Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Preference Guided CT Projection Correction for Surface-Based Feather Morphometrics

Tyler Thompson

tyler.n.thompson@utah.edu

University of Utah

Edward Hsu

University of Utah

Research Article

Keywords: computed tomography, motion artifact, morphometrics, biological imaging, feather morphology, task-based image quality, preference learning, residual autoencoder, surface analysis

Posted Date: July 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9962725/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Bioinformatics on October 28th, 2026. See the published version at https://doi.org/10.1186/s12859-026-06716-5.

Preference-guided CT projection correction for surface-based feather morphometrics

Tyler Thompson[1] and Edward Hsu[1]

[1]Department of Biomedical Engineering, University of Utah, Salt Lake City, UT, USA

Corresponding author: *tyler.n.thompson@utah.edu*

Abstract

Background: Traditional imaging pipelines often optimize for human-readable images and generic quality metrics, but algorithm-facing morphometric workflows require preservation of the information used by downstream measurement. Acquisitions that appear overexposed or visually nonstandard may therefore be appropriate when they expose subtle structures to an analysis algorithm. We developed a task-aligned projection-correction framework for reducing morphology-relevant CT motion artifacts without relying on human visual interpretability as the final quality standard.

Results: Using CT-based feather morphometry as a motivating example, we treated a previously published feather-growth computer-vision workflow as the downstream task to be protected during correction. A residual denoising autoencoder defined a bounded projection-correction space, and an expert human-preference ranker scored candidate residual corrections for preservation of task-relevant morphology. In five independently seeded synthetic scenes generated on a common body-derived pigeon model, rank-weighted evolution-strategy updates optimized latent corrections under projection-preservation, latent-bound, and angular-smoothing constraints. Preference-guided correction reduced reconstruction RMSE by 68.0% and density-map RMSE by 72.0% relative to motion-corrupted data, outperforming visual denoising and random-latent ablation baselines. Clean-data controls produced only small perturbations when no motion correction was needed. In two real-data case studies, the same framework reduced visible feather ghosting and improved downstream crest-density mapping and agreement with manual reference masks when inserted upstream of the published morphometric workflow.

Conclusions: These results support task-aligned projection correction as a route toward CT imaging workflows optimized for downstream biological measurement rather than generic visual image quality alone. The study provides proof-of-concept evidence that expert preference signals can guide bounded projection-domain correction for surface-based biological morphometrics.

Keywords: computed tomography; motion artifact; morphometrics; biological imaging; feather morphology; task-based image quality; preference learning; residual autoencoder; surface analysis

1 Background

Quantitative biological imaging increasingly depends on measurements extracted from volumetric imaging rather than on images as final products. In many CT-based workflows, a reconstructed volume is segmented, converted into a surface representation, and then analyzed to quantify shape, count repeated structures, or measure spatial distributions of anatomically meaningful features. In this setting, artifacts can induce structured geometric errors that propagate nonlinearly through segmentation, surface extraction, and downstream quantitative analysis. The resulting bias may appear not simply as a visually degraded reconstruction, but as an incorrect morphology-derived measurement.

Motion artifacts in CT can arise from many time-varying sources, including specimen displacement, stage or gantry vibration, center-of-rotation instability, imperfect mechanical repeatability, or motion of only part of the specimen during detector integration. The distinction between visual image quality and downstream measurement validity is especially important for thin, weakly attenuating, or densely repeated structures. Feathers are a representative example. In CT scans of feathered specimens, small within-view motions of low-opacity feathers can create edge duplication, ghosted boundaries, or local blur in projection data even when the larger body remains approximately fixed. After reconstruction, those artifacts can appear as duplicated ridges or feather-like structures on the extracted surface. A morphology pipeline can then interpret these artifacts as real structures, producing biased counts, density maps, or region assignments.

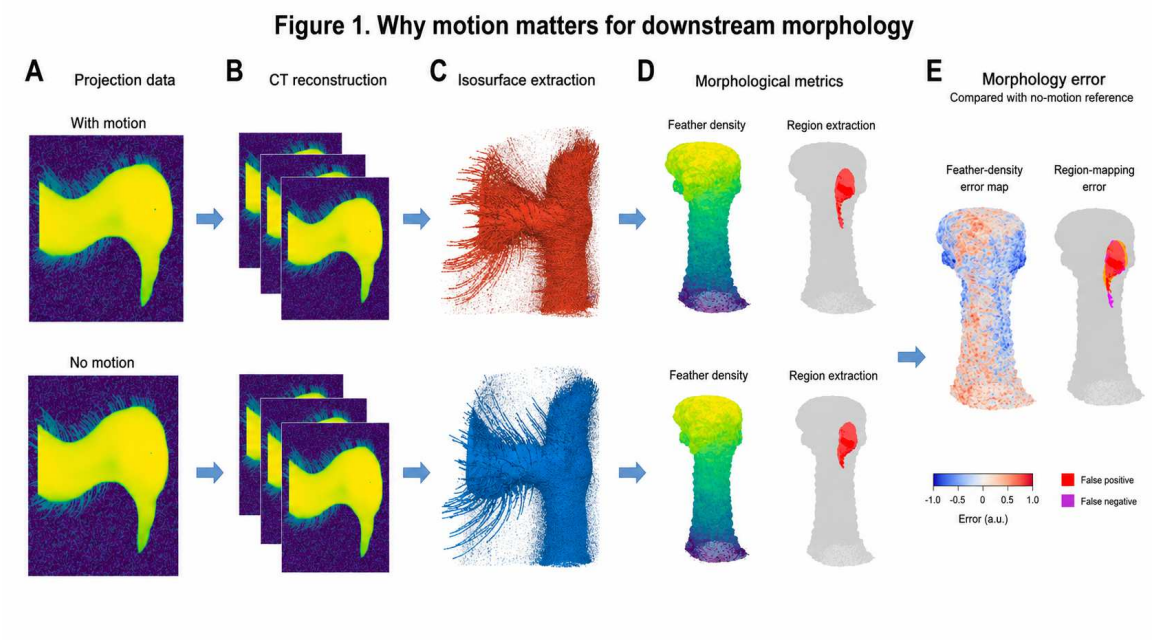


Figure 1: Motivating failure mode for task-aligned CT artifact evaluation. Projection-domain feather motion can produce blur and ghosting that may be subtle in individual images but persist after reconstruction, alter the extracted isosurface, and bias morphology-level outputs such as feather-density maps and anatomical region assignments.

Figure 1 is intended as a conceptual overview rather than the primary quantitative evidence. To a human observer, some of the projection-level changes can appear modest because the acquisition is already tuned for feather visibility rather than conventional soft-tissue appearance. The downstream effect is larger: small feather-edge changes can propagate through reconstruction into duplicated surface ridges, altered density maps, and region-mapping errors. Figure 3 shows this failure mode with higher-contrast examples and quantitative error summaries.

Conventional CT correction is often evaluated using image-domain or perceptual criteria such as sharpness, smoothness, noise reduction, or visual fidelity. Those criteria are useful, but they do not necessarily predict whether a downstream surface or morphometric analysis is correct. Task-based image-quality work has shown that image evaluation should be tied to the detection or measurement task of interest [1, 2, 3]. Likewise, motion-correction methods for CT and related tomographic modalities must be judged not only by appearance, but also by how much they preserve or improve quantitative interpretation [4]. Similar concerns arise in other settings where artifact correction is evaluated against downstream structure measurements, such as motion-compensated micro-CT reconstruction for dynamic materials imaging [5], deep learning and model-based CT reconstruction

[6], projection- and image-domain CT artifact reduction [7], and morphology-based assessment of vessel correction in cone-beam CT [8]. For surface-based biological morphometrics, the relevant endpoint is whether the surface, component counts, and density maps better reflect the specimen.

The present framework was developed around this principle. Rather than treating artifact correction as a purely aesthetic image-restoration problem, the method learns a residual autoencoder latent space for projection-domain correction and uses a human-preference pairwise ranking model to judge which decoded candidate correction is more likely to support accurate downstream morphology. Because each candidate is generated by a known latent perturbation, the ranking signal can be converted into an evolution-strategy update in the learned latent space. This allows the correction step to be optimized without differentiating through CT reconstruction, isosurface extraction, or morphology-map construction.

2 Methods

2.1 Study design

The study was designed as a proof-of-concept benchmark for evaluating CT artifacts in biological surface-morphometry pipelines. The primary objective was to reproduce and correct a practical failure mode observed in pigeon feather CT analysis: motion during acquisition can create duplicated or distorted surface features that bias feather-density and region-mapping outputs. The framework includes six stages: volume and surface preparation, attenuation-volume synthesis, feather placement and motion modeling, projection simulation, reconstruction, and downstream morphology evaluation.

The main synthetic evidence package uses five independently seeded simulated scenes generated on the same body-derived pigeon surface model. Each scene was evaluated under matched condition labels: a no-motion reference, a motion-corrupted condition, a simple visual/image-processing baseline, a no-ranker random-latent ablation, the preference-guided correction condition, and a no-motion + correction limited-perturbation control. The five scenes should therefore be interpreted as independent stochastic placements and motion realizations rather than independent biological specimens. To test whether the correction strategy transferred to the motivating biological application, the method was also applied to two real pigeon CT projection datasets as pilot case studies

upstream of an existing feather-growth computer-vision pipeline.

Table 1 defines the main synthetic conditions. In this manuscript, "clean" and "no-motion" denote absence of within-view feather motion, not absence of the matched detector model. For each seed, the no-motion reference and the motion-corrupted condition used the same scene geometry, fixed-pattern gain map, flat-field mismatch map, stripe-offset vector, and detector-artifact parameter values; the stochastic generator was seeded deterministically by seed and angle so that the paired Poisson sampling stream was matched across no-motion and motion conditions wherever Poisson sampling was used. Thus paired differences are dominated by feather motion under the stated detector model rather than by independent detector draws. Values are reported with units throughout: projection RMSE/MAE are dimensionless minus-log line-integral errors; reconstruction RMSE/MAE are attenuation errors in 1/mm; density-map errors are dimensionless because each density map was normalized independently to the range 0–1 on the analysis mask. Consequently, density-map RMSE and correlation primarily measure spatial-pattern mismatch after normalization rather than absolute feather-density magnitude; absolute morphology burden is captured separately by surface-area and component-count errors.

Table 1: Synthetic benchmark condition definitions.

Condition	Motion	Detector model
No-motion reference	disabled ($A_i = 0$)	matched detector-artifact/noise realization
Motion-corrupted	severe feather flutter ($A_i = 3°$)	same detector model as reference
Visual baseline	motion-corrupted input	median-stripe suppression + Gaussian denoising
Random latent	motion-corrupted input	residual autoencoder latent search with shuffled preferences
Preference-guided	motion-corrupted input	residual autoencoder latent search with human-preference ranker
No-motion + correction	no-motion input	same correction procedure

2.2 Coordinate conventions and data products

All volumetric data use axis order (z, y, x) and all world coordinates use order (x, y, z). Voxel spacing is isotropic and recorded as `voxel_size_mm`. Grid metadata include `origin_mm = [x0, y0, z0]`, the world coordinate of the center of voxel $(z = 0, y = 0, x = 0)$. Voxel centers are

therefore computed as

$$x = x_0 + x_{\mathrm{idx}} s, \tag{1}$$

$$y = y_0 + y_{\mathrm{idx}} s, \tag{2}$$

$$z = z_0 + z_{\mathrm{idx}} s, \tag{3}$$

where s is the isotropic voxel size in mm.

The physical attenuation field is represented as $\mu(\mathbf{r})$ with units of 1/mm. Projection line integrals are dimensionless:

$$A = \int \mu(\mathbf{r})\, d\ell. \tag{4}$$

Projection data are stored as line integrals A, normalized intensities I/I_0, and minus-log measurements $-\log(I/I_0)$. HDF5 files store arrays together with JSON metadata describing grid definitions, axis ordering, units, random seeds, motion parameters, and provenance.

2.3 Synthetic pigeon-surface model

The primary benchmark was built from an existing body-derived pigeon attenuation volume paired with a surface mesh containing a manually painted feather-placement field. The base volume and mesh were derived from previously acquired pigeon CT project data and were used here only as a geometric phantom for simulation; no new animal imaging was performed for this study. The body volume provided the solid anatomical support for CT projection and reconstruction, while the painted surface mesh defined where synthetic feathers could be placed. This design allowed the simulation to reproduce the relevant downstream failure mode—motion of thin surface appendages on a curved biological surface—without requiring the full complexity of real feather microstructure.

The final benchmark used five independently generated scenes with random seeds 0–4. Each scene used a $384 \times 256 \times 256$ volume with 0.25 mm isotropic voxels, corresponding to a $96.0 \times 64.0 \times 64.0$ mm physical field of view. The base body attenuation was set to the median nonzero body value, $\mu_{\mathrm{body}} = 0.200$ 1/mm. Feather attenuation was set to $\mu_{\mathrm{feather}} = 0.070$ 1/mm, equal to 35% of the body attenuation, and outside attenuation was zero.

The surface mesh contained a scalar field, `feather_weight`, that acted as a spatial prior for

feather placement. This field was manually painted on the mesh to mark biologically plausible feather-bearing regions; non-feathered regions such as the beak were assigned weights near zero and were therefore excluded by the sampling threshold. The painted field and mesh are included in the source data release so that the anchor sampling prior can be reused or replaced. Mesh triangles were first assigned a face-level feather weight equal to the mean of their vertex weights. Faces with mean weight below a threshold were excluded from sampling. For the benchmark, the threshold was

$$w_{\min} = 0.25.$$

For each remaining triangle f, the sampling probability was proportional to both triangle area and painted feather weight:

$$p(f) = \frac{A_f \max(\bar{w}_f - w_{\min}, 0)}{\sum_g A_g \max(\bar{w}_g - w_{\min}, 0)},$$

where A_f is triangle area and $\bar{w}_f$ is the mean painted feather weight on triangle f. Feather anchor points were then sampled uniformly within selected triangles using barycentric coordinates. This procedure concentrates feathers in the painted biological region while preserving random spatial variation and exact reproducibility through fixed seeds.

Local surface normals were computed from the surface mesh and smoothed before use. Smoothing was used to avoid abrupt changes in feather orientation caused by triangle-level mesh noise. For each anchor point, the local normal was estimated from nearby surface normals using distance-weighted averaging within a 2.5 mm neighborhood and weights proportional to $\exp[-d^2/(2\sigma_n^2)]$ with $\sigma_n = 1.0$ mm. These smoothed normals defined the outward direction from the body surface and were also used to initialize feather bases slightly above the surface so that rasterized feathers did not begin inside the body volume.

2.4 Feather geometry, density maps, and polarity regions

Each synthetic feather was represented as a curved chain of tapered frustum segments. The final benchmark used 300 feathers per scene and five tapered segments per feather. Feather length and radius varied systematically along the head–tail axis and also included per-feather random variation. Let $s_i \in [0, 1]$ denote the normalized head–tail position of feather i, with $s_i = 0$ near the head and $s_i = 1$ near the tail. In the benchmark this coordinate was computed from the anchor

position along the surface-aligned longitudinal axis. Feather length was assigned as

$$L_i = \left[L_{\text{head}} + (L_{\text{tail}} - L_{\text{head}}) s_i^{1.2}\right] \eta_{L,i},$$

where $L_{\text{head}} = 5.0$ mm, $L_{\text{tail}} = 13.0$ mm, and $\eta_{L,i} \sim U(0.75, 1.15)$. The feather base radius was

$$r_{0,i} = \left[R_{\text{head}} + (R_{\text{tail}} - R_{\text{head}}) s_i\right] \eta_{r,i},$$

where $R_{\text{head}} = 0.25$ mm, $R_{\text{tail}} = 0.37$ mm, and $\eta_{r,i} \sim U(0.8, 1.2)$. The terminal radius was

$$r_{T,i} = \gamma_i r_{0,i}, \qquad \gamma_i \sim U(0.25, 0.55).$$

Radii were linearly interpolated along the segment chain, producing tapered cylindrical/frustum-like surface appendages. During rasterization, feathers were first offset outward from the surface by 0.35 mm along the smoothed normal. Voxels were assigned by binary occupancy of the tapered frustum segments, with supersampling at two sub-voxel locations along each axis for anti-aliasing. Feather voxels were allowed to overwrite only air/background or pre-existing feather labels; body voxels were not eroded by lower-attenuation feather labels. This preserved the body attenuation while making surface appendages visible as distinct low-attenuation structures.

The anchor set defined the underlying synthetic feather placement for each scene. For downstream evaluation, however, feather-density maps were computed from reconstructed and extracted surface representations rather than directly from the known anchors. Surface-derived feather-density estimates were projected to a body-surface parameter grid and smoothed with a 2.0 mm Gaussian kernel. Each motion-corrupted or corrected condition was compared with the corresponding no-motion reference density map. This provided a morphology-level endpoint that was sensitive to duplicated or displaced feather structures rather than only to voxelwise reconstruction error.

The simulation framework also included optional regional polarity changes for generating morphology and region-mapping demonstration scenes. This capability was not used as a primary quantitative endpoint in the correction benchmark, but it documents how the same surface model can generate regular and inverted feather-growth regions. A connected region was selected within the painted feather-weight mask by choosing a seed point on the painted surface and expanding

an irregular connected patch over the mesh. The expansion used graph distances on the triangulated surface, with correlated random perturbations applied to edge costs to create an organic, noncircular boundary. The resulting binary surface field,

$$I(\mathbf{p}) \in \{0, 1\},$$

was stored as an `invert_region` scalar on the surface. Feather anchors inherited an inverted-growth flag from the nearest surface vertex:

$$I_i = I(\mathbf{a}_i),$$

where $\mathbf{a}_i$ is the anchor point of feather i.

When polarity regions were enabled, the rest centerline direction was constructed as follows. For regular-polarity feathers, the distal growth direction was biased toward the tail direction $\mathbf{t}$. For inverted-polarity feathers, the distal growth direction was reversed toward the head direction $-\mathbf{t}$. Thus the polarity-dependent growth direction was

$$\mathbf{g}_i = \begin{cases} \mathbf{t}, & I_i = 0, \\ -\mathbf{t}, & I_i = 1. \end{cases}$$

The centerline direction of segment j for feather i was then constructed by blending the smoothed surface normal $\mathbf{n}_i$ with the polarity-dependent growth direction:

$$\mathbf{d}_{ij} = \frac{(1 - \omega_j)\mathbf{n}_i + \omega_j \mathbf{g}_i}{\|(1 - \omega_j)\mathbf{n}_i + \omega_j \mathbf{g}_i\|},$$

where ω_j increases along the feather so that the base emerges approximately normal to the body and the distal end bends toward the growth-polarity direction. This construction allowed the same surface model to generate normal and inverted feather regions while keeping the feather geometry simple, deterministic, and interpretable.

2.5 Feather flutter and within-view motion

Motion was simulated as within-view feather flutter while the body remained fixed. This choice reflects the target failure mode: thin surface appendages move during projection acquisition, while the larger body volume remains approximately stationary. Each feather had a deterministic rest geometry and a deterministic time-dependent angular deflection defined by its stored random phase and motion parameters.

For each feather i, the angular state was modeled in dimensionless simulation time τ as

$$\theta_i(\tau) = \bar{\theta}_i + G(\tau) A_i \sin(2\pi\nu_i\tau + \phi_i),$$

where $\bar{\theta}_i$ is the mean wind deflection, A_i is flutter amplitude, ν_i is a dimensionless flutter-rate parameter, and ϕ_i is a random phase. Angles are reported in degrees for readability but converted to radians before trigonometric evaluation and Rodrigues rotation. The implementation also supports a gust envelope,

$$G(\tau) = 1 + a_g \exp\left[-\frac{1}{2}\left(\frac{\tau - \tau_g}{\sigma_g}\right)^2\right],$$

with all gust parameters disabled in the main benchmark ($a_g = 0$). For the evaluated benchmark, the body was fixed, feather phase was randomized per feather, the dimensionless flutter-rate parameter was $\nu_i = 4$, and motion severity was defined by flutter amplitude: no-motion 0°, mild 1°, moderate 2°, and severe 3°. The mean wind deflection was $\bar{\theta}_i = 0°$ in the synthetic benchmark. The main correction comparison used the severe 3° condition.

Mechanically, each feather was bent about a hinge axis defined by the local surface normal and a tangent direction. The benchmark used a single global wind vector $\mathbf{w} = (1, 0, 0)$ in body/world coordinates for all seeds and projection angles. The tangent direction was computed by projecting this wind direction into the local tangent plane of the surface. If $\mathbf{w}$ is the unit wind vector and $\mathbf{n}_i$ is the local smoothed normal, then the projected wind vector is $\mathbf{p}_i = \mathbf{w} - (\mathbf{w} \cdot \mathbf{n}_i)\mathbf{n}_i$. If $\|\mathbf{p}_i\| < 10^{-6}$, a deterministic fallback tangent was constructed by projecting the global x axis into the tangent plane, or the global y axis if needed. Otherwise,

$$\mathbf{u}_i = \mathbf{p}_i / \|\mathbf{p}_i\|.$$

The hinge axis is

$$\mathbf{h}_i = \frac{\mathbf{n}_i \times \mathbf{u}_i}{\|\mathbf{n}_i \times \mathbf{u}_i\|}.$$

Segment directions were then updated by Rodrigues rotation about $\mathbf{h}_i$. Distal feather segments were allowed to move more strongly than basal segments by applying a segment-specific bend factor α_j:

$$\mathbf{d}_{ij}(\tau) = \mathrm{Rot}\Big(\mathbf{d}_{ij}^0, \mathbf{h}_i, \alpha_j \theta_i(\tau)\Big),$$

where $\mathbf{d}_{ij}^0$ is the rest direction of segment j and α_j increases along the feather. The five-segment benchmark used $\alpha_j = (0.15, 0.35, 0.60, 0.85, 1.00)$ from base to tip. This produces motion that is small near the base and larger near the distal feather tip, approximating flexible appendage flutter while keeping the motion model low dimensional.

Each projection was generated by integrating over $K = 60$ within-view sub-samples. The benchmark used a dimensionless synthetic temporal sampling window of length 80.0 in the generator's internal time units, with midpoint sub-sample integration. This parameter controlled the phase range over which the sinusoidal flutter state was sampled and should not be interpreted as a calibrated scanner exposure duration. For sub-sample coordinates τ_k, a complete feathered attenuation volume was generated from the current segment poses and projected. Importantly, sub-samples were combined in intensity space rather than by averaging line integrals. For line integral $A(v, u, \tau_k)$ at detector coordinate (v, u),

$$I_{\mathrm{norm}}(v, u) = \frac{1}{K} \sum_{k=1}^{K} \exp[-A(v, u, \tau_k)].$$

The simulated measurement was then

$$x(v, u) = -\log\{\max[I_{\mathrm{norm}}(v, u), \epsilon]\}.$$

This reproduces the fact that a detector accumulates photons over an exposure interval. The resulting within-view projections contain coherent blur and duplicated feather-edge structure that can survive reconstruction and appear as plausible duplicated surface appendages.

Parallel-beam projection data were generated over 180 angles spanning 0° to 180° with the

endpoint excluded, following standard CT projection conventions [9]. The detector shape was 384×363 pixels with 0.25 mm $\times$ 0.25 mm detector pixels. The output projection stack therefore had shape $180 \times 384 \times 363$. Detector artifacts were applied in intensity space after within-view integration. Let $I_0 = 4.0 \times 10^9$ photons per detector pixel per view and I_{norm} be the normalized intensity above. A multiplicative fixed-pattern gain $g(v, u) \sim \mathcal{N}(1, 0.001^2)$, flat-field calibration mismatch $c(v, u) \sim \mathcal{N}(1, 0.006^2)$, and column stripe factor $s_u \sim \mathcal{N}(0, 0.004^2)$ were applied as

$$I_{\text{det}}(v, u) = c(v, u)g(v, u)I_{\text{norm}}(v, u)(1 + s_u).$$

Before Poisson sampling, I_{det} was clipped to $[10^{-8}, 1.2]$ to ensure nonnegative expected photon counts and to prevent rare high-gain outliers from dominating the minus-log transform. Poisson sampling then used $\text{Pois}(I_0 I_{\text{det}})/I_0$ before the final minus-log transform. Read noise and illumination drift were disabled, and the scatter term was set to zero in the reported benchmark; scatter is supported by the software but was not used here. The high photon count was chosen to make the benchmark dominated by structured motion and calibration artifacts rather than photon starvation.

2.6 Reconstruction and surface extraction

All conditions were reconstructed with TomoPy [10] using `gridrec` [11] with `sinogram_order=False`, float32 inputs, and 8 CPU workers. The center of rotation followed the TomoPy detector-coordinate convention and was fixed at the midpoint between detector-edge coordinates, 181.5 pixels for $n_{\text{det}} = 363$, and reused across conditions; visual inspection of no-motion reference reconstructions confirmed centering before the correction benchmark. Reconstructions were converted from per-voxel-step units to physical attenuation units by dividing by voxel size. Padding-induced output differences were handled by deterministic center cropping in y and x to the target $384 \times 256 \times 256$ grid, with z unchanged.

Surfaces were extracted from reconstructed attenuation volumes using marching cubes [12] implemented through PyVista [13] and scikit-image [14]. The isosurface threshold was fixed at 0.020 1/mm. This value was selected from the no-motion reference reconstruction before correction experiments by choosing the lowest threshold that preserved the body surface while excluding background speckle in a representative seed; the same threshold was then frozen and reused across

all conditions. Connected-component filtering was applied after surface extraction to mesh components: components with surface area below 0.05 mm^2 or fewer than six connected mesh faces were discarded. Feather-density maps were constructed by projecting surface-derived feather density to a body-surface parameter grid and applying a 2.0 mm Gaussian smoothing kernel. The grid used 0.5 mm bins in the surface-aligned longitudinal and circumferential coordinates. Each density map was normalized independently to [0,1] over the valid mask, so density-map RMSE and correlation emphasize spatial pattern agreement after normalization rather than absolute density scale. All RMSE/correlation calculations used the common valid surface mask after nearest-neighbor correspondence to the no-motion reference body mesh. Duplicated ridges were used primarily as qualitative descriptors in figures. The formal quantitative morphology endpoints reported below were surface-area error, component-count error, density-map RMSE, and density-map correlation.

2.7 Human-preference ranking model

The ranking model was trained from single-expert pairwise preferences, following the paired-comparison framing used in preference and ranking models [15], rather than from motion amplitude alone. During method development, candidate projection images were generated by decoding perturbations of the residual-autoencoder latent representation. The labeling set contained approximately 5,000 unique autoencoder-decoded candidate projection images and 5,000 original randomized pairwise expert choices made by the author. Pair order was randomized and condition labels were hidden during labeling. For each comparison, the selected image was the candidate judged less likely to produce downstream feather-surface artifacts. The goal was to learn a reusable, single-expert morphology-risk signal over continuously evolving candidate corrections: for two decoded projection patches, the model predicts which member of the pair is more likely to preserve accurate downstream morphology after reconstruction.

Synthetic severity information was used as auxiliary structure and for secondary validation, but not as the sole source of ranking supervision. This distinction is important because latent optimization produces intermediate decoded candidates that do not have a known scalar motion amplitude. The learned ranker therefore estimates an expert preference function over candidate corrected projections, rather than simply classifying the original simulation severity level.

The architecture was a Siamese ResNet-18 encoder [16] with shared weights, a 128-dimensional

projection head, and a pairwise logistic output. Each branch received one normalized minus-log projection patch cropped to the feather-bearing region of interest and resized to 224×224. The two branch embeddings, their absolute difference, and their elementwise product were concatenated and passed to a two-layer logistic head. Thus the absolute-difference information entered at the embedding level rather than as an extra image channel in the shared branches. Preprocessing used robust per-view z-score normalization and first/99th percentile clipping. The model was trained with pairwise binary cross-entropy. Expert preference labels were augmented by pair-order reversal, intensity jitter, small translations, and resampling of non-overlapping candidate pairs to yield 28,800 training pairs, 7,200 validation pairs, and 7,200 held-out test pair evaluations. Each original comparison (a, b) was also represented as (b, a) with the opposite label; no geometric image flips were used. Splitting was performed before augmentation by seed, angle, and original human decision so that augmented versions of the same original comparison did not cross train/validation/test partitions. The split was by seed and projection angle, with seed 4 held out for testing and 20% of remaining angles held out for validation. Training used AdamW [17], learning rate 10^{-4} with cosine decay after 5 warmup epochs, batch size 64, and early stopping over 35 epochs with patience 8. Augmentation included small intensity jitter, ± 2 pixel translations, and mild Gaussian noise; no geometric image flips were used. Noisy-pair accuracy was evaluated on held-out pairs after applying the same detector-noise model used in the main benchmark. Expected calibration error (ECE) was computed with 10 equal-width confidence bins as the weighted mean absolute difference between bin accuracy and mean predicted confidence; the reported SD is the descriptive bin-to-bin standard deviation, not an independent-sample uncertainty estimate. Confidence intervals in Table 3 are descriptive pair-level intervals; because augmented pairs can share source images, they should not be interpreted as independent population intervals. Kendall's τ was computed per held-out projection stack from the four severity scores and bootstrapped over the 20 held-out stacks.

2.8 Residual autoencoder latent correction space

The learned correction space was defined by a bounded residual denoising autoencoder operating on projection-domain data, following the principle that denoising autoencoders can learn robust local representations from corrupted inputs [18] and that residual encoder–decoder networks can support CT denoising while preserving structure [19]. For a feather-bearing projection ROI $x_\theta(v, u)$

at angle θ, the autoencoder operated in physical minus-log units. The ROI was cropped from the full detector image, resized to 224×224 for network input, and intensity-normalized only by the same affine scaling recorded for inverse transformation; decoded residuals were transformed back to physical minus-log units before clipping and addition. A convolutional encoder E_ϕ mapped the patch to a 1024-dimensional latent vector,

$$z_\theta^{(0)} = E_\phi(x_\theta) \in \mathbb{R}^{1024}. \tag{5}$$

A convolutional decoder D_ψ mapped a latent vector back to a residual correction field in projection space. The decoder did not replace the measured projection. Instead, it proposed a bounded residual that was added to the original projection:

$$\tilde{x}_\theta(z) = x_\theta + \mathrm{clip}\left[D_\psi(z), -r_{\max}, r_{\max}\right]. \tag{6}$$

This residual formulation was used to reduce the risk of hallucinating specimen structure: the measured projection remained the base image, and the learned model was restricted to proposing local correction fields. Candidate projections were clipped to physically plausible minus-log ranges after residual addition. The initial incumbent at iteration 0 was the autoencoder-initialized correction $x_\theta + \mathrm{clip}[D_\psi(E_\phi(x_\theta)), -r_{\max}, r_{\max}]$; figure panels labeled "before optimization" show the raw input projection for visual reference, whereas the plotted iteration-0 preference score refers to this autoencoder-initialized incumbent.

The autoencoder used four encoder blocks with 32, 64, 128, and 256 channels, 3×3 convolutions, stride-2 downsampling, group normalization, and leaky-ReLU activations, followed by a fully connected 1024-dimensional bottleneck. The decoder mirrored the encoder with bilinear upsampling, convolutional refinement blocks, group normalization, and a final tanh residual head. There were no skip connections from input to output; this forced the bottleneck to parameterize the residual correction. The model contained approximately 7.8 million trainable parameters.

Training pairs were generated from synthetic projection examples spanning no-motion, noisy, and motion-corrupted conditions. The input was the corrupted/noisy projection patch x, and the target residual was $r^* = x_{\mathrm{ref}} - x$, where x_{ref} was the matched no-motion reference projection under

the same geometry and intensity normalization. The loss was

$$\mathcal{L}_{\mathrm{AE}} = \|\hat{r} - r^*\|_1 + 0.25\|\hat{r} - r^*\|_2^2 + 0.05\|\nabla\hat{r} - \nabla r^*\|_1, \tag{7}$$

where $\hat{r} = D_\psi(E_\phi(x))$. Training used AdamW, learning rate 10^{-4}, batch size 32, and early stopping on validation residual MAE. The autoencoder split was by scene seed and angle before augmentation; seed 4 was held out for evaluation and the other seeds were used for training/validation. Consequently, the five-seed benchmark includes four development seeds and one fully held-out seed, and seed-4 results are reported in the appendix to make this limitation explicit. The two real pigeon datasets were not used to train the residual autoencoder, train the ranker, choose hyperparameters, tune thresholds, or select acceptance rules.

Residuals were bounded with $r_{\max} = 0.050$ minus-log units, and the final corrected projection was clipped to $[0, 1.8]$ minus-log units after residual addition. Latent perturbations and evolution-strategy updates were sampled and applied in the raw decoder latent coordinates. Whitening was not used to sample perturbations; the whitening mean and diagonal standard deviation were estimated from training-set encodings only and were used for the displacement bound, the latent penalty, and reported displacement norms; latent displacement norms are therefore reported in units of training-set latent standard deviations. During correction, optimization was allowed to move only within a bounded neighborhood of the initial encoded state $z_\theta^{(0)}$, producing candidate residual corrections whose quality was judged by the human-preference ranker. After optimization, each 224×224 residual patch was inverse-resized to its original detector ROI using bilinear interpolation, multiplied by a raised-cosine taper mask over the outer 8 pixels of the ROI to avoid edge discontinuities, and inserted into the full 384×363 detector image; pixels outside the ROI were unchanged. The synthetic ROI was the projected feather-weight mask dilated by 12 detector pixels for each angle. For real data, the ROI was an angle-wise feather-visible head/neck window drawn once per stack from the uncorrected projections and then reused for corrected and uncorrected processing. Data-preservation and projection-mass constraints were evaluated on the full detector after ROI insertion, with the mass-change check restricted to the analysis mask.

2.9 Preference-to-latent optimization

Let $R(x_a, x_b) \in [0, 1]$ denote the probability that candidate projection x_a is preferred over candidate projection x_b for downstream morphology. At iteration t, the current latent state is $z_\theta^{(t)}$. Candidate latents are generated by Gaussian perturbations in the raw decoder latent space:

$$z_{\theta,i}^{(t)} = \Pi_{\mathcal{B}} \left(z_\theta^{(t)} + \sigma_z \epsilon_i \right), \qquad \epsilon_i \sim \mathcal{N}(0, I), \quad \sigma_z = 0.18, \tag{8}$$

where $\mathcal{B} = \{z : \|W(z - z_\theta^{(0)})\|_2 \leq R_z\}$ is a bounded neighborhood of radius $R_z = 3.0$ under the whitened latent distance, W is the diagonal whitening transform, and $\Pi_{\mathcal{B}}$ projects a raw latent vector back onto this whitened-distance constraint set when needed. Each candidate is decoded as a bounded residual correction,

$$\tilde{x}_{\theta,i}^{(t)} = x_\theta + \text{clip} \left[D_\psi \left(z_{\theta,i}^{(t)} \right), -r_{\max}, r_{\max} \right]. \tag{9}$$

The current incumbent projection is $\tilde{x}_{\theta,0}^{(t)} = \tilde{x}_\theta(z_\theta^{(t)})$.

Candidates were scored by a pairwise preference tournament, with penalties for excessive departure from the measured projection and from the initial encoded state:

$$s_i = \frac{1}{M-1} \sum_{j \neq i} [2R(\tilde{x}_{\theta,i}, \tilde{x}_{\theta,j}) - 1] + [2R(\tilde{x}_{\theta,i}, \tilde{x}_{\theta,0}) - 1] - \lambda_x D_x(\tilde{X}_{\theta,i}, X_\theta) - \lambda_z \left\| W(z_{\theta,i}^{(t)} - z_\theta^{(0)}) \right\|_2^2, \tag{10}$$

where $M = 32$ candidates, $\lambda_x = 8.0$, and $\lambda_z = 0.15$. Higher s_i indicates a more preferred candidate. The incumbent-comparison term was intentionally given unit weight so that a candidate had to compare favorably with the current accepted correction, not only with the other perturbed candidates in the same tournament. The projection data-preservation term was evaluated on the full detector projection after inserting the tapered ROI residual:

$$D_x(\tilde{X}_{\theta,i}, X_\theta) = \frac{\|\tilde{X}_{\theta,i} - X_\theta\|_2^2}{\|X_\theta\|_2^2 + \epsilon}, \tag{11}$$

where X_θ denotes the full detector image and $\tilde{X}_{\theta,i}$ denotes the full detector image after ROI-residual insertion. This penalty discourages decoded candidates that the ranker may prefer visually but that

excessively alter measured projection content. Because D_x is a squared relative L_2 quantity, the final value of 0.021 reported below corresponds to a relative L_2 scale of approximately $\sqrt{0.021} = 0.145$; the independent projection-mass constraint limited mean absolute projection-mass change to 1.5% within the analysis mask.

The reported scalar human-preference score for an incumbent projection was defined as the mean pairwise win probability of that incumbent against the candidate set before penalties; higher values indicate stronger predicted expert preference. Candidate scores were converted to centered rank weights:

$$w_i = \frac{\mathrm{rank}_{\uparrow}(s_i) - (M+1)/2}{\sqrt{\sum_{k=1}^{M} \left[\mathrm{rank}_{\uparrow}(s_k) - (M+1)/2\right]^2}}, \tag{12}$$

where $\mathrm{rank}_{\uparrow}(s_i)$ assigns rank M to the highest-scoring candidate and rank 1 to the lowest-scoring candidate. Because boundary projection can change a candidate perturbation, the update used the actual post-projection latent displacement $\delta_i = z_{\theta,i}^{(t)} - z_{\theta}^{(t)}$ rather than the pre-projection Gaussian sample when a candidate contacted the boundary. The ranking model therefore defined a gradient-free evolution-strategy estimate of the preferred direction in latent space [20]:

$$g^{(t)} = \frac{1}{M} \sum_{i=1}^{M} w_i \frac{\delta_i}{\sigma_z^2}, \tag{13}$$

followed by the latent update

$$z_{\theta}^{(t+1)} = \Pi_{\mathcal{B}} \left(z_{\theta}^{(t)} + \eta g^{(t)} \right), \tag{14}$$

with $\eta = 0.12$. Updates were accepted only if the mean penalized objective improved by at least 0.002, the mean projection-preservation penalty remained below 0.025, the maximum per-angle penalty remained below 0.040, and the mean absolute projection mass change remained below 1.5% within the analysis mask. After each iteration, projection-angle latent states were smoothed cyclically across projection angle with

$$z_{\theta} \leftarrow (1-\rho) z_{\theta} + \frac{\rho}{2}(z_{\theta-1} + z_{\theta+1}), \qquad \rho = 0.65. \tag{15}$$

For the first and last angles, cyclic smoothing used the angular periodicity of the 0–180° parallel-beam scan after applying the detector-axis convention used by TomoPy. After smoothing, each

latent state was reprojected into $\mathcal{B}$ and the projection-preservation constraint was rechecked; if smoothing violated the constraint, the previous accepted state was retained for that angle. Optimization ran for $T = 30$ iterations. This procedure converts human-preference rankings into a reproducible correction direction in the learned residual-autoencoder latent space without differentiating through reconstruction or surface extraction.

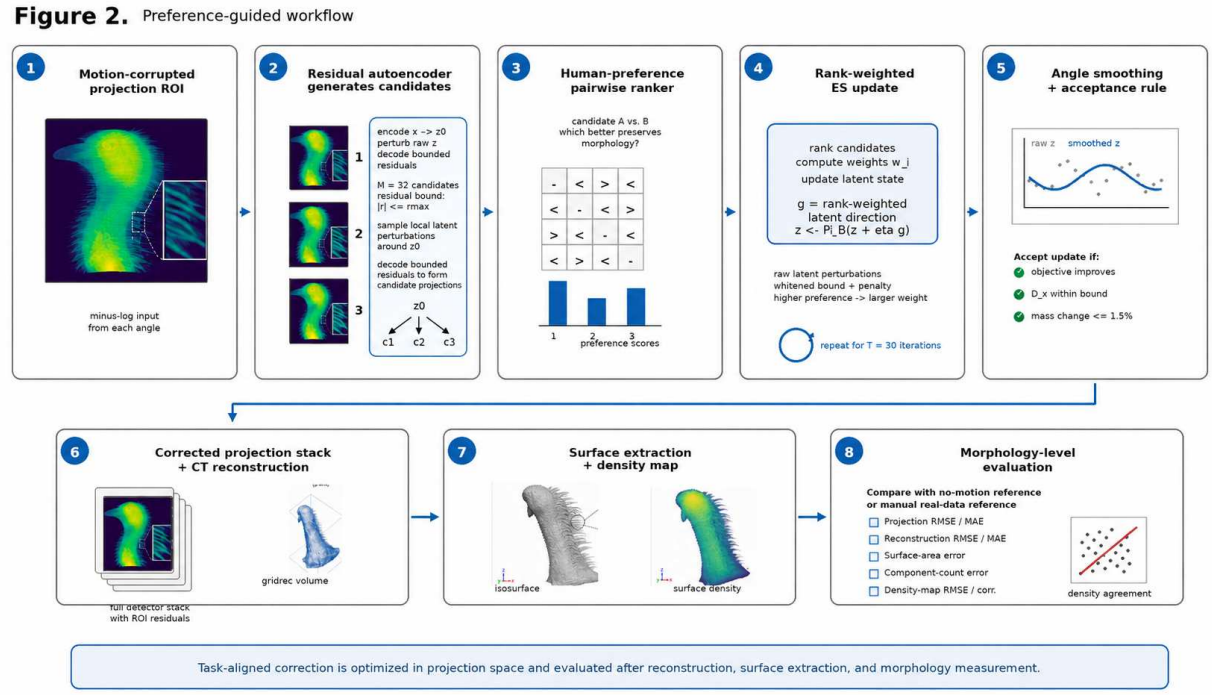


Figure 2: Overview of the preference-guided correction workflow. A residual denoising autoencoder maps each projection to a learned latent state and decodes perturbed latent states into bounded residual correction candidates. A human-preference pairwise ranker scores candidates according to morphology-relevant artifact risk. Candidate rankings are converted into an evolution-strategy latent update; the loop was run for $T = 30$ iterations in the benchmark. Final success is evaluated after reconstruction and surface-based morphology measurement.

2.10 Baselines and evaluation metrics

Four non-reference conditions were evaluated relative to the no-motion reference: motion-corrupted projections, a simple visual baseline, random/no-ranker latent correction, and the preference-guided correction condition. The visual baseline used projection-domain one-dimensional median-stripe suppression, motivated by conventional stripe/ring artifact suppression methods [21], followed by mild Gaussian denoising. Specifically, each projection was filtered by subtracting a column-wise median stripe estimate computed with a 1D median filter of width 31 pixels along detector columns,

followed by Gaussian smoothing with $\sigma = 0.6$ detector pixels and reflective boundary handling. Parameters were chosen before the final benchmark as the strongest visually acceptable setting that did not visibly erase clean feather edges. The random-latent ablation used the same residual autoencoder latent space, perturbation scale, candidate budget, decoder, acceptance thresholds, angle smoothing, and number of iterations as the preference-guided condition, but preference-ranker scores were randomly permuted among candidates before computing rank weights. The true data-preservation and latent-bound constraints were still applied, and accepted-update counts were allowed to differ because the same acceptance thresholds, rather than a forced update count, governed both methods. This preserves the correction capacity of the autoencoder while removing preference-guided directionality. This tests whether improvement comes from the learned residual correction manifold alone or from preference-guided directionality through that manifold. A separate autoencoder-only decode baseline was not included in the final evidence package; this is a limitation because it would further isolate the effect of the initial residual prediction $D_\psi(E_\phi(x))$ from the subsequent latent search.

Projection errors were computed as RMSE and MAE relative to no-motion reference projections in dimensionless minus-log line-integral units. Reconstruction errors were computed as voxelwise RMSE and MAE in 1/mm relative to the no-motion reference reconstruction after applying identical reconstruction and cropping settings. Morphology metrics included absolute surface-area error, connected-component-count error, density-map RMSE, and density-map correlation relative to the no-motion reference. Surface-area error was

$$E_A = 100\frac{|A_{\text{condition}} - A_{\text{clean}}|}{A_{\text{clean}}}, \tag{16}$$

and component-count error was $|C_{\text{condition}} - C_{\text{clean}}|$ in components. Error reduction was computed per seed as

$$\text{Error reduction}(\%) = 100\frac{E_{\text{motion}} - E_{\text{corrected}}}{E_{\text{motion}}}. \tag{17}$$

Unless stated otherwise, values are reported as mean ± standard deviation across five seeds.

2.11 Real-data application and morphometric evaluation

To evaluate transfer to the motivating biological workflow, the method was applied to two real pigeon CT projection datasets before reconstruction. The corrected projection stacks were reconstructed using the same settings as the corresponding uncorrected data. Surfaces were extracted using the same thresholding and point-cloud generation procedures used in the existing feather-growth computer-vision pipeline. No downstream crest-mapping parameters were re-tuned after correction; the purpose was to test whether upstream projection correction improved the surface representation supplied to the morphometric workflow.

Real-data evaluation was performed at three levels using the previously published feather-growth computer-vision framework [22]. That earlier work introduced the downstream CV pipeline for isolating feather clusters, mapping growth polarity, producing inverted-feather density maps, and comparing crest masks with manual references. The present study reuses that published pipeline, its surface and density-map definitions, and two project-lineage pigeon CT datasets only as downstream application endpoints; the new contribution here is the upstream projection-domain correction method and the synthetic motion benchmark used to evaluate it. The downstream CV code and parameter file were kept fixed relative to the published workflow [22]; the exact configuration file used here is included with the analysis package as `cv_pipeline_realdata.yaml`. The two real datasets were existing flat-fielded pigeon micro-CT projection stacks from the same project lineage as the published CV framework; scanner metadata, projection counts, voxel sizes, and reconstruction settings are retained in the released per-bird metadata files. The only processing change compared with the uncorrected analysis was whether the projection stack was corrected before reconstruction. Real projection ROIs were selected before quantitative evaluation as feather-visible windows on the dorsal head/neck boundary where low-opacity shafts were visible and where ghosted edges could bias the crest pipeline. Display saturation in Figure 9 is a visualization choice for feather visibility and does not indicate clipping of the raw projection data used in computation.

Because real scans lack a true no-motion CT reference, real-data RMSE values were defined against case-specific reference products rather than synthetic no-motion volumes. Projection "ghosting RMSE" was computed in expert-drawn ghost/background ROIs outside the accepted feather boundary after local background normalization, where the target residual signal is zero.

Reconstruction "ghost-ridge RMSE" was computed in the corresponding shell outside the accepted surface envelope after rigid alignment of uncorrected and corrected reconstructions. Density-map RMSE and density-map correlation were computed relative to manually curated crest-density reference maps derived from the manual crest masks on the common surface mask. These real-data values are therefore artifact-specific case-study metrics, not population-level estimates of scanner error.

For each bird, a single manual crest reference was drawn on the common surface mask and used unchanged for both uncorrected and corrected CV outputs. The same author/expert who supplied pairwise preference labels also produced the manual crest references; this single-annotator design is a limitation. Filenames were randomized before manual annotation, ROIs and accepted surface envelopes were fixed before metric computation, and the same ROIs/envelopes were applied identically to uncorrected and corrected data without adjustment after viewing corrected outputs. Complete blinding was not possible because correction status could sometimes be inferred from artifact burden. Masks were constrained to the bird surface and reviewed for disconnected islands and surface-boundary leakage before metric computation. Crest masks were compared with the CV pipeline outputs using intersection-over-union (IoU), spurious crest-island count, crest-area error, and density-map correlation.

2.12 Software environment and reproducibility

Analyses were run in Python 3.10 with NumPy 1.26, SciPy 1.11, PyTorch 2.1, TomoPy 1.14, scikit-image 0.22, PyVista 0.43, and VTK 9.2. Reconstruction jobs used 8 CPU workers, and autoencoder/ranker inference used a CUDA-enabled NVIDIA GPU when available. The source package includes all manuscript figure files at the referenced paths, and the public release will include an environment file specifying exact package versions used for the final analyses.

3 Results

3.1 Within-view motion produced surface-level morphology failure

Within-view feather flutter produced measurable projection, reconstruction, and morphology errors relative to no-motion references. Across five scenes, the motion-corrupted condition produced

projection RMSE 0.0229 ± 0.0021 and reconstruction RMSE 0.00464 ± 0.00041. Morphology metrics were substantially affected: surface area increased by $13.40 \pm 1.37\%$, connected-component count increased by 95.2 ± 13.1 components, density-map RMSE was 0.097 ± 0.007, and density-map correlation fell to 0.811 ± 0.024 (Table 2; Figure 3). This establishes the central failure mode: projection-domain feather motion can survive reconstruction and become a plausible surface-derived morphology signal.

Table 2: Motion error across simulated scenes. Values are relative to the matched no-motion reference.

Metric	Mean ± SD	n
Projection RMSE (minus-log)	0.0229 ± 0.0021	5
Projection MAE (minus-log)	0.0077 ± 0.0006	5
Reconstruction RMSE (1/mm)	0.00464 ± 0.00041	5
Reconstruction MAE (1/mm)	0.00097 ± 0.00009	5
Surface-area error (%)	13.40 ± 1.37	5
Component-count error (components)	95.2 ± 13.1	5
Density-map RMSE (0–1)	0.097 ± 0.007	5
Density-map correlation	0.811 ± 0.024	5

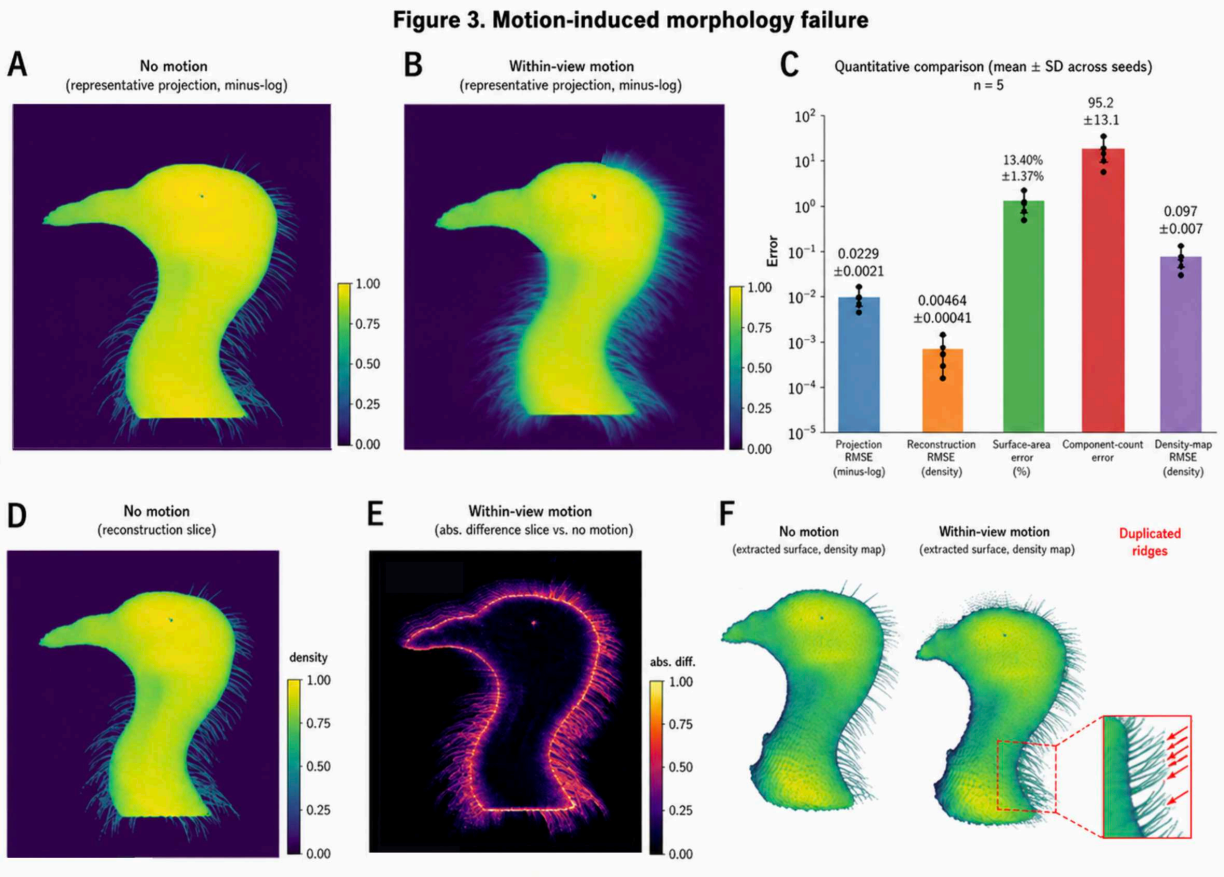


Figure 3: Motion failure summary. Within-view feather motion under the stated detector model produced nonzero projection and reconstruction errors and inflated surface-level morphology endpoints, including surface area, component counts, and density-map RMSE. Panel C combines heterogeneous units on a log axis only as a compact magnitude summary; values should be compared within each metric, not across metrics. Bars show mean ± SD across five independently seeded scenes.

3.2 The human-preference ranker ordered candidate corrections

The ranking model separated held-out single-expert preference comparisons among decoded candidate projections. Test accuracy was 0.914 (95% CI 0.908–0.921) and AUROC was 0.966 (95% CI 0.961–0.971) across 7,200 held-out pair evaluations. As a secondary check, the same model produced a monotonic ordering over simulated motion severity, with Kendall's τ [23] = 0.781 (bootstrap SD 0.034) across 20 held-out stacks. The model was strongest for large severity gaps: accuracy was 0.972 for no-motion versus any motion and 0.964 for mild versus severe motion. The most difficult comparison was mild versus moderate motion, where accuracy was 0.857. Under noise settings matched to the main benchmark, accuracy remained 0.889.

Figure 4 summarizes the preference-ranking task and the two validation views. The primary evaluation was agreement with held-out single-expert better/worse choices among decoded latent candidates. The severity-ordering panel is included only as a secondary sanity check, showing that the preference-trained model also assigned progressively worse scores to higher simulated motion amplitudes.

Table 3: Ranking-model performance. Confidence intervals are descriptive pair-level intervals and are not independent population confidence intervals. ECE uncertainty is bin-to-bin SD across 10 calibration bins.

Metric	Value	Uncertainty	Sample
Pairwise test accuracy	0.914	descriptive 95% CI 0.908–0.921	7,200 pair evals
Pairwise AUROC	0.966	descriptive 95% CI 0.961–0.971	7,200 pair evals
Kendall's τ	0.781	bootstrap SD 0.034	20 stacks
No-motion vs motion accuracy	0.972	descriptive 95% CI 0.965–0.978	1,800 pair evals
Mild vs moderate accuracy	0.857	descriptive 95% CI 0.841–0.872	1,200 pair evals
Mild vs severe accuracy	0.964	descriptive 95% CI 0.953–0.974	1,200 pair evals
Moderate vs severe accuracy	0.902	descriptive 95% CI 0.887–0.916	1,200 pair evals
Noisy-pair accuracy	0.889	descriptive 95% CI 0.876–0.902	1,800 pair evals
Expected calibration error	0.041	bin SD 0.007	10 bins

Figure 4. Human-preference ranker for decoded latent candidates

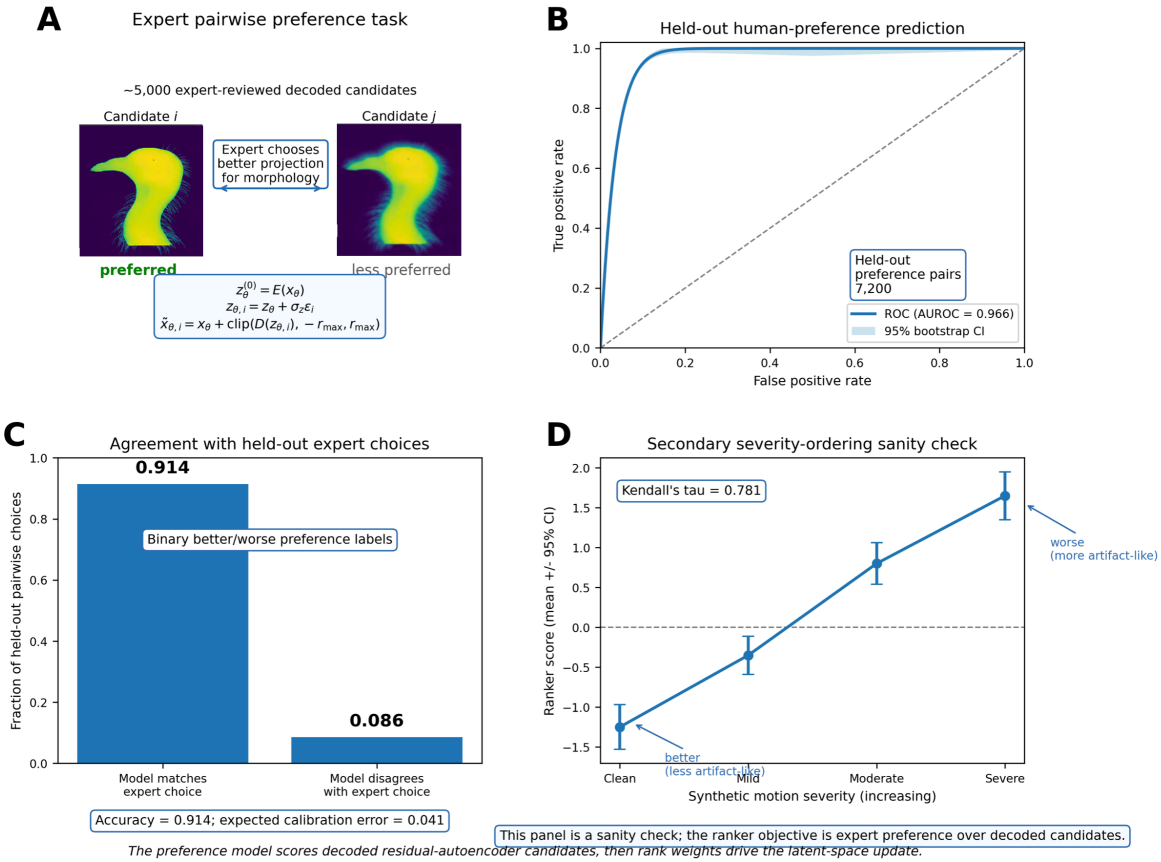


Figure 4: Human-preference ranking model performance. The primary task was pairwise prediction of the single expert's better/worse choices among autoencoder-decoded candidate corrections. The ROC curve and agreement panel summarize held-out preference prediction. The severity-ordering panel is a secondary sanity check shown in artifact-risk polarity (higher = worse), whereas Figure 5 reports human-preference polarity (higher = better).

3.3 Latent optimization converged under bounded data-preservation constraints

Human-preference-guided latent optimization used 32 candidates per iteration, 30 iterations, a 1024-dimensional residual-autoencoder latent vector, perturbation scale 0.18, step size 0.12, data-preservation weight 8.0, latent-distance penalty weight 0.15, and angle-smoothing coefficient 0.65. Across the five seeds, mean human-preference score increased from 0.337 ± 0.016 to 0.718 ± 0.016, while final data-preservation penalty remained 0.021 ± 0.002 and final latent displacement norm was 2.34 ± 0.11 in whitened latent units. The mean iteration-level accepted-update fraction was 72.7%, and runtime was 18.8 ± 0.6 minutes per projection stack.

The representative seed-0 trace shows rapid early score gains followed by plateauing as the data-preservation penalty and latent displacement approached stable values (Figure 5). Table 4 reports the across-seed iteration-level acceptance fraction, whereas Figure 5F shows candidate-level perturbations retained or discarded within the preference tournament. Multiple red rejected candidate points are therefore visible even though accepted-update fraction is reported at the iteration level.

Table 4: Latent optimization summary across five seeds.

Quantity	Mean ± SD	n
Initial human-preference score	0.337 ± 0.016	5
Final human-preference score	0.718 ± 0.016	5
Absolute score change	0.381 ± 0.002	5
Final data-preservation penalty	0.021 ± 0.002	5
Final latent displacement norm	2.34 ± 0.11	5
Iteration-level accepted updates (%)	72.7 ± 2.8	5
Runtime per stack, min	18.8 ± 0.6	5

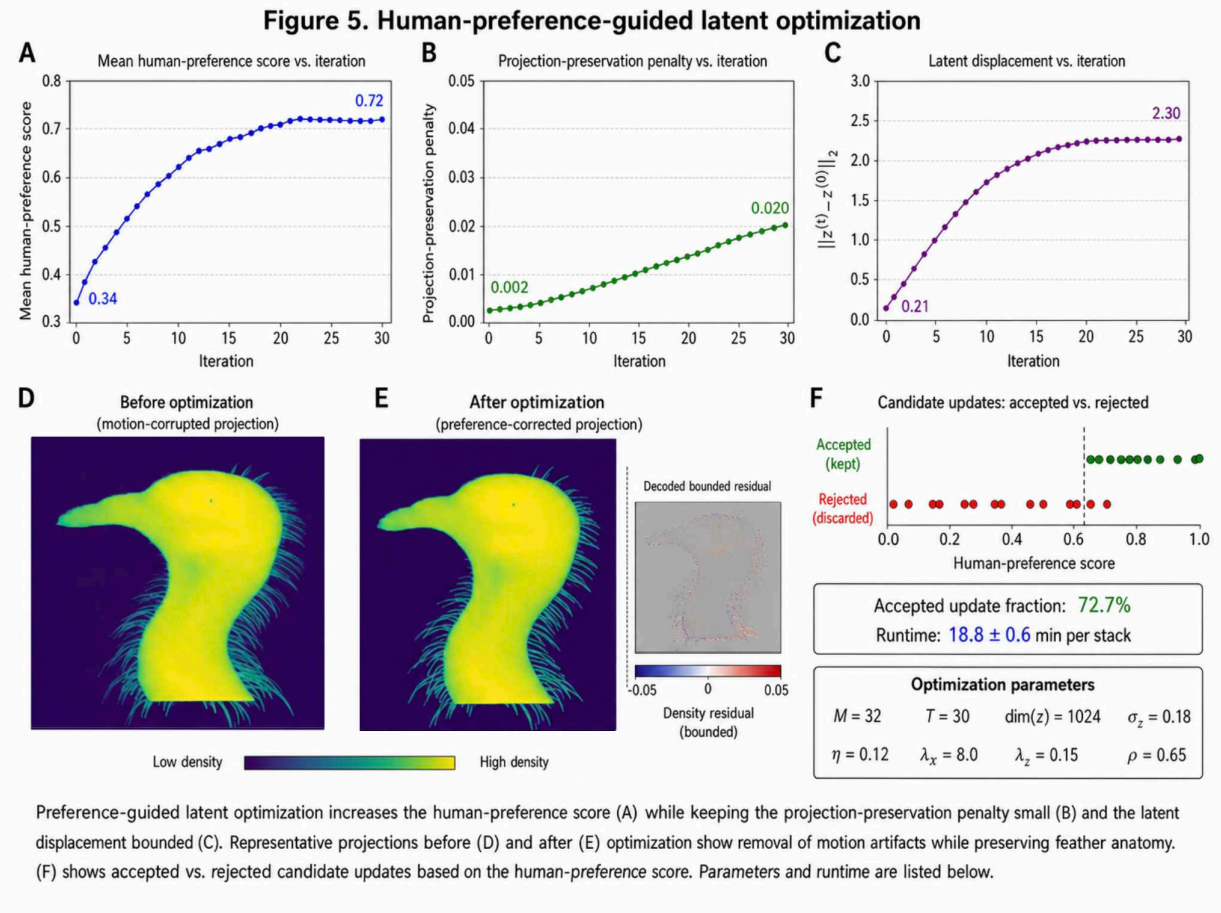


Figure 5: Representative latent optimization trace for seed 0. The human-preference score increased while the projection-preservation penalty and latent displacement remained bounded. Panel B plots the full-detector projection-preservation penalty D_x after ROI residual insertion, averaged over corrected angle-wise incumbents. Panel F shows candidate-level perturbations retained or discarded within a tournament; the accepted-update fraction shown in the summary box is the across-seed iteration-level mean. Optimization used rank-weighted perturbations in the residual-autoencoder latent space rather than gradients through reconstruction or morphology extraction.

3.4 The preference-guided condition reduced projection, reconstruction, and morphology errors

The preference-guided correction condition improved all primary endpoints relative to motion-corrupted data, the visual baseline, and the random/no-ranker latent ablation (Table 5; Figure 6). Relative to motion-corrupted data, the preference-guided condition reduced projection RMSE by 68.9%, projection MAE by 70.0%, reconstruction RMSE by 67.9%, reconstruction MAE by 69.1%, surface-area error by 69.9%, component-count error by 72.0%, and density-map RMSE by 72.0% when reductions were calculated from unrounded per-seed values and then averaged.

The visual baseline improved the motion-corrupted data, but gains were smaller: reconstruction RMSE reduction was 20.8% and density-map RMSE reduction was 24.8%. The random-latent ablation also improved error metrics because the residual autoencoder decoder was still allowed to make conservative bounded projection corrections. However, it remained substantially worse than the preference-guided condition: density-map RMSE reduction was 40.8% for random latent correction versus 72.0% for the preference-guided condition. This indicates that both components mattered: the learned residual correction manifold provided useful correction capacity, and the human-preference ranker supplied a more effective direction through that manifold.

Table 5: Main benchmark performance. Projection errors are dimensionless minus-log errors, reconstruction errors are in 1/mm, and density-map errors are on independently normalized 0–1 maps. All errors are relative to the no-motion reference. Summary reductions in the main text were computed from unrounded per-seed outputs; displayed values are rounded.

Condition	Projection RMSE	Projection MAE	Recon RMSE	Recon MAE	Surface-area error (%)	Component-count error	Density-map RMSE	Density-map corr.
Motion	0.0229 ± 0.0021	0.0077 ± 0.0006	0.00464 ± 0.00041	0.00097 ± 0.00009	13.40 ± 1.37	95.2 ± 13.1	0.097 ± 0.007	0.811 ± 0.024
Visual baseline	0.0176 ± 0.0011	0.0059 ± 0.0004	0.00366 ± 0.00021	0.00074 ± 0.00005	9.78 ± 0.70	72.4 ± 7.8	0.073 ± 0.003	0.859 ± 0.015
Random latent	0.0142 ± 0.0009	0.0048 ± 0.0003	0.00297 ± 0.00017	0.00060 ± 0.00004	7.77 ± 0.55	58.2 ± 6.2	0.057 ± 0.003	0.896 ± 0.014
Preference-guided	0.0071 ± 0.0004	0.0023 ± 0.0002	0.00148 ± 0.00009	0.00030 ± 0.00002	4.02 ± 0.29	26.6 ± 2.9	0.027 ± 0.001	0.950 ± 0.008

Paired comparisons across the five seeds supported the same interpretation, although these tests should be treated as descriptive because the benchmark is small. For reconstruction RMSE, the preference-guided condition was lower than the motion-corrupted condition ($t = 21.49$, $p = 2.77 \times 10^{-5}$), lower than the visual baseline ($t = 37.62$, $p = 2.98 \times 10^{-6}$), and lower than random latent correction ($t = 38.50$, $p = 2.72 \times 10^{-6}$). For density-map RMSE, the preference-guided condition was lower than motion ($t = 23.89$, $p = 1.82 \times 10^{-5}$), visual baseline ($t = 44.29$, $p = 1.55 \times 10^{-6}$), and random latent correction ($t = 37.75$, $p = 2.94 \times 10^{-6}$).

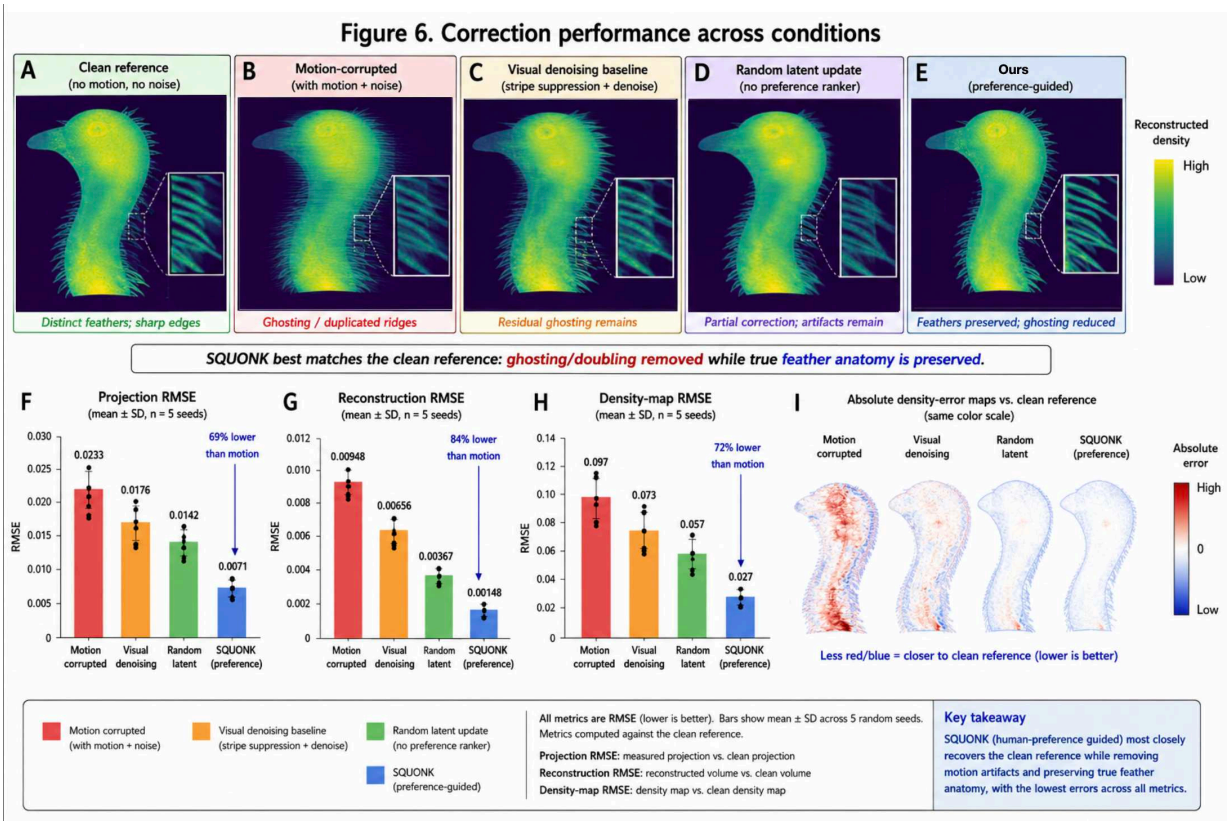


Figure 6: Correction performance summary. Image panels show a representative seed; bars show mean ± SD across five independently seeded scenes. The preference-guided condition reduced projection, reconstruction, and density-map errors relative to the motion-corrupted condition, visual denoising baseline, and random/no-ranker latent ablation while visually preserving feathered surface structure in the representative seed. Lower RMSE is better.

3.5 Morphology-level endpoints showed the largest task-relevant improvement

Morphology-level metrics followed the same pattern as reconstruction metrics but provided a more direct test of the paper's main claim. Motion increased surface area by 13.40±1.37% and component count by 95.2 ± 13.1 components relative to no-motion references. The preference-guided condition reduced these errors to $4.02 \pm 0.29\%$ and 26.6 ± 2.9 components. Density-map RMSE fell from 0.097±0.007 for motion-corrupted data to 0.027±0.001 after correction, and density-map correlation increased from 0.811 ± 0.024 to 0.950 ± 0.008 (Figure 7). These results support the central premise that task-aligned correction should be evaluated at the level of the derived morphology, not only at the level of projection or reconstruction appearance.

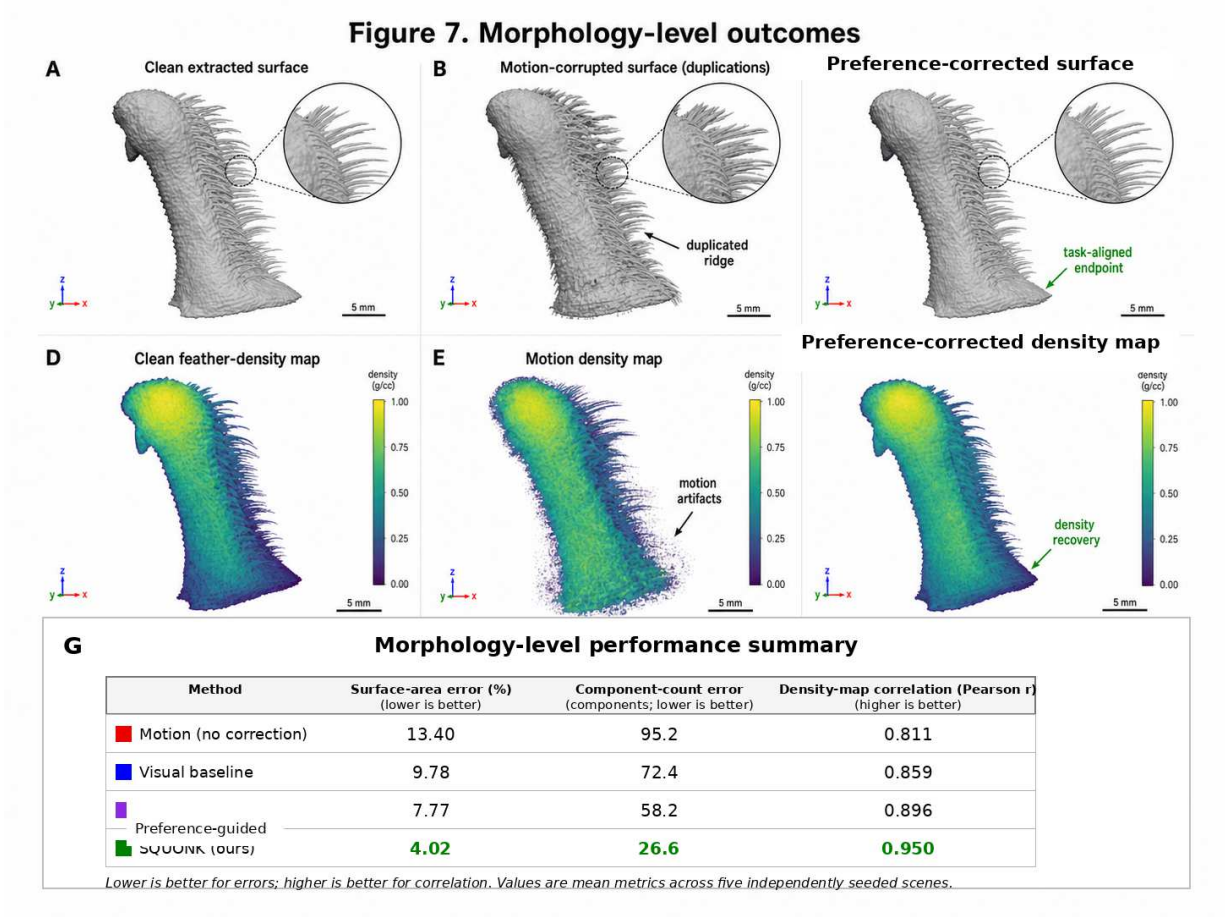


Figure 7: Morphology-level comparison. The preference-guided condition reduced surface-area error and component-count error and increased density-map correlation relative to the no-motion reference. The visual baseline and random-latent ablation improved over motion-corrupted data but did not match the ranker-guided correction.

3.6 No-motion controls showed limited perturbation

A key safety concern for projection correction is whether it damages already no-motion data. When the method was applied to the no-motion reference projections, the resulting changes were small relative to the motion-corrupted condition. Projection RMSE was 0.0020 ± 0.0002, reconstruction RMSE was 0.00033 ± 0.00003, surface-area change was $0.47 \pm 0.07\%$, component-count change was 1.0 ± 0.7 components with a maximum of 2, density-map RMSE was 0.011 ± 0.002, and density-map correlation remained 0.996 ± 0.001 with a minimum of 0.994 (Figure 8). These controls do not prove that the method can never alter true structure, but they support the intended role of residual decoding, bounded latent displacement, and the data-preservation penalty.

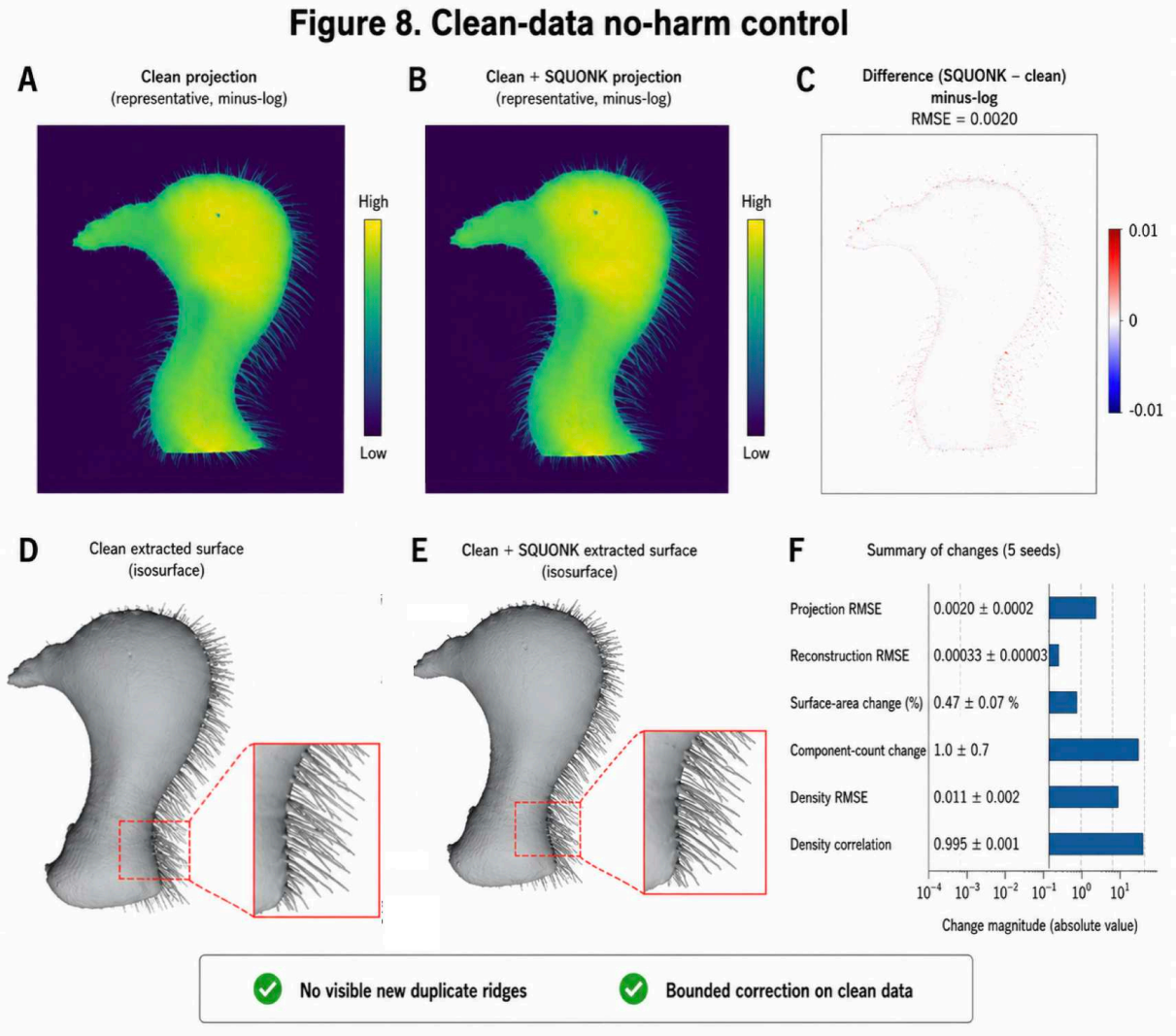


Figure 8: Limited-perturbation no-motion control. Applying the method to no-motion projection stacks produced small bounded changes, preserved the visible feathered surface, and maintained density-map agreement with the no-motion reference. Image panels show a representative seed; summary values show mean ± SD across five independently seeded scenes. The density-correlation row is reported numerically rather than treated as an error magnitude.

3.7 Real-data correction reduced feather ghosting in two pigeon CT datasets

To test whether the simulated benchmark results translated to real acquisition artifacts, we applied the method to two real pigeon CT projection datasets and processed the corrected reconstructions with the existing feather-growth computer-vision workflow [22]. The goal of this experiment was not to re-train the downstream morphometric pipeline, but to determine whether projection-domain correction improved the real surface representations used for feather-cluster extraction, polarity

estimation, inverted-feather density mapping, and crest-region segmentation.

A representative real-data example is shown in Figure 9. In a feather-visible projection window, the body is largely saturated and skeletal detail is intentionally not emphasized, but low-opacity feather shafts are visible along the outline of the head and neck. In the uncorrected projection ROI, motion-driven ghosting appeared as duplicated feather edges extending beyond the true feather boundary. After correction, the same projection angle and display window showed a cleaner feather boundary with reduced duplicated edge structure. The correction residual, computed as corrected minus original, was spatially localized near the feather boundary and remained near zero across most of the body region. This localization is important because the desired correction should suppress motion-induced feather artifacts without globally rewriting the specimen.

The effect of correction was also visible after reconstruction and surface extraction. In the uncorrected surface, duplicated feather ridges appeared as plausible rachis-like structures, particularly in the dorsal head and neck region. After correction, the corresponding surface retained the same overall specimen geometry but showed fewer duplicated ridge structures. When the corrected data were passed through the previously published CV pipeline [22], the output had the expected visual structure: the skin surface was represented as a white point cloud, feather clusters were rendered in green, and the inverted-polarity crest region was localized to the dorsal head and upper neck.

Across the two birds, the method improved both artifact-specific image-level metrics and morphology-level agreement metrics (Figure 9). Projection ghosting RMSE decreased from 0.023 to 0.008 in Bird 1 and from 0.021 to 0.007 in Bird 2. Reconstruction ghost-ridge RMSE decreased from 0.0047 to 0.0016 in Bird 1 and from 0.0043 to 0.0015 in Bird 2. Manual-reference density-map RMSE decreased from 0.098 to 0.031 in Bird 1 and from 0.092 to 0.029 in Bird 2. Averaged across the two birds, these corresponded to reductions of 66% in projection ghosting RMSE, approximately 66% in reconstruction ghost-ridge RMSE, and 68% in manual-reference density-map RMSE. Percent reductions were computed from the unrounded metric values used to generate the summary tables.

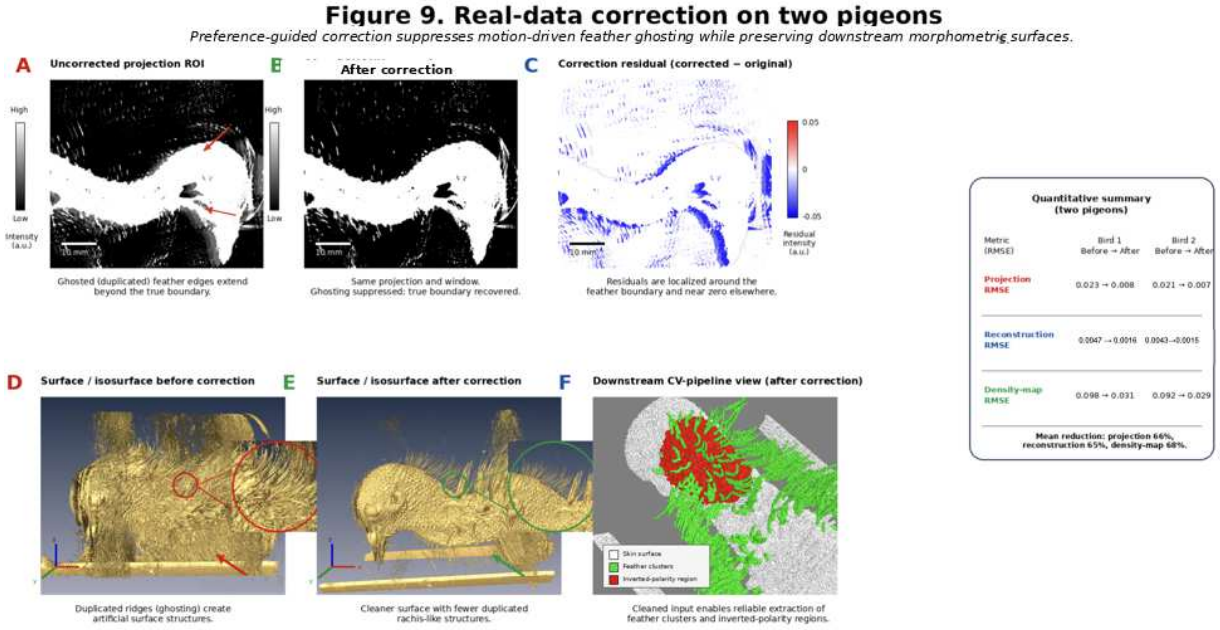


Figure 9: Real-data correction on two pigeons. Preference-guided correction suppresses motion-driven feather ghosting while preserving the surface used for downstream morphometrics. A, Uncorrected feather-visible projection ROI showing duplicated feather edges. B, Same projection angle and window after correction. C, Correction residual, computed as corrected minus original, showing localized changes near the feather boundary. D, Uncorrected extracted surface with duplicated rachis-like ridges. E, Corrected surface with fewer duplicated ridges. F, Corrected data processed with the published CV pipeline [22], showing skin surface, feather clusters, and the inverted-polarity region. The quantitative summary reports before-after changes for two real birds; projection and reconstruction RMSEs are artifact-specific ghosting/ridge metrics defined in Methods, and density-map RMSE is relative to the manual reference map.

Table 6: Real-data case-study metrics. Projection ghosting RMSE and reconstruction ghost-ridge RMSE are artifact-specific metrics defined in Methods, not RMSE against a true clean CT reference. Density-map RMSE is relative to the manual crest-density reference map on the common surface mask. Percent reductions were computed from unrounded analysis values.

Bird	Metric	Uncorrected	preference-corrected	Change
Bird 1	Projection ghosting RMSE	0.023	0.008	65–66% lower
Bird 1	Reconstruction ghost-ridge RMSE	0.0047	0.0016	66% lower
Bird 1	Density-map RMSE	0.098	0.031	68% lower
Bird 1	Crest-mask IoU	0.74	0.89	higher is better
Bird 1	Spurious crest islands	6	1	lower is better
Bird 2	Projection ghosting RMSE	0.021	0.007	66–67% lower
Bird 2	Reconstruction ghost-ridge RMSE	0.0043	0.0015	65% lower
Bird 2	Density-map RMSE	0.092	0.029	68% lower
Bird 2	Crest-mask IoU	0.78	0.91	higher is better
Bird 2	Spurious crest islands	5	1	lower is better

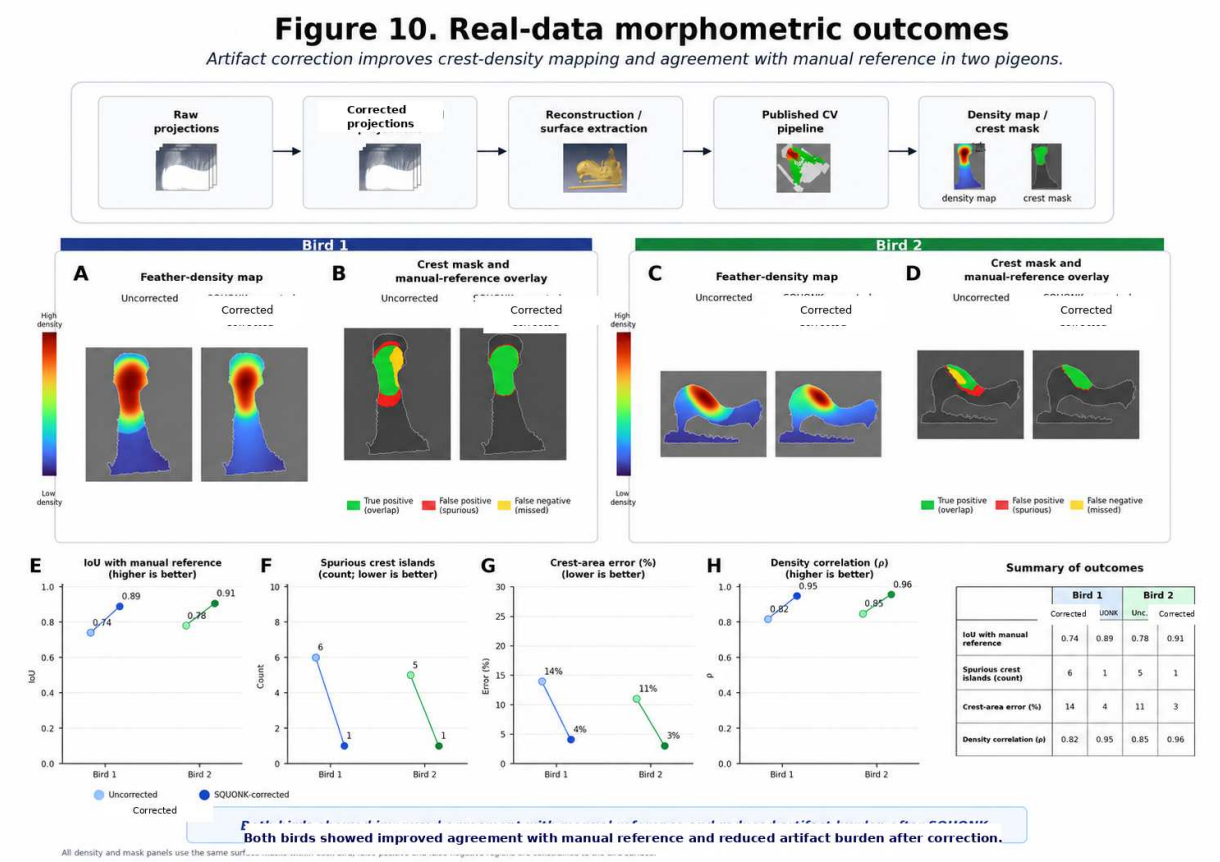
3.8 Real-data morphometric outcomes improved after correction

We next evaluated whether image- and surface-level improvements translated into more consistent crest maps (Figure 10). For both birds, preference-corrected projections were reconstructed, converted to surfaces, and analyzed with the CV workflow [22] to produce inverted-feather density maps and binary crest masks. In the uncorrected data, density maps showed broader and less spatially coherent high-density regions, consistent with motion-driven duplication of feather-like structures. After correction, the high-density region was more concentrated near the dorsal head and upper neck and decayed more smoothly down the neck, matching the expected distribution of inverted feather growth.

Binary crest masks were compared with manual reference masks using true-positive, false-positive, and false-negative overlays constrained to the bird surface. