OsteoFlow: Lyapunov-Guided Flow Distillation for Predicting Bone Remodeling after Mandibular Reconstruction

Hamidreza Aftabi[1,4*], Faye Yu[2], Brooke Switzer[3], Zachary Fishman[4], Eitan Prisman[3], Antony Hodgson[2], Cari Whyne[4], Sidney Fels[1+], and Michael Hardisty[4+]

[1] Department of Electrical and Computer Engineering, University of British Columbia, Canada, *aftabi@student.ubc.ca
[2] Department of Mechanical Engineering, University of British Columbia, Canada
[3] Department of Surgery, University of British Columbia, Canada
[4] Sunnybrook Research Institute, University of Toronto, Canada

Abstract. Predicting long-term bone remodeling after mandibular reconstruction would be of great clinical benefit, yet standard generative models struggle to maintain trajectory-level consistency and anatomical fidelity over long horizons, particularly in low-data regimes. We introduce *OsteoFlow*, a flow-based framework predicting Year-1 post-operative CT scans from Day-5 scans. Our core contribution is Lyapunov-guided trajectory distillation: Unlike one-step distillation, our method distills a *continuous trajectory* over transport time from a registration-derived stationary velocity field teacher. Combined with a resection-aware image loss, this enforces geometric correspondence without sacrificing generative capacity. Evaluated on 344 paired regions of interest, *OsteoFlow* significantly outperforms state-of-the-art baselines, reducing mean absolute error in the surgical resection zone by ∼20%. This highlights the promise of trajectory distillation for long-term prediction in low-data clinical settings. Code is available on GitHub: *OsteoFlow*.

Keywords: Mandibular Reconstruction · Flow-based Modeling · Surgical Planning · Lyapunov-guided Distillation

1 Introduction

Mandibular reconstruction after segmental resection aims to restore jaw form and function in patients with oral and maxillofacial disease [1]. Despite advances in surgical planning and fixation, bone healing at the graft-host interface remains a major clinical challenge. Nonunion at the graft-host interface occurs up to 37% of cases in some sites [2], and can lead to pain, impaired function, hardware complications, and revision surgery. Although union is influenced by multiple

* Corresponding author.
+ Joint senior authors.

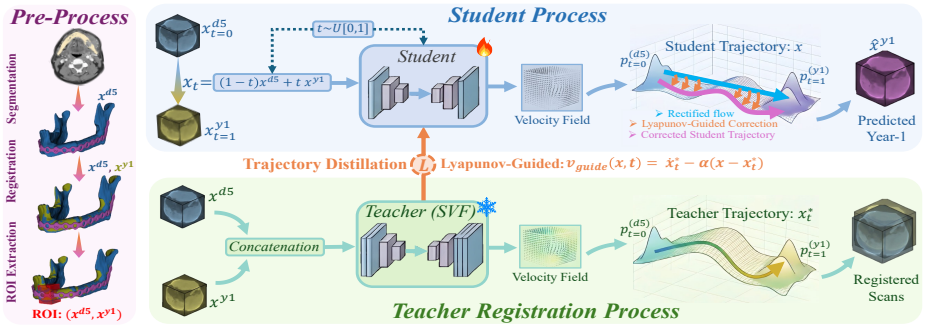

Fig. 1: Illustration of the preprocessing steps and the teacher–student Lyapunov-guided distillation framework for guiding the student velocity field.

patient and treatment-related factors (e.g., radiation exposure), previous studies suggest that local graft-host apposition (interface gap and contact geometry) is a key determinant of healing [3,4]. This motivates local image-based prediction at the graft-host interface, where remodeling most directly relates to union.

Prior studies on mandible reconstruction prediction have largely used physics-based and biomechanical simulations, including bone remodeling [5,6,7]. While mechanistically grounded and interpretable, these methods often require case-specific assumptions and parameterization, and can be computationally expensive at deployment. Even physics-informed machine learning models rely on extensive simulation guidance for bone remodeling, given the lack of a canonical governing PDE [8]. In parallel, generative AI, including diffusion-based models and recent flow-matching formulations, has shown strong performance across medical imaging modalities, including anomaly detection [9], modalities translation [10] and fusion [11], image synthesis [12,13], and reconstruction [14]. These advances motivate a data-driven approach for modeling post-operative anatomical change directly from imaging. However, most existing medical image generation methods are not designed for longitudinal remodeling prediction, where the goal is not only realistic appearance but biologically plausible evolution from an early post-operative scan to a later healing state. Some approaches also rely on segmentation/mask-based conditioning [15,16]. While these constraints can improve anatomically consistent structure, they provide limited supervision for the transport dynamics governing longitudinal anatomical evolution. Moreover, prior distillation strategies in generative modeling commonly focus on one-step teacher matching [17,12]. This approach is not well suited to longitudinal transport settings, where the key challenge is trajectory-level consistency over the full evolution, beyond ODE solver or sampler improvements.

