour interpolation, denoted $W_{\downarrow}$. With the hinge loss $L_{hinge}(a,b) = \frac{1}{2}(\mathbb{E}[\max(0, 1-a)] + \mathbb{E}[\max(0, 1+b)])$, the discriminator and generator adversarial objectives are

$$L_D = L_{hinge}(D(x), D(\hat{x})) + \lambda_{roi} \cdot L_{hinge}(W_{\downarrow} \odot D(x),\ W_{\downarrow} \odot D(\hat{x})), \tag{2}$$

$$\mathcal{L}_{GAN}^{ROI} = -\mathbb{E}[D(\hat{x})] - \lambda_{roi} \cdot \mathbb{E}[W_{\downarrow} \odot D(\hat{x})]. \tag{3}$$

The adversarial term is adaptively scaled by $\gamma = \|\nabla_\theta \mathcal{L}_{rec}\| / (\|\nabla_\theta \mathcal{L}_{GAN}^{ROI}\| + 10^{-4})$, computed on the final decoder layer and clipped to $[0, 10^4]$; it is activated only after a warm-up iteration t_{warm}.

3.3 Mask-Conditioned Latent Diffusion Model

For topologically controllable fracture generation, we propose a pure mask-guided conditional latent diffusion architecture (MC-LDM). During training, real 3D CT images and their corresponding fracture masks serve as paired supervision.

Because the input fracture mask y_{mask} contains discrete labels with distinct anatomical semantics, it is first converted into a multi-channel one-hot encoding to decouple the semantics of different bone fragments from the background. This encoded mask is subsequently processed by a domain-specific encoder τ_{mask}, which shares the VAE architecture of the first stage, mapping the mask to a continuous latent condition vector:

$$c = \tau_{mask}(y_{mask}) \tag{4}$$

The clean volume x is encoded to the latent $z_0 = \mathcal{E}(x)$ by the first-stage VAE, and the forward diffusion process produces the noised latent

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{5}$$

where $\bar{\alpha}_t$ is the standard cumulative noise schedule; z_T denotes the fully noised latent at the terminal diffusion step, i.e., the $\mathcal{N}(0, I)$ sample that initializes reverse sampling. The mask condition c is injected into the 3D U-Net (Çiçek et al., 2016; Ronneberger et al., 2015) denoising backbone by *channel-wise concatenation*. Because τ_{mask} shares the first-stage VAE architecture, the condition features c are spatially aligned with the noisy latent z_t; the two are concatenated along the channel dimension to form the input to the denoising network. The timestep t is injected separately through a sinusoidal timestep embedding added to the residual blocks. The noise prediction is thus

$$\epsilon_\theta(z_t, t, c) = \mathrm{UNet}\big([\, z_t \,\|\, c\,],\ \mathrm{emb}(t)\big), \tag{6}$$

where $[\,\cdot\,\|\,\cdot\,]$ denotes concatenation along the channel dimension. Conditioning the denoiser on the spatially aligned mask features in this way anchors the synthesized content to the fracture topology encoded in c, so that the generated fracture conforms to the boundaries specified by the mask. The training objective is formulated as a mask-constrained noise prediction task:

$$\mathcal{L}_{MC\text{-}LDM} = \mathbb{E}_{z_0,\, \epsilon\sim\mathcal{N}(0,1),\, t,\, y_{mask}}\big[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\big] \tag{7}$$

By optimizing this objective, MC-LDM learns the conditional distribution $p(z_0 \mid y_{mask})$, enabling high-fidelity image generation precisely governed by the fracture mask.

3.4 GEOMETRIC FRACTURE SIMULATION MODULE

Because existing generative models struggle with explicit topological disruptions, the GFSM algorithmically implants geometrically simulated fractures into healthy bone masks. Driven by a clinical text prompt T, this module transforms a healthy anatomical mask into a counterfactual fracture mask via three stages.

Stage 1: Fracture prior mining. To construct a high-quality fracture surface atlas $\mathcal{L}_{surf}$, we mine fracture-surface templates from real fractured bones. For a real fractured bone, let M_{prox}^{real} and M_{dist}^{real} denote its proximal and distal fragments, respectively. We first compute the axis-aligned bounding boxes (AABBs) of the fragment pair and take their spatial intersection Ω_{int}, i.e., the region where the two fragments are adjacent and the fracture surface must lie; Ω_{int} serves as a cheap pre-filter that restricts the subsequent KD-tree search. Within Ω_{int}, a KD-tree nearest-neighbour search is employed for voxel-level retrieval. For any voxel v on the proximal fragment boundary ∂M_{prox}^{real}, its Euclidean distance to the distal fragment boundary ∂M_{dist}^{real} is:

$$d(v,\, \partial M_{dist}^{real}) = \min_{u \in \partial M_{dist}^{real}} \|v - u\|_2 \tag{8}$$

Regions satisfying $d(v, \partial M_{dist}^{real}) < \tau$ (with $\tau = 3\,\text{mm}$) are extracted, normalized as raw fracture surface templates S_{raw}, and stored in $\mathcal{L}_{surf}$ alongside anatomical semantic labels (e.g., "Pubic", "Sacrum Zone II").

Stage 2: Text-guided semantic-geometric retrieval. A retrieval mapping $\mathcal{R}(T) \to (S_{ref}, \theta_{trans})$ bridges clinical text descriptions with geometric priors. The input text T is parsed for anatomical site and fracture type, enabling the retrieval of a semantically matched fracture surface prior S_{ref} from $\mathcal{L}_{surf}$; retrieval is performed by exact label matching on anatomical site and fracture type, and when multiple templates share the same label, one is chosen uniformly at random. Here θ_{trans} denotes the displacement/rotation parameter range associated with the retrieved template, i.e., the distribution from which the rigid-transform parameters $(\alpha, \beta, \gamma, \mathbf{t})$ in Eq. 9 are sampled once a template is selected. This S_{ref} encapsulates the fracture surface geometry of the target injury and acts as a reference template for the subsequent cut-and-displace operation.

Stage 3: Topological deformation. S_{ref} is mapped to the coordinate system of the target healthy mask M_h via centroid alignment, topologically splitting M_h into a proximal fragment M_{prox} and a distal fragment M_{dist}. To simulate trauma-induced displacement, M_{dist} is treated as a rigid body and subjected to a rigid transformation parameterized by Euler angles (α, β, γ) and translation vector $\mathbf{t} = [t_x, t_y, t_z]^T$. Since a rotation cannot be left-multiplied onto a label volume, the transformation is applied by warping the coordinate grid and resampling with nearest-neighbour interpolation:

$$M'_{dist}(x) = M_{dist}\big(R_x(\alpha)^T R_y(\beta)^T R_z(\gamma)^T\,(x - c - \mathbf{t}) + c\big)\,, \tag{9}$$

where x is a voxel coordinate and c is the fragment centroid about which the rotation is applied. The sampling distributions for $(\alpha, \beta, \gamma, \mathbf{t})$ are constrained by the injury severity specified in T. The final counterfactual fracture mask is thus $M_{frac} = M_{prox} \cup M'_{dist}$.

3.5 ANATOMY-GUIDED COUNTERFACTUAL GENERATION AND SEAMLESS FUSION

Having established the fracture topology via GFSM and synthesized the latent features via MC-LDM, the remaining challenge is seamlessly fusing the generated fracture into the original healthy CT without introducing visible artifacts. We address this through a two-step pipeline, as shown in Figure 2.

Healthy bone ghost erasure. Direct projection overlaps the original healthy bone; the partial-volume cortical halo survives simple masking and causes whitening. We therefore use a 'Disguise–Erase–Restore' pipeline (Boolean difference + dilation, soft-tissue median disguise, TELEA inpainting, z-axis Gaussian smoothing) to obtain a ghost-free background X_{base}. Detailed formulations are provided in Appendix C.

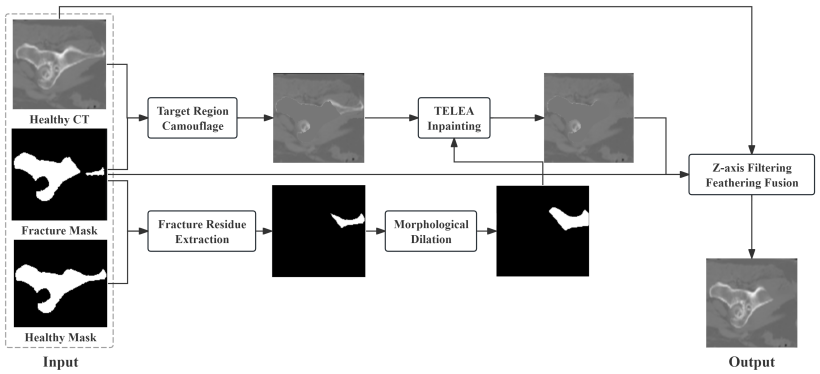

Figure 2: The "Disguise–Erase–Restore" 3D bone ghost cleaning pipeline. The pipeline applies workspace disguise, morphological dilation, TELEA inpainting, z-axis Gaussian filtering, and feathering fusion to produce a clean soft-tissue background free of cortical bone halos.

Mask-conditioned inference and pixel-level seamless compositing. Given the clean background X_{base}, MC-LDM performs guided denoising sampling to yield the synthesized image X_{pred}. The binary fracture mask M_{frac} is then spatially smoothed via a 3D Gaussian convolution kernel $\mathcal{G}_{\sigma_{blend}}$ to generate a soft blend mask:

$$M_{soft}(x) = (M_{frac} * \mathcal{G}_{\sigma_{blend}})(x), \quad M_{soft}(x) \in [0, 1] \tag{10}$$

The final counterfactual image is computed via element-wise weighted fusion:

$$X_{final} = M_{soft} \odot X_{pred} + (1 - M_{soft}) \odot X'_{base} \tag{11}$$

where X'_{base} denotes the background post-VAE encoding and decoding. This soft-boundary alignment reduces stitching seams, anchors the synthesized fracture at its intended position, and preserves the background texture of X_{origin} outside the edit region.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We evaluate on two 3D CT datasets comprising variable-resolution volumetric scans. **PENGWIN2024** (Sang et al., 2025) contains 100 pelvic fracture CT scans. After strict filtering—splitting each volume into sacrum, left hip, and right hip sub-units; removing non-fractured portions; and excluding extreme comminuted cases (>4 fragments)—we obtain 153 valid fracture samples (43 sacrum, 55 left hip, 55 right hip). **CTPelvic1k** (Liu et al., 2021) provides large-scale healthy pelvic CT data. We select 153 anatomically matched healthy samples for joint training, and reserve an independent healthy subset for counterfactual inference evaluation.

Data partitioning and leakage control. L_{surf} is mined only from the PENGWIN2024 training split; evaluation counterfactuals use an independent CTPelvic1k healthy subset disjoint from all PENGWIN data, so Table 3 involves no template leakage.

Preprocessing. To standardize the heterogeneous native spacings of the clinical scans, all CT volumes are cropped to skeletal mask bounding boxes and resampled to a uniform $128 \times 128 \times 128$ voxel grid via trilinear interpolation. Intensity values are clipped to $[-1000, 1200]$ HU to emphasize bone structures and linearly normalized to $[-1, 1]$. Data are partitioned into 7:2:1 splits for training, validation, and testing.

Implementation details. The 3D VAE-GAN is trained for 120 epochs (compression factor $(8, 8, 8)$, learning rate 10^{-4}, batch size 2). MC-LDM is trained for 1000 epochs with T=1000 diffusion timesteps, utilizing the AdamW optimizer (Loshchilov & Hutter, 2019) with a linear warm-up to 5×10^{-5} after 26,000 steps, bfloat16 mixed-precision, and gradient accumulation. Inference employs 50-step DDIM (Song et al., 2021) sampling on a single NVIDIA RTX 4090 GPU (model parameter size: 1.77 GB).

Evaluation metrics. Generation quality is rigorously measured by Peak Signal-to-Noise Ratio (PSNR) (Wang & Bovik, 2009), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), and Fréchet Inception Distance (FID) (Heusel et al., 2017). Downstream segmentation performance is evaluated using the Dice Similarity Coefficient (DSC) and the 95th-percentile Hausdorff Distance (HD95).

4.2 Synthesis Quality Evaluation

Qualitative visualization of GFSM-simulated fracture masks across all six subtypes is provided in Appendix B.

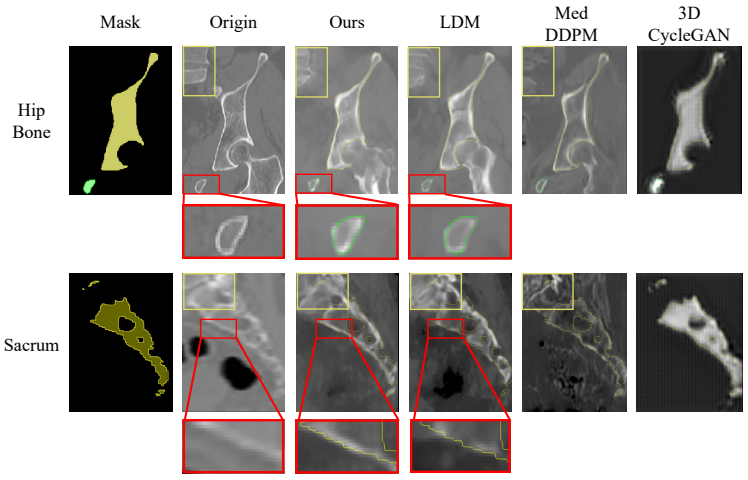

Figure 3: Qualitative comparison of 3D fracture generation in hip and sacrum regions. Columns: input fracture mask, real reference CT, and results from our method, LDM, Med-DDPM, and 3D CycleGAN. Yellow boxes highlight background reconstruction; red boxes highlight bone boundary detail.

Quantitative comparison. Our method is benchmarked against **3D CycleGAN** (Zhu et al., 2017), **Med-DDPM** (Dorjsembe et al., 2024), and **LDM** (Rombach et al., 2022) under a consistent training and evaluation framework. For a fair comparison, all models are conditioned exclusively on identical fracture masks, thereby eliminating potential biases arising from disparate input signals. We note that PSNR, SSIM, LPIPS, and FID are all computed over the entire volume rather than being restricted to the fracture ROI, and this scope is applied identically to all methods.

Table 1: Quantitative comparison of synthesis quality on pelvic CT. Bold: best; underline: second best.

Metric	Region	3D CycleGAN	Med-DDPM	LDM	Ours
PSNR ↑	Left Hip	11.0387	19.3973	21.8512	**22.0139**
	Sacrum	12.9010	19.7252	22.3507	**22.4087**
	Right Hip	11.4911	18.0496	19.6408	**20.4489**
	Avg.	11.8103	19.0574	21.2809	**21.6238**
SSIM ↑	Left Hip	0.5132	0.7946	**0.8149**	0.8120
	Sacrum	0.5816	0.7224	0.7335	**0.7455**
	Right Hip	0.5570	0.7956	**0.8193**	0.8160
	Avg.	0.5506	0.7709	0.7892	**0.7912**
LPIPS ↓	Left Hip	0.4283	0.2583	0.2366	**0.2337**
	Sacrum	0.4634	0.3169	**0.2969**	**0.2969**
	Right Hip	0.4490	0.2400	0.2136	**0.2164**
	Avg.	0.4469	0.2717	**0.2490**	**0.2490**
FID ↓	Left Hip	635.8308	63.2149	27.5727	**27.0205**
	Sacrum	621.7141	34.4943	16.2226	**14.6014**
	Right Hip	638.7555	60.3726	19.0045	**17.3581**
	Avg.	632.1001	52.6939	20.9332	**19.6600**

As shown in Table 1, Diffusion models clearly outperform 3D CycleGAN (avg. PSNR 11.81 dB, FID 632.10). Against the strongest baseline LDM, our MC-LDM is comparable on rendering metrics (avg. LPIPS identical at 0.2490, $p > 0.6$), but more consistent in the complex sacrum, where its per-sample LPIPS IQR is roughly half that of LDM (Figure 8). Our central claims rest on GFSM and the downstream augmentation gain (Table 3), not on the rendering backbone alone.

Qualitative comparison. Figure 3 shows checkerboard artifacts and global blurring for 3D CycleGAN. Med-DDPM and LDM recover basic bone morphology but show blurred boundaries in these examples. Our method more closely follows the mask-defined discontinuities, with sharper trabecular and boundary details.

4.3 Counterfactual Generation with Background Preservation

Comparison with edge-conditioned baseline. To rigorously validate our "pure mask generation + image-space fusion" strategy, we compare against an edge-map baseline. This alternative concatenates a 3D Canny edge map (with the target fracture region masked out) with the fracture mask, feeding both as dual conditions end-to-end into the diffusion model.

Table 2: Counterfactual generation quality: pure-mask (Ours) vs. edge-map baseline.

Region	Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Hip Bone	Edge-map	23.1403	0.7983	0.2231	50.5112
	Ours	**29.6288**	**0.9155**	**0.1142**	**24.0380**
Sacrum	Edge-map	21.9846	0.7570	0.2782	**7.4459**
	Ours	**29.8759**	**0.9012**	**0.1475**	12.0114

As detailed in Table 2, the pure-mask design improves PSNR, SSIM, and LPIPS in both regions. In the hip, it also improves FID (24.04 vs. 50.51), whereas the edge baseline has lower FID in the sacrum (7.45 vs. 12.01). Thus, the results support improved voxel-level/background fidelity but not uniform dominance on every distributional metric. The lower PSNR and SSIM of the edge baseline are consistent with conflicts between noisy inter-slice edges, local fracture topology, and background reconstruction.

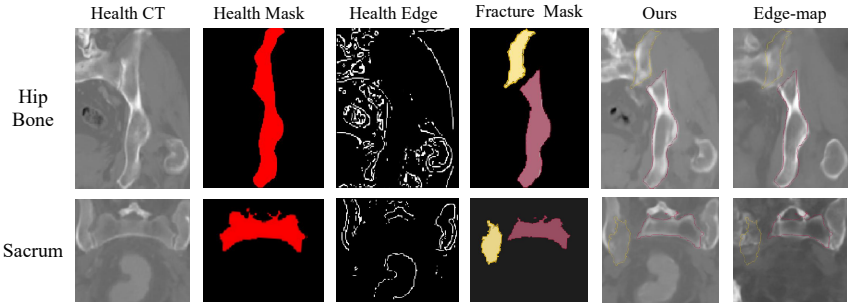

Figure 4: Qualitative comparison of counterfactual generation under different anatomical prior mechanisms. Rows: hip bone (top) and sacrum (bottom). Columns: healthy source CT, healthy mask/edge map with the target fracture region suppressed, input fracture mask, our pure-mask result, and edge-map result.

Qualitative results. As demonstrated in Figure 4, our method strictly adheres to the counterfactual principle of "local modification with global preservation." Soft tissues, muscle textures, and unaffected bone structures remain visually consistent with the original healthy CT in these examples. This qualitative evidence is consistent with local modification and global preservation, but does not by itself establish generalization beyond the evaluated distribution. Additional multi-planar examples across all fracture subtypes are provided in Appendix F.

4.4 Long-Tail Augmentation and Downstream Segmentation

To address the severe class imbalance inherent in PENGWIN2024 (e.g., 33 pubic fractures vs. only 2 Denis type-III sacral fractures), we design an adaptive augmentation strategy targeting 30 samples per category, yielding 111 counterfactual fracture samples. Specifically, 28 samples are synthesized for Denis type-III (a $15\times$ minority class expansion), while adequately represented categories receive no additional samples. We subsequently evaluate downstream generalization using a 3D U-Net (Çiçek et al., 2016) framework, comparing a *Baseline* model (real data only) against an *Augmented* model (real + synthetic data).

As presented in Table 3, the baseline model severely struggles to segment small fracture fragments, particularly within the rare Denis type-III category (DSC = 0.0291). Following data augmentation,

Table 3: Quantitative segmentation performance before and after augmentation on pelvic CT. Bold: improved performance. Δ: Dice (After−Before); HD95 (Before−After).

Fragment	Region	Dice ↑			HD95 (mm) ↓		
		Before Aug	After Aug	Δ	Before Aug	After Aug	Δ
Main fragment	Pubic	0.8756	0.9239	**0.0483**	26.3283	15.2787	**11.0496**
	Iliac	0.8329	0.8780	**0.0451**	40.6188	26.6596	**13.9592**
	Acetabular	0.9402	0.9233	-0.0169	30.7994	29.1000	**1.6994**
	I	0.8423	0.9527	**0.1104**	6.8893	3.3167	**3.5726**
	II	0.6998	0.8897	**0.1899**	13.7534	14.4702	-0.7168
	III	0.7038	0.6659	-0.0379	26.3805	10.9839	**15.3966**
Small fragment	Pubic	0.6768	0.6885	**0.0117**	39.0263	69.1633	-30.1370
	Iliac	0.4806	0.5225	**0.0419**	81.6265	80.5805	**1.0460**
	Acetabular	0.5425	0.6798	**0.1373**	49.1200	50.2102	-1.0902
	I	0.2123	0.7901	**0.5778**	33.9953	12.1493	**21.8460**
	II	0.1414	0.7779	**0.6365**	56.9510	34.1470	**22.8040**
	III	0.0291	0.6342	**0.6051**	94.7068	26.3188	**68.3880**

our model dramatically elevates the Denis type-III small fragment DSC to 0.6342 (an approximately $21.8\times$ relative improvement) and substantially reduces the HD95 boundary error from 94.71 mm to 26.32 mm. Significant performance gains are consistently observed across other challenging regions, such as Denis I and II small fragments (absolute DSC improvements of 0.5778 and 0.6365, respectively). These results indicate that the synthetic data can improve downstream segmentation for rare fracture types. The experiment evaluates benchmark task performance; it does not establish diagnostic generalization or clinical effectiveness. A few saturated categories regress marginally (acetabular main fragment -0.0169 DSC; Denis III main fragment -0.0379 DSC), the expected trade-off of class-rebalancing toward the rare target classes.

4.5 Ablation Study

Table 4: Ablation study of loss functions in the VAE-GAN reconstruction module. Bold: best.

Version	$\mathcal{L}_{GAN}^{ROI}$	$\mathcal{L}_{grad}$	SSIM ↑				PSNR ↑			
			L.Hip	R.Hip	Sacrum	Avg	L.Hip	R.Hip	Sacrum	Avg
v0			0.9417	0.9424	0.9330	0.9390	32.7108	32.8090	31.2671	32.2623
v1	✓		0.9429	0.9442	0.9393	0.9421	32.8206	32.8206	**32.4808**	32.7073
Full	✓	✓	**0.9473**	**0.9492**	**0.9434**	**0.9466**	**32.9868**	**33.4534**	32.0356	**32.8253**

Table 4 shows incremental gains from the region-aware adversarial loss and the 3D gradient loss. The Full model achieves the best average SSIM and PSNR, as well as the best hip-region values. Its sacrum PSNR is lower than v1 (32.04 vs. 32.48 dB) but remains higher than v0 (31.27 dB), an improvement of 0.77 dB over the baseline. These results support the composite loss while indicating that its effects vary by anatomical region.

5 Conclusion and Limitation

We identified reliance on scarce fracture masks as a bottleneck in traumatic injury synthesis. Our geometry-grounded framework couples GFSM with a region-aware mask-conditioned LDM, decouples topology generation from real annotations, and improves downstream segmentation in the evaluated benchmark. Augmentation improves Dice for Denis type-III small fragments from 0.0291 to 0.6342 (+21.8×) and reduces HD95, supporting task utility for long-tail augmentation; this is not a clinical validation.

However, the fracture surfaces produced by geometric simulation are regular "smooth-cut" boundaries that lack the surface irregularity of real fractures, as the current GFSM does not incorporate biomechanical factors such as bone density, patient age, or the direction of impact force. Furthermore, our evaluation is limited to pelvic CT and does not yet cover other skeletal sites. We leave the incorporation of richer physiological priors and the extension to broader anatomical regions for future work.

Ethics Statement

This work relies exclusively on two publicly available, de-identified pelvic CT datasets, PENGWIN2024 (Sang et al., 2025) and CTPelvic1k (Liu et al., 2021); we collected no new patient data, and informed consent and ethical approval for the original acquisitions are the responsibility of

the respective dataset providers. Our method synthesizes counterfactual fracture CTs strictly for data augmentation in research settings. We emphasize that the synthetic images are not intended for clinical diagnosis or treatment decisions, and that any deployment in a medical context would require dedicated clinical validation and regulatory approval. We foresee no additional risks to human subjects, privacy, or fairness arising from this work.

AI USE STATEMENT

In this work, generative AI tools were used to aid and polish writing, including grammar, clarity, and readability. They were also used to assist with research ideation and execution, including brainstorming methodological directions and assisting with parts of the experimental implementation. All research ideas, experimental choices, code, analyses, results, and claims were checked and validated by the authors. The authors take full responsibility for the correctness and integrity of the final content and artifacts. See the ICLR 2027 AI Policy for Authors for more details.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide comprehensive implementation details in Section 4.1, including data filtering criteria, preprocessing steps, network training schedules, optimizer settings, and evaluation metrics. The GFSM algorithm is fully specified with mathematical formulations in Section 3.4, and the ghost-erasure pipeline is detailed in Appendix C. We use two publicly available datasets, PENGWIN2024 (Sang et al., 2025) and CTPelvic1k (Liu et al., 2021), with data splits described in Section 4.1. Our code and the geometric fracture simulation module are released at the anonymous repository `https://anonymous.4open.science/r/Counterfactual-Fracture-Synthesis-6DD3/`.

REFERENCES

Mounir Ben Alaya, Daniel M. Lang, Benedikt Wiestler, Julia A. Schnabel, and Ivan Ezhov. MedEdit: Counterfactual diffusion-based image editing on brain MRI. In *Simulation and Synthesis in Medical Imaging (SASHIMI), MICCAI Workshop*, pp. 167–176, 2024.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. *Medical Image Analysis*, 69:101952, 2021.

Yuan Bi, Lucie Huang, Ricarda Clarenbach, Reza Ghotbi, Angelos Karlas, Nassir Navab, and Zhongliang Jiang. Synomaly noise and multi-stage diffusion: A novel approach for unsupervised anomaly detection in medical images. *Medical Image Analysis*, 105:103737, 2025.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

Joseph Cho, Mrinal Mathur, Cyril Zakka, et al. MediSyn: A generalist text-guided latent diffusion model for diverse medical image synthesis. *arXiv preprint arXiv:2405.09806*, 2025.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 424–432, 2016.

Linrui Dai, Rongzhao Zhang, Yongrui Yu, et al. GuideGen: A text-guided framework for paired full-torso anatomy and CT volume generation. *arXiv preprint arXiv:2403.07247*, 2025.

Francis Denis, Sigrid Davis, and Tom Comfort. Sacral fractures: An important problem. retrospective analysis of 236 cases. *Clinical Orthopaedics and Related Research*, 227:67–81, 1988.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3D brain MRI synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, 2021.

Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. *IEEE Transactions on Medical Imaging*, 44(9):3574–3585, 2025.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yu Gu, Jianwei Yang, Naoto Usuyama, et al. BiomedJourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. *arXiv preprint arXiv:2310.10765*, 2023.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, et al. GenerateCT: Text-conditional generation of 3D chest CT volumes. In *European Conference on Computer Vision (ECCV)*, pp. 126–143, 2024.

Kun Han, Yifeng Xiong, Chenyu You, Pooya Khosravi, Shanlin Sun, Xiangyi Yan, James S. Duncan, and Xiaohui Xie. MedGen3D: A deep generative framework for paired 3D image and mask generation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 759–769, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Krithika Iyer, Alan Morris, Brian Zenger, Karthik Karanth, Nawazish Khan, Benjamin A. Orkild, Oleksandre Korshak, and Shireen Elhabian. Statistical shape modeling of multi-organ anatomies with shared boundaries. *Frontiers in Bioengineering and Biotechnology*, 10:1078800, 2023.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13(1):7303, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Amar Kumar, Anita Kriz, Mohammad Havaei, and Tal Arbel. PRISM: High-resolution and precise counterfactual medical image generation using language-guided stable diffusion. *arXiv preprint arXiv:2503.00196*, 2025.

Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, and Hollis Potter. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 115(45):11591–11596, 2018.

Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, et al. Deep learning to segment pelvic bones: large-scale CT datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):749–756, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Eric G. Meinberg, Julie Agel, Craig S. Roberts, Matthew D. Karam, and James F. Kellam. Fracture and dislocation classification compendium—2018. *Journal of Orthopaedic Trauma*, 32:S1–S170, 2018.

A. H. M. Mennen, A. S. Blokland, M. Maas, F. W. Bloemers, L. M. G. Geeraedts Jr., and N. W. L. Schep. Imaging of pelvic ring fractures in older adults and its clinical implications: a systematic review. *Osteoporosis International*, 34(9):1549–1559, 2023.

Joost Z. Moolenaar, Nazli Tümer, and Sara Checa. Computer-assisted preoperative planning of bone fracture fixation surgery: A state-of-the-art review. *Frontiers in Bioengineering and Biotechnology*, 10:1037048, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8162–8171, 2021.

Wei-zhi Nie, Zichun Zhang, Weijie Wang, Bruno Lepri, An-An Liu, and Niculae Sebe. Causal disentanglement for robust long-tail medical image generation. *arXiv preprint arXiv:2504.14450*, 2025.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Yudi Sang, Yanzhen Liu, Sutuke Yibulayimu, et al. Benchmark of segmentation techniques for pelvic fracture in CT and X-ray: Summary of the PENGWIN 2024 challenge. *arXiv preprint arXiv:2504.02382*, 2025.

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

Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1):23–34, 2004.

Ashish Kumar Upadhyay and Ashish Kumar Bhandari. Advances in deep learning models for resolving medical image segmentation data scarcity problem: A topical review. *Archives of Computational Methods in Engineering*, 31(3):1701–1719, 2024.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3D MedDiffusion: A 3D medical diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging*, 44(12):4960–4972, 2025.

Tonghe Wang, Yang Lei, Yabo Fu, Jacob F. Wynne, Walter J. Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of Applied Clinical Medical Physics*, 22(1):11–36, 2021.

Yuran Wang, Zhijing Wan, Yansheng Qiu, and Zheng Wang. Devil is in details: Locality-aware 3D abdominal CT volume generation for self-supervised organ segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10640–10648, 2024.

Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

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.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Diffusion models for medical anomaly detection. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 35–45, 2022.

Yanwu Xu, Li Sun, Wei Peng, Shyam Visweswaran, and Kayhan Batmanghelich. MedSyn: Text-guided anatomy-aware synthesis of high-fidelity 3-d CT images. *IEEE Transactions on Medical Imaging*, 43(10):3648–3660, 2024.

Yongrui Yu, Hanyu Chen, Zitian Zhang, Qiong Xiao, Wenhui Lei, Linrui Dai, Yu Fu, Hui Tan, Guan Wang, Peng Gao, and Xiaofan Zhang. CT synthesis with conditional diffusion models for abdominal lymph node segmentation. *arXiv preprint arXiv:2403.17770*, 2024a.

Yongrui Yu, Yannian Gu, Shaoting Zhang, and Xiaofan Zhang. MedDiff-FM: A diffusion-based foundation model for versatile medical image applications. *arXiv preprint arXiv:2410.15432*, 2024b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023a.

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.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023b.

Qiubao Zheng. Research progress on the evaluation indicators for the efficacy of pelvic fracture treatment. *Chinese Journal of Bone and Joint Injury*, 37(12):1338–1340, 2022.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.

Yan Zhuang, Benjamin Hou, Tejas Sudharshan Mathai, Pritam Mukherjee, Boah Kim, and Ronald M. Summers. Semantic image synthesis for abdominal CT. In *Deep Generative Models Workshop (DGM4MICCAI)*, pp. 214–224, 2024.

A Pelvic Anatomy and Fracture Classification

Figure 5 summarizes the anatomical structure of the pelvis and the fracture classification systems used in this work. The hip bone (os coxae) consists of three fused bones: the ilium (upper wing), the ischium (posterior-inferior), and the pubis (anterior-inferior), meeting at the acetabulum—the cup-shaped socket of the hip joint. Fractures of these regions are classified according to the AO/OTA Fracture and Dislocation Classification Compendium (Meinberg et al., 2018).

The sacrum is the triangular bone at the base of the spine, and sacral fractures are classified using the Denis system (Denis et al., 1988) based on the anatomical zone of injury: Zone I (lateral sacral ala, not involving the foramina), Zone II (crossing the sacral foramina, risking L5/S1 nerve root injury), and Zone III (involving the central spinal canal, highest risk of neurological deficit). Denis type-III fractures are the rarest and most clinically severe, with only 2 annotated training samples in the PENGWIN2024 dataset—making them the primary target of our long-tail augmentation strategy.

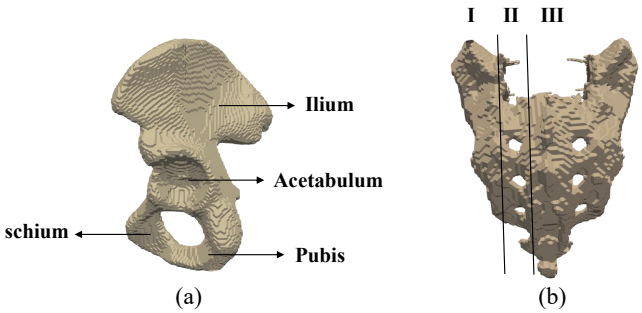

Figure 5: Pelvic anatomy and fracture classification. (a) Hip bone anatomy showing iliac, ischial, and pubic regions with the acetabular joint surface. (b) Denis classification of sacral fractures: Zone I (sacral ala), Zone II (sacral foramina), Zone III (central canal).

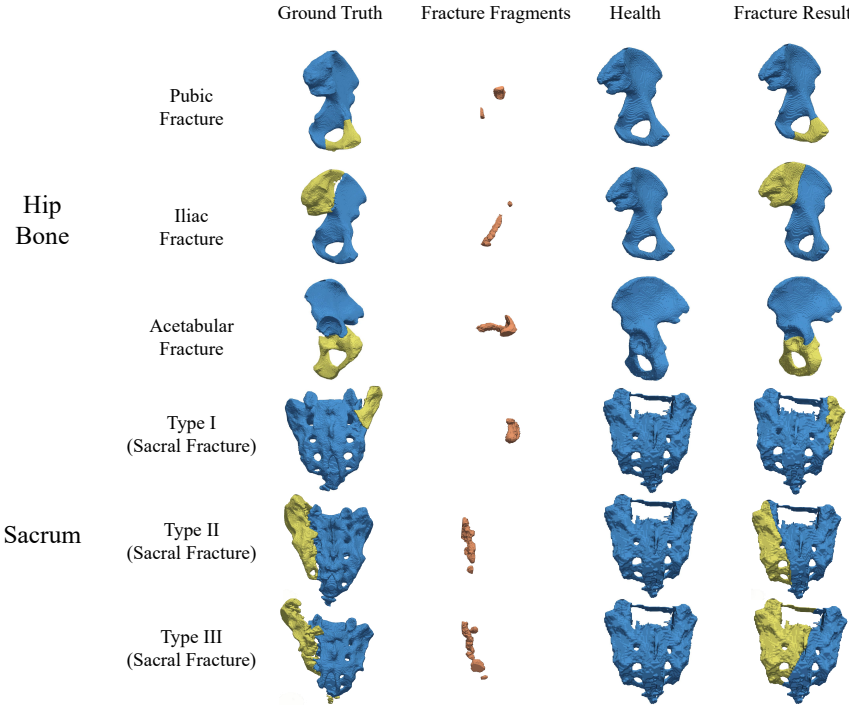

Figure 6: GFSM output across six fracture subtypes. Rows (top to bottom): pubic, iliac, acetabular (hip bone); Denis type-I, type-II, type-III (sacrum). Columns (left to right): real reference fracture bone, extracted fracture surface template S_{ref}, original healthy bone mask M_h, and simulated counterfactual fracture mask M_{frac}.

B Geometric Fracture Simulation Results

Following the AO/OTA classification (Meinberg et al., 2018) for hip bone and the Denis classification (Denis et al., 1988) for sacrum (see Appendix A for the anatomical reference), GFSM simulates six fracture subtypes: pubic, iliac, and acetabular fractures for the hip, and Denis type-I, II, and III for the sacrum.

Figure 6 shows the full pipeline output for each subtype: the real reference fracture bone (retrieved from the atlas), the extracted fracture surface template S_{ref}, the original healthy bone mask M_h, and the final simulated fracture mask M_{frac} after the Cut-and-Displace procedure (Eq. 9). The simulated masks qualitatively cover the intended anatomical sites from the pubic ramus to the sacral foramina and reproduce rigid fragment displacement and cortical separation consistent with the specified classes. In particular, the pubic and iliac subtypes exhibit characteristic linear separation, the acetabular subtype shows oblique splitting along the joint surface, and the Denis type-III sacral fracture places the fracture line through the intended central spinal canal region in this example. Although GFSM produces regular "smooth-cut" surfaces that lack the micro-irregular texture of real

fracture interfaces, the resulting topology provides sufficiently geometric guidance for MC-LDM to synthesize CT appearance, as evaluated by the synthesis metrics in Table 1.

C Detailed Formulation of Ghost Erasure

This appendix provides the complete mathematical formulation of the "Disguise–Erase–Restore" bone ghost cleaning algorithm summarized in Section 3.5.

Stage 1: Ghost region extraction and dilation. The ghost region to be erased is obtained via Boolean difference between the original healthy bone region Ω_{prior} and the target fracture region Ω_{frac}:

$$\Omega_{ghost} = \Omega_{prior} \setminus \Omega_{frac} = \{x \in \mathbb{R}^3 \mid x \in \Omega_{prior} \wedge x \notin \Omega_{frac}\}. \tag{12}$$

To completely wrap the cortical halo caused by the partial volume effect, the algorithm applies 3D morphological dilation with a spherical structural element B_r of radius r:

$$\Omega_{dilate} = \Omega_{ghost} \oplus B_r = \{x + b \mid x \in \Omega_{ghost},\, b \in B_r\}. \tag{13}$$

Stage 2: Workspace disguise. Before erasure, to prevent gradient contamination from bright bone pixels during inpainting, the algorithm adaptively extracts the gray-value median v_{bg} of the annular soft-tissue region surrounding the lesion, and temporarily "disguises" the target fracture region Ω_{frac} as this soft-tissue baseline to construct a working image:

$$X_{work}(x) = \begin{cases} v_{bg}, & x \in \Omega_{frac} \\ X_{origin}(x), & \text{otherwise.} \end{cases} \tag{14}$$

Stage 3: Inpainting and feathered restoration. The TELEA inpainting algorithm (Telea, 2004) is applied to Ω_{dilate} on X_{work}, propagating soft-tissue texture inward along iso-contour normals to produce X_{erased}. A 1D Gaussian filter G_{σ_z} is then applied along the z-axis to smooth inter-slice discontinuities. The final ghost-free background image is obtained by feathered fusion:

$$X_{base} = W_{replace} \odot (X_{erased} * G_{\sigma_z}) + (1 - W_{replace}) \odot X_{origin}, \tag{15}$$

where $W_{replace}$ is a feathering weight map generated by convolving the replacement mask $M_{replace} = \Omega_{dilate} \setminus \Omega_{frac}$ with a 2D Gaussian kernel $G_{\sigma_{xy}}$. This soft-boundary blending prevents visible discontinuities at the boundary between erased and original regions.

D Multi-Planar Analysis of Simulated Fracture Displacement

Figure 7 provides multi-planar verification of the rigid-body Euler transformation applied by GFSM (Eq. 9). We show one representative case before and after Cut-and-Displace in coronal, sagittal, and axial views. Across all three planes, the distal fragment exhibits coherent rotation and translation consistent with the parameters sampled from the clinical text prompt. In this example, no visually apparent overlap occurs between the proximal and distal fragments. The AABB intersection check and KD-tree retrieval define the separating surface, while the post-transform mask should be checked quantitatively for overlap and volume changes.

E LPIPS Distribution Comparison

Figure 8 shows the per-sample LPIPS distributions for our method and LDM across all three anatomical regions (left hip, sacrum, right hip). While the average LPIPS values of the two methods are equal at 0.2490 (with no statistically significant difference, $p > 0.6$ by two-sample t-test), the distribution analysis reveals a clear advantage of our method in generation *consistency*: across all regions, our method exhibits smaller interquartile range (IQR) and fewer high-error outliers. This advantage is most pronounced in the sacrum, where our IQR is approximately half that of LDM. The smaller IQR is consistent with more stable generation in this evaluation, but the distribution alone does not isolate which loss component causes the difference.

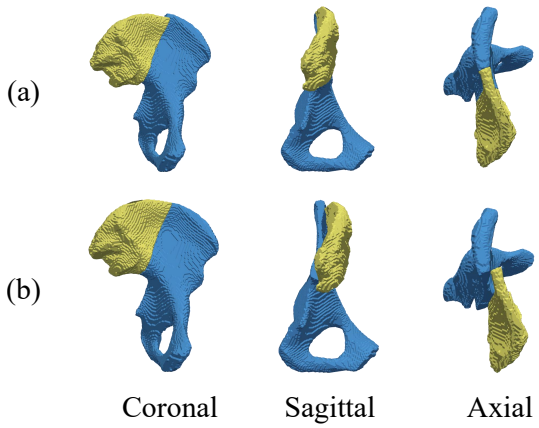

Figure 7: Multi-planar analysis of simulated fracture displacement. (a) Pre-transform healthy bone mask; (b) post-transform counterfactual fracture mask shown in coronal, sagittal, and axial views, illustrating fragment misalignment and rotation; quantitative overlap checks are not shown.

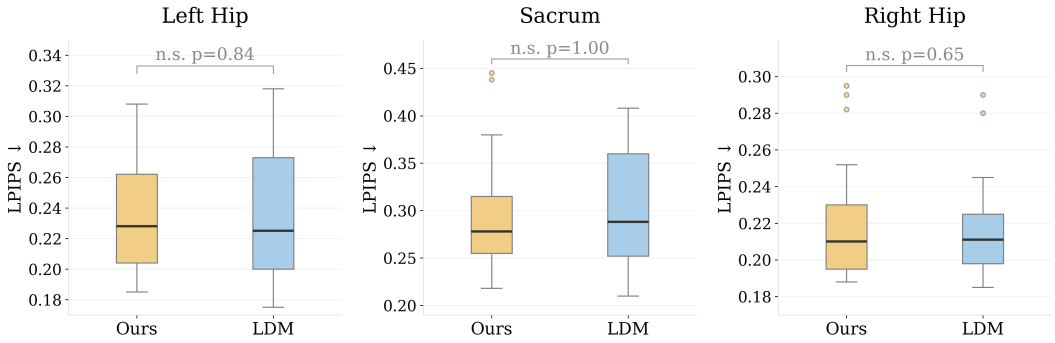

Figure 8: Per-sample LPIPS distributions for our method vs. LDM across left hip, sacrum, and right hip regions. Box boundaries indicate the interquartile range (IQR); whiskers extend to 1.5×IQR; dots denote outliers. Our method consistently achieves smaller IQR, indicating more stable generation quality, with the most notable improvement in the structurally complex sacrum region.

F Counterfactual Generation Across Fracture Subtypes

Figure 9 presents the full visualization of counterfactual fracture generation results across all six fracture subtypes: pubic, iliac, and acetabular fractures for the hip bone, and Denis type-I, II, and III fractures for the sacrum. For each subtype, we show the 3D fracture mask produced by GFSM alongside the generated counterfactual CT in coronal, sagittal, and axial views.

The generated fracture boundaries visually follow the GFSM masks across the six subtypes. Soft tissues and unaffected structures also remain visually consistent with the healthy source in these examples. The Denis type-III example illustrates how GFSM can specify a rare topology without requiring a new real fracture annotation; these visualizations do not establish clinical realism.

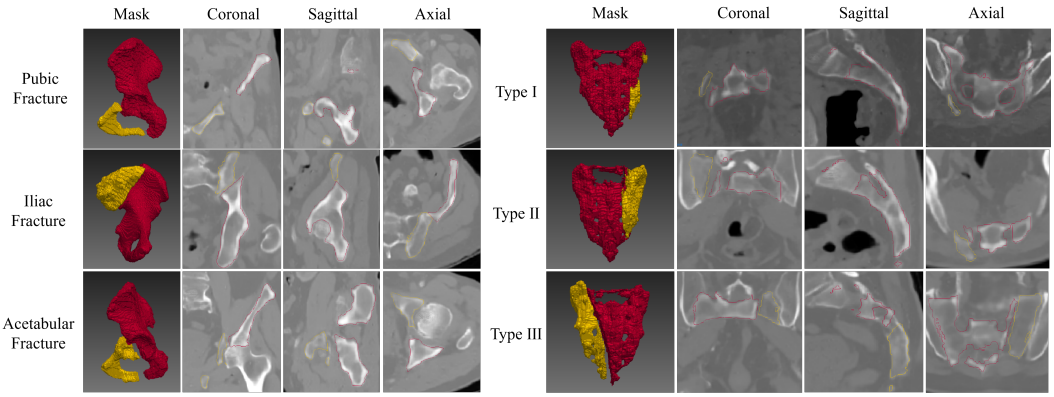

Figure 9: Counterfactual fracture generation results across all six fracture subtypes. (a) Hip bone: pubic (row 1), iliac (row 2), and acetabular (row 3) fractures. (b) Sacrum: Denis type-I (row 1), type-II (row 2), and type-III (row 3) fractures. For each subtype, the leftmost column shows the 3D GFSM fracture mask; the remaining columns show coronal, sagittal, and axial slice views of the corresponding generated counterfactual CT.