In this work, we formulate mandibular bone remodeling prediction as conditional generative transport between paired Day-5 and Year-1 CT regions of interest (ROIs) centered at the resection interface (Figure 1). Our main contributions are as follows: (i) We develop a flow-based generative model, *OsteoFlow*, that predicts a time-dependent image-space velocity field and integrates it over

Algorithm 1 Flow-Based Trajectory Distillation with Lyapunov Guidance

Require: (x^{d5}, x^{y1}); student v_θ; teacher v_{SVF}; α; λ_{img}; $\lambda_{\max}$; Δt; τ_{HU}; σ

Training: repeat ▷ Optimize v_θ

for each iteration k (sample (x^{d5}, x^{y1}) and $t \sim \mathcal{U}(0,1)$) **do** ▷ One update

$p_{\mathrm{on}} \leftarrow \textsc{Ramp}(k)$, $\lambda_{\mathrm{Lyap}} \leftarrow \lambda_{\max} \cdot \textsc{Ramp}(k)$ ▷ Joint ramps

$x_t^{\mathrm{anchor}} \leftarrow (1-t)x^{d5} + t x^{y1}$ ▷ RF anchor (off-policy)

$\{x_{\mathrm{roll}}(s)\}_{s\in[0,1]} \leftarrow \textsc{Rollout}(v_\theta, x^{d5})$ ▷ Entire rollout

$(x_t, \hat{x}^{y1}) \leftarrow \begin{cases} (x_{\mathrm{roll}}(t),\ x_{\mathrm{roll}}(1)), & \text{with prob. } p_{\mathrm{on}} \\ (x_t^{\mathrm{anchor}},\ x_{\mathrm{roll}}(1)), & \text{otherwise} \end{cases}$ ▷ off-policy → on-policy

$x_t^\star \leftarrow \mathcal{W}\big(x^{d5}, \exp(t \cdot v_{\mathrm{SVF}})\big)$ ▷ Teacher warp

$x_{t+\Delta t}^\star \leftarrow \mathcal{W}\big(x^{d5}, \exp((t+\Delta t) \cdot v_{\mathrm{SVF}})\big)$ ▷ Next warp

$\dot{x}_t^\star \leftarrow (x_{t+\Delta t}^\star - x_t^\star)/\Delta t$ ▷ Teacher velocity (fwd diff.)

$v_{\mathrm{pred}} \leftarrow v_\theta(x_t, t; x^{d5})$ ▷ Student velocity

$\mathcal{L}_{\mathrm{RF}} \leftarrow \|v_{\mathrm{pred}} - (x^{y1} - x^{d5})\|^2$ ▷ RF supervision

$v_{\mathrm{guide}} \leftarrow \dot{x}_t^\star - \alpha(x_t - x_t^\star)$ ▷ Lyapunov guidance

$\mathcal{L}_{\mathrm{Lyap}} \leftarrow \|v_{\mathrm{pred}} - v_{\mathrm{guide}}\|^2$ ▷ Trajectory distillation

$M_{\mathrm{bone}}(\omega) \leftarrow \mathbf{1}\{x^{d5}(\omega) > \tau_{\mathrm{HU}}\}$ ▷ Bone mask

$W(\omega) \propto \exp\big(-d(\omega)^2/(2\sigma^2)\big)\, M_{\mathrm{bone}}(\omega)$ ▷ Resection-aware weight

$\mathcal{L}_{\mathrm{img}} \leftarrow \|W \odot (\hat{x}^{y1} - x^{y1})\|^2$ ▷ Bone-focused image loss

$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{RF}} + \lambda_{\mathrm{img}}\mathcal{L}_{\mathrm{img}} + \lambda_{\mathrm{Lyap}}\mathcal{L}_{\mathrm{Lyap}}$ ▷ Total loss

Update θ by minimizing $\mathcal{L}$ ▷ Gradient step

end for

Inference: integrate $\dot{x} = v_\theta(x, t; x^{d5})$ from $t = 0$ to $t = 1$ ▷ Teacher off

transport time to obtain the Year-1 prediction. (ii) To improve longitudinal transport learning, we cast teacher guidance as learning using privileged information (LUPI) [18], where a registration-derived teacher computed from both scans is used as a prior to supervise the student trajectory. (iii) We introduce an effective yet simple control-inspired, Lyapunov-based trajectory distillation that goes beyond one-step teacher matching by distilling a *continuous teacher trajectory* to improve rollout stability, anatomical coherence, and endpoint prediction. We further combine this trajectory-level supervision with rectified-flow supervision and a bone-focused image-space loss to improve endpoint fidelity near the union interface. Next, we describe our methodology in detail.

2 Methods

Overview. Let $(x^{d5}, x^{y1}) \sim p_{\mathrm{data}}$ denote paired Day-5 and Year-1 post-operative regions of interest (ROIs), and let $t \sim \mathcal{U}(0,1)$ denote the transport time. We model bone remodeling evolution by transporting x^{d5} to x^{y1} under a conditional velocity field $v_\theta(\cdot, t; x^{d5})$. The training objective has three terms: (i) a rectified-flow (RF) loss that provides paired supervision and learns the global remodeling transport direction, (ii) a Lyapunov-guided trajectory distillation loss that uses a registration-derived teacher trajectory as a LUPI-style prior to correct transport and improve endpoint accuracy, and (iii) a bone-focused image-space loss that improves endpoint fidelity in bone regions. Together, RF initializes and stabilizes transport, the teacher-guided term supervises the trajectory, and the image-space loss enforces Year-1 endpoint fidelity.

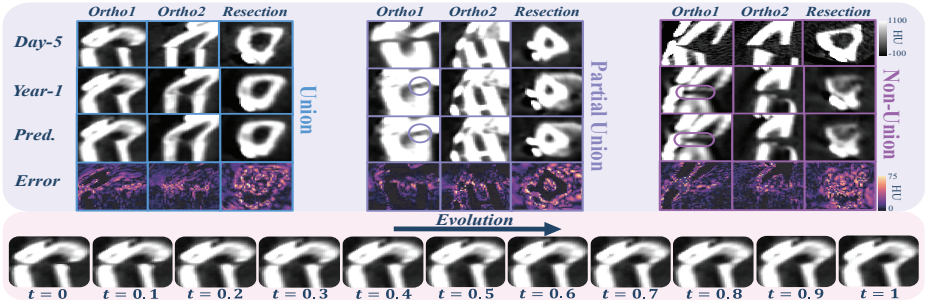

Fig. 2: Model predictions for three representative cases (union, partial union, and nonunion), shown on the resection plane and two other central orthogonal slices. The purple circle shows the partial union near the resection site, while the pink ellipsoid depicts the nonunion.

Rectified-flow supervision. Training starts from the RF anchor state $x_t = (1-t)x^{d5}+tx^{y1}$, which defines the training states and initializes the model toward the endpoint direction $(x^{y1} - x^{d5})$. The RF loss learns the conditional velocity $v_\theta(x_t, t; x^{d5})$ toward the endpoint [19]:

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{p_{\mathrm{data}},\, t}\left[\left\| v_\theta(x_t, t; x^{d5}) - (x^{y1} - x^{d5})\right\|_2^2\right]. \tag{1}$$

Lyapunov-guided trajectory distillation. During training, we gradually inject guidance from a teacher obtained via diffeomorphic stationary-velocity-field (SVF) [20] registration using privileged information from both scans; the teacher is used only during training and discarded at inference. It strictly provides deformation-based trajectory guidance as a *prior* during training and does not model actual osteogenesis. Unlike common one-step distillation in generative AI, which matches a single teacher update, our method distills a *continuous trajectory* over transport time. A simple baseline is mean-squared matching of student and teacher velocities, but this local objective does not enforce trajectory-level consistency in longitudinal transport. We therefore use a Lyapunov-guided formulation [21] that imposes principled reference-tracking dynamics: the teacher path defines the guidance trajectory, and the Lyapunov term provides corrective feedback to keep the student close when local predictions are imperfect. We denote the teacher trajectory by $x_t^\star$, constructed from a teacher SVF $v_{\mathrm{SVF}}(\xi)$, and approximate its image-space velocity as

$$x_t^\star = \mathcal{W}\big(x^{\mathrm{d5}},\, \exp(t \cdot v_{\mathrm{SVF}})\big) = x^{\mathrm{d5}} \circ \exp(t \cdot v_{\mathrm{SVF}})\,, \qquad \dot{x}_t^\star \approx \frac{x_{t+\Delta t}^\star - x_t^\star}{\Delta t}. \tag{2}$$

Since the SVF is defined in coordinate space (i.e., it parameterizes spatial deformation), it does not directly provide the image-space transport velocity needed for supervision. We therefore exponentiate the SVF using the scaling-and-squaring algorithm (warp operator $\mathcal{W}(\cdot,\cdot)$) at times t and $t + \Delta t$, then apply the finite-difference approximation above to obtain the image-space velocity $\dot{x}_t^\star$. To distill

this trajectory into the student dynamics, we define a quadratic Lyapunov energy centered on the teacher path,

$$V(x,t) = \frac{\alpha}{2}\,\|x - x_t^\star\|_2^2, \qquad \nabla_x V(x,t) = \alpha\,(x - x_t^\star). \tag{3}$$

This Lyapunov function is positive semidefinite (zero on the reference path and positive elsewhere), and the feedback is designed so that its derivative along the closed-loop dynamics is non-increasing ($\dot{V} \le 0$) [21]. Let u denote the control input. With running cost $\ell(x,t,u) = \frac{1}{2}\|u - \dot{x}_t^\star\|_2^2$, the induced guidance field is $v_{\text{guide}}(x,t) = \dot{x}_t^\star - \alpha\big(x - x_t^\star\big)$, where $\dot{x}_t^\star$ is the teacher trajectory signal and $-\alpha(x - x_t^\star)$ is corrective feedback toward the guidance path with stiffness $\alpha > 0$. Unlike critic-based methods (e.g., actor-critic), this guidance is derived in closed form from a Lyapunov energy function, without requiring a separate critic network. We then distill this guidance into the student velocity field through

$$\mathcal{L}_{\text{Lyap}} = \mathbb{E}_{p_{\text{data}},\,t}\Big[\big\|v_\theta(x_t,t;x^{d5}) - v_{\text{guide}}(x_t,t)\big\|_2^2\Big]. \tag{4}$$

Bone-focused image-space supervision. To improve endpoint fidelity in clinically relevant regions, we add a bone-focused image-space loss with a fixed, case-specific spatial weight map $W \in [0,1]$ that emphasizes bone voxels near the resection region. We define a Day-5 Hounsfield unit (HU)-derived bone mask as $M_{\text{bone}}(\omega) = \mathbf{1}\{x^{d5}(\omega) > \tau_{\text{HU}} = 300\}$. Let $d(\omega)$ denote the distance from voxel ω to the resection plane, and let $\hat{x}^{y1}$ denote the predicted Year-1 image obtained by integrating the learned ODE from x^{d5}. We then define the resection-aware weight and image-space loss as

$$W(\omega) \propto \exp\!\big(-d(\omega)^2/(2\sigma^2)\big)\,M_{\text{bone}}(\omega), \quad \mathcal{L}_{\text{img}} = \mathbb{E}_{p_{\text{data}}}\Big[\big\|W \odot \big(\hat{x}^{y1} - x^{y1}\big)\big\|_2^2\Big]. \tag{5}$$

Final objective. The student objective combines three losses:

$$\mathcal{L} = \mathcal{L}_{\text{RF}} + \lambda_{\text{img}}\,\mathcal{L}_{\text{img}} + \lambda_{\text{Lyap}}\,\mathcal{L}_{\text{Lyap}}, \tag{6}$$

where λ_{img} is fixed and λ_{Lyap} weights Lyapunov distillation. The teacher is trained only with $\mathcal{L}_{\text{img}}$ and then frozen for student training. Since $\mathcal{L}_{\text{RF}}$ and $\mathcal{L}_{\text{Lyap}}$ can impose competing velocity supervision, we use a joint curriculum rather than fixed weighting (Algorithm 1). Training starts with RF supervision on off-policy anchor states for stability, then linearly increases both the Lyapunov-distillation weight and the probability of on-policy sampling. On-policy losses are evaluated on rollout states, progressively shifting training toward inference-time dynamics.

3 Experiments and Results

3.1 Dataset and Augmentation Pipeline

We curated an internal longitudinal mandibular CT dataset from 120 patients with scans at approximately postoperative Day-5 (x^{d5}) and Year-1 (x^{y1}). The

mandible was segmented only to standardize landmarking. At each resection site, four fiducials were placed on x^{d5} (where osteotomy margins are clear), while x^{y1} often shows remodeling and bridging bone. The titanium reconstruction plate was segmented at both time points and used as a rigid reference to register x^{y1} to x^{d5}. A plane fit to each 4-fiducial group defined a local frame, and an oblique cubic ROI centered on the resection site was resampled to 48^3 voxels (0.5 mm isotropic), yielding 344 paired ROIs. All ROIs were windowed to $[-100, 1100]$ HU and affine-registered to suppress metal-driven outliers and focus on bone-remodeling-relevant intensities. The upper bound preserves cortical bone, and bridging callus near the resection typically falls below ∼1000 HU [22], leaving a 100 HU safety margin. To expand training data while preserving longitudinal correspondence, we used paired augmentation by applying the same transform to both ROIs in each (x^{d5}, x^{y1}) pair. Geometric changes were limited to medically plausible perturbations: left-right mirroring, small rotations (up to 10°), and translations (up to 2 voxels). Pose augmentation was implemented by transforming the ROI sampling frame and re-extracting from the original volumes. Intensity augmentation included additive Gaussian noise (10–25 HU), mild brightness/contrast shifts (about ±20–30 HU and ±3%), and light Gaussian blur ($\sigma = 0.8$) to mimic reconstruction-kernel variability. Each ROI pair yielded 41 variants in total, resulting in 14,104 paired samples. We evaluated the model using three random patient-level train/test splits (85% train, 15% test per split). Augmentation was applied only to the training set, while the test sets used only original ROIs, preventing data leakage.

3.2 Implementation Details and Metrics

The student is a time-conditioned 3D residual U-Net [23] that predicts an image-space velocity field on 48^3 ROIs (single-channel in/out). It uses two strided downsampling stages (48→24→12) with symmetric transposed-convolution upsampling, base width 48 channels (48/96/192 across scales), and residual blocks with $3 \times 3 \times 3$ convolutions, GroupNorm, and SiLU. Time is injected in every residual block via FiLM-style modulation (per-channel scale/shift) [24] after the first convolution, and the block output is added to the skip path. The teacher is an SVF 3D residual U-Net with the same backbone (base width 48; depth 48→24→12). It takes the concatenated pair (x^{d5}, x^{y1}) as a 2-channel input and outputs a 3-channel SVF. Teacher parameters are frozen during student training, and to prevent data leakage, the teacher is retrained separately for each random split. In Eq. (6), we set $\lambda_{\text{img}} = 0.2$ and linearly ramp λ_{Lyap} to $\lambda_{\max} = 1$ over the first 10 epochs. In Eq. (3), we set $\alpha = 1$ empirically and leave sensitivity analysis for future work. Training used batch size 1 and AdamW (10^{-4} learning rate, 10^{-5} weight decay) on a single A100 GPU (40 GB) for up to 50 epochs (∼ 600,000 steps). Early stopping used a patience of 10 epochs based on validation plateau. Flow-based inference used RK4 ODE solver with step size 0.1.

We compared *OsteoFlow* against baselines spanning direct (x^{d5}, x^{y1}) comparison, regression (ResUNet++ [25]), GANs (Pix2Pix [26], GRIT [27]), diffu-

Table 1: Baseline comparison (mean ± std across runs). Among non-teacher methods, the best and second-best results are shown in **bold** and underlined, respectively. An asterisk (*) indicates a significant difference from the second-best result using a two-sided Wilcoxon signed-rank test on paired ROI-level test metrics ($p < 0.05$).

Method	Mid slab				Full volume			
	Dice (%)↑	MS-SSIM (%)↑	MAE All HU↓	MAE Bone HU↓	Dice (%)↑	MS-SSIM (%)↑	MAE All HU↓	MAE Bone HU↓
Day5↔Year1	85.60 ±0.18	64.18 ±0.36	95.89 ±2.93	177.94 ±5.40	94.36 ±0.12	87.13 ±0.26	41.78 ±1.75	70.32 ±3.27
ResUNet++ [25]	92.65 ±0.11	78.10 ±0.22	56.85 ±1.64	102.67 ±4.20	96.57 ±0.08	91.82 ±0.17	27.76 ±1.12	45.80 ±2.50
Pix2Pix [26]	92.51 ±0.11	78.26 ±0.25	56.96 ±1.80	103.55 ±4.50	96.45 ±0.08	91.86 ±0.18	27.75 ±1.26	46.38 ±2.71
GRIT [27]	92.45 ±0.13	77.05 ±0.31	60.07 ±2.21	100.63 ±4.10	96.48 ±0.09	91.32 ±0.22	30.31 ±1.56	45.78 ±2.63
cDDPM(Δ) [28,29]	93.64 ±0.15	80.02 ±0.34	50.45 ±2.44	87.15 ±4.81	97.01 ±0.11	92.21 ±0.24	26.14 ±1.62	39.71 ±3.19
SegGuidedDiff [16]	93.03 ±0.14	78.00 ±0.29	54.10 ±2.10	95.58 ±4.60	96.93 ±0.10	91.57 ±0.23	27.40 ±1.40	43.06 ±2.88
MedVAE-1 [30]	88.86 ±0.19	63.94 ±0.44	81.76 ±3.10	145.39 ±5.90	92.41 ±0.14	73.66 ±0.35	57.99 ±2.22	96.00 ±3.80
MedVAE-2 [30]	89.63 ±0.17	67.56 ±0.39	73.78 ±2.74	107.29 ±5.40	92.80 ±0.12	76.67 ±0.31	51.12 ±2.00	66.98 ±3.50
RecFlow [19]	92.86 ±0.13	79.25 ±0.28	55.24 ±2.05	97.28 ±4.45	96.70 ±0.12	92.32 ±0.22	26.98 ±1.54	45.06 ±3.04
***OsteoFlow* (Ours)**	**94.55*** **±0.09**	**83.49*** **±0.21**	**44.12*** **±1.34**	**69.88*** **±3.75**	**97.23*** **±0.07**	**93.82*** **±0.15**	**22.88*** **±0.92**	**36.74*** **±2.30**
Teacher (LUPI)	96.87 ±0.07	90.33 ±0.16	37.17 ±1.09	40.63 ±3.28	98.45 ±0.05	96.55 ±0.12	19.31 ±0.76	18.43 ±1.98

sions (cDDPM(Δ) [28,29], SegGuidedDiff [16]), VAEs (MedVAE [30]), and flow (Rectified Flow [19]). Except for MedVAE, all baselines were matched to our backbone depth and capacity, differing only by method-specific architecture or losses. For MedVAE, we fine-tuned the released pretrained model with its original setup (L1 + LPIPS + PatchGAN; MedVAE-1); due to poor performance, we trained a second variant with the image-space loss in Eq. 5 (MedVAE-2). For cDDPM(Δ), following Jiang *et al.* [29], the model predicts the inter-scan intensity difference (Δ) conditioned on x^{d5}. In SegGuidedDiff, we conditioned the model on x^{d5} and the bony mask. We report Dice (bone, $HU > 300$), multi-scale structural similarity (MS-SSIM), and mean absolute error (MAE; all HU and bone-only HU) for the full ROI and a mid slab (central 12 slices around the resection plane).

3.3 Results and Ablation Studies

Table 1 summarizes the comparison between our method (*OsteoFlow*) and state-of-the-art baselines. *OsteoFlow* achieves the best overall performance with the most important gains in the mid-slab region, which contains the graft-host zone and the main remodeling changes. Relative to the second best method (cDDPM(Δ)), *OsteoFlow* improves mid-slab Dice from 93.64 to 94.55 (+0.91 points, ∼0.97%), improves MS-SSIM from 80.02 to 83.49 (+3.47 points, ∼4.3%), and reduces MAE (All HU) from 50.45 to 44.12 HU (−6.33 HU, ∼12.5%). The largest gain is in mid-slab bone MAE, which drops from 87.15 to 69.88 HU (−17.27 HU, ∼19.8%), indicating markedly better fidelity of bone remodeling

Table 2: Ablation study (mean ± std across runs). R = rectified flow; T_1 = our Lyapunov-guided distillation; T_2 = MSE-based distillation; A = augmentation; I = image-space loss; W = resection-aware weight. Best results are in **bold**.

	R	A	I	T_1	T_2	W	Mid slab				Full volume			
							Dice (%)↑	MS-SSIM (%)↑	MAE All HU↓	MAE Bone HU↓	Dice (%)↑	MS-SSIM (%)↑	MAE All HU↓	MAE Bone HU↓
Teacher Effect	✓	×	×	×	×	×	88.97 ±0.20	66.76 ±0.48	82.47 ±3.08	139.84 ±5.86	95.25 ±0.13	87.45 ±0.30	40.07 ±1.92	63.89 ±3.96
	×	×	×	✓	×	×	90.79 ±0.17	71.23 ±0.41	77.00 ±2.74	123.85 ±5.33	95.72 ±0.11	89.05 ±0.27	37.08 ±1.70	55.91 ±3.58
	✓	×	×	✓	×	×	**91.54 ±0.16**	**73.20 ±0.39**	**69.79 ±2.41**	**115.66 ±5.02**	**96.00 ±0.10**	**89.68 ±0.26**	**34.51 ±1.49**	**52.91 ±3.37**
	✓	×	×	×	✓	×	89.67 ±0.22	69.22 ±0.50	79.29 ±2.96	137.79 ±5.71	95.49 ±0.14	88.48 ±0.33	37.60 ±1.83	60.03 ±3.80
Full Pipeline	✓	✓	×	×	×	×	92.86 ±0.13	79.25 ±0.28	55.24 ±2.05	97.28 ±4.45	96.70 ±0.12	92.32 ±0.22	26.98 ±1.54	45.06 ±3.04
	✓	✓	✓	×	×	×	93.65 ±0.12	81.39 ±0.25	49.47 ±1.74	83.65 ±4.06	97.00 ±0.10	93.09 ±0.19	24.95 ±1.26	40.02 ±2.92
	✓	✓	✓	✓	×	×	94.46 ±0.10	83.12 ±0.22	45.37 ±1.41	71.85 ±3.79	97.21 ±0.08	93.73 ±0.17	23.10 ±1.02	37.12 ±2.49
	✓	✓	✓	✓	×	✓	**94.55 ±0.09**	**83.49 ±0.21**	**44.12 ±1.34**	**69.88 ±3.75**	**97.23 ±0.07**	**93.82 ±0.15**	**22.88 ±0.92**	**36.74 ±2.30**

Table 3: Backbone ablation study (mean ± std across runs). A = augmentation; I = image-space loss. Both backbones have ∼12M parameters and the same network depth for a fair comparison. Best results are in **bold**.

Backbone	A	I	Mid slab				Full volume			
			Dice (%)↑	MS-SSIM (%)↑	MAE All HU↓	MAE Bone HU↓	Dice (%)↑	MS-SSIM (%)↑	MAE All HU↓	MAE Bone HU↓
SwinUNETR	✓	×	92.30 ±0.14	77.79 ±0.30	58.83 ±1.86	103.92 ±4.62	96.58 ±0.10	91.44 ±0.21	28.28 ±1.30	46.82 ±2.82
Res-UNet	✓	×	**92.86 ±0.13**	**79.25 ±0.28**	**55.24 ±2.05**	**97.28 ±4.45**	**96.70 ±0.12**	**92.32 ±0.22**	**26.98 ±1.54**	**45.06 ±3.04**
SwinUNETR	✓	✓	93.27 ±0.13	80.24 ±0.27	51.95 ±1.78	89.21 ±4.36	96.86 ±0.10	92.74 ±0.20	25.49 ±1.22	41.24 ±2.94
Res-UNet	✓	✓	**93.65 ±0.12**	**81.39 ±0.25**	**49.47 ±1.74**	**83.65 ±4.06**	**97.00 ±0.10**	**93.09 ±0.19**	**24.95 ±1.26**	**40.02 ±2.92**

near the union site. These gains are notable because they occur in a high-performance regime (e.g., mid-slab Dice > 93%), where further improvements are typically harder to obtain. As expected, the teacher, constructed using privileged information from both scans, acts as a structural upper bound. While the teacher is limited to spatial deformation for guiding the student's trajectory and cannot predict true osteogenesis, *OsteoFlow* narrows this performance gap. Figure 2 also shows plausible union, nonunion, and partial union prediction, rather than smooth, averaged outputs, despite union being the dominant class.

Ablation studies in Tables 2 and 3 show clear and consistent contributions from the proposed components, with the largest gains in the clinically relevant mid-slab (resection/union) region. In Table 2, the proposed Lyapunov-guided trajectory distillation (T_1) improves performance over the rectified-flow-only setting and clearly outperforms the MSE-based teacher distillation variant (T_2), which indicates that the gain comes from the guidance formulation, not from teacher supervision alone. The full-pipeline ablation also shows a steady improvement as augmentation (A), image-space supervision (I), trajectory distillation

(T_1), and resection weighting (W) are added, with the largest effect on bone-focused error near the union site (e.g., mid-slab bone MAE decreases from 97.28 to 69.88 HU). Table 3 further confirms that these gains do not depend on model size: with matched depth and parameter count ($\sim$12M), Res-UNet outperforms SwinUNETR [31] both with and without I, which justifies our backbone choice.

4 Conclusion

This work presents a trajectory-distillation framework for bone remodeling prediction in a low-data, class-imbalanced setting. The model leverages privileged training information from a registration-derived teacher with Lyapunov-guided trajectory supervision. Our proposed approach differs in nature from one-step distillation or mask-based anatomical supervision, with guidance applied over the entire evolution trajectory to promote biologically plausible outputs. Despite the difficulty of predicting long-term graft-host healing from early post-operative geometry alone, the model generated realistic union, nonunion, and partial-union patterns rather than smooth, averaged outputs. These findings suggest that geometry-centered trajectory modeling can capture clinically meaningful remodeling structure even when outcome determinants are only partially observed. Key limitations of our work include the absence of important patient-level covariates (e.g., radiation exposure [4]), incorporation of exact scan time intervals, and the need for external validation beyond our internal dataset.

Acknowledgments. We gratefully acknowledge financial support from the Terry Fox Research Institute and the UBC Friedman Award for Scholars in Health. We also thank the UBC ISTAR Group and the Holland Bone and Joint Research Team at Sunnybrook.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Hamidreza Aftabi, Katrina Zaraska, Atabak Eghbal, Sophie McGregor, Eitan Prisman, Antony Hodgson, and Sidney Fels. Computational models and their applications in biomechanical analysis of mandibular reconstruction surgery. *Comput. Biol. Med.*, 169:107887, 2024.
2. Brian Swendseid, Ayan Kumar, Larissa Sweeny, Tingting Zhan, Richard A. Goldman, Howard Krein, Ryan N. Heffelfinger, Adam J. Luginbuhl, and Joseph M. Curry. Natural history and consequences of nonunion in mandibular and maxillary free flaps. *Otolaryngol. Head Neck Surg.*, 163(5):956–962, 2020.
3. Hamidreza Aftabi, John E. Lloyd, Amanda Ding, Benedikt Sagl, Eitan Prisman, Antony Hodgson, and Sidney Fels. Osteoopt: A bayesian optimization framework for enhancing bone union likelihood in mandibular reconstruction surgery. In *MICCAI*, pages 448–458, 2025.
4. Farahna Sabiq, Abhiram Cherukupalli, Mohammad Khalil, Linh K. Tran, Jamie JY Kwon, Thomas Milner, James S. Durham, and Eitan Prisman. Evaluating the benefit of virtual surgical planning on bony union rates in head and neck reconstructive surgery. *Head Neck*, 46(6):1322–1330, 2024.

5. Hamidreza Aftabi, Benedikt Sagl, John E. Lloyd, Eitan Prisman, Antony Hodgson, and Sidney Fels. To what extent can mastication functionality be restored following mandibular reconstruction surgery? a computer modeling approach. *Comput. Methods Programs Biomed.*, 250:108174, 2024.
6. Hamidreza Aftabi, John E. Lloyd, Benedikt Sagl, Amanda Ding, Eitan Prisman, Antony Hodgson, and Sidney Fels. Optimizing bone cuts enhances predicted bone union propensity in mandibular body reconstruction. In *ISBI*, pages 1–4, 2025.
7. Hamidreza Aftabi, John E Lloyd, Amanda Ding, Benedikt Sagl, Eitan Prisman, Antony Hodgson, and Sidney Fels. Patient-specific optimization for mandibular reconstruction planning with enhanced bone union. *arXiv preprint arXiv:2605.01084*, 2026.
8. Chi Wu, Ali Entezari, Keke Zheng, Jianguang Fang, Hala Zreiqat, Grant P Steven, Michael V Swain, and Qing Li. A machine learning-based multiscale model to predict bone formation in scaffolds. *Nat. Comput. Sci.*, 1:532–541, 2021.
9. Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *CVPR*, pages 650–656, 2022.
10. Zhaohu Xing, Sicheng Yang, Sixiang Chen, Tian Ye, Yijun Yang, Jing Qin, and Lei Zhu. Cross-conditioned diffusion model for medical image-to-image translation. In *MICCAI*, pages 201–211, 2024.
11. Meng Zhou and Farzad Khalvati. Clinicalfmamba: Advancing clinical assessment using mamba-based multimodal neuroimaging fusion. In *MLMI*, pages 245–255, 2025.
12. Heng Chang, Yu Shang, Haifeng Wang, Yuxia Liang, Haoyu Wang, Fan Wang, Chen Niu, and Chunfeng Lian. Controllable flow matching for 3d contrast-enhanced brain MRI synthesis from non-contrast scans. In *MICCAI*, pages 119–128, 2025.
13. Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical AI for synthetic imaging. In *WACV*, pages 4430–4441, 2025.
14. Yiran Sun, Hana Baroudi, Tucker Netherton, Laurence Court, Osama Mawlawi, Ashok Veeraraghavan, and Guha Balakrishnan. Difr3ct: Latent diffusion for probabilistic 3d CT reconstruction from few planar x-rays. *arXiv preprint arXiv:2408.15118*, 2024.
15. Yuetan Chu, Changchun Yang, Gongning Luo, Zhaowen Qiu, and Xin Gao. Anatomic-constrained medical image synthesis via physiological density sampling. In *MICCAI*, pages 69–79, 2024.
16. Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *MICCAI*, pages 88–98, 2024.
17. Huaisheng Zhu, Teng Xiao, Shijie Zhou, Zhimeng Guo, Hangfan Zhang, Siyuan Xu, and Vasant G. Honavar. Simple distillation for one-step diffusion models. In *NeurIPS*, 2025.
18. Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Netw.*, 22(5–6):544–557, 2009.
19. Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
20. Akshay Pai, Stefan Sommer, Lauge Sørensen, Sune Darkner, Jon Sporring, and Mads Nielsen. Kernel bundle diffeomorphic image registration using stationary ve-

locity fields and wendland basis functions. *IEEE Trans. Med. Imaging*, 35(6):1369–1380, 2015.
21. Hassan K. Khalil and Jessy W. Grizzle. *Nonlinear Systems*, volume 3. Prentice Hall, 2002.
22. Abhay Datarkar, Bhavana Valvi, Suraj Parmar, and Jagadish Patil. Qualitative assessment of newly formed bone after distraction osteogenesis of mandible in patients with facial asymmetry using 3-dimensional computed tomography. *J. Oral Biol. Craniofac. Res.*, 11(3):410–414, 2021.
23. Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432, 2016.
24. Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, volume 32, 2018.
25. Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *ISM*, pages 225–2255, 2019.
26. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.
27. Saksham Suri, Moustafa Meshry, Larry S. Davis, and Abhinav Shrivastava. Grit: Gan residuals for paired image-to-image translation. In *WACV*, pages 4965–4975, 2024.
28. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.*, 33:6840–6851, 2020.
29. Caiwen Jiang, Xiaodan Xing, Zaixin Ou, Mianxin Liu, Walsh Simon, Guang Yang, and Dinggang Shen. CIResDiff: A clinically-informed residual diffusion model for predicting idiopathic pulmonary fibrosis progression. In *MLMI*, pages 83–93, 2024.
30. Maya Varma, Ashwin Kumar, Rogier Van der Sluijs, Sophie Ostmeier, Louis Blankemeier, Pierre Chambon, Christian Bluethgen, Jip Prince, Curtis Langlotz, and Akshay Chaudhari. Medvae: Efficient automated interpretation of medical images with large-scale generalizable autoencoders. *arXiv preprint arXiv:2502.14753*, 2025.
31. Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *MICCAI BrainLesion Workshop*, pages 272–284, 2021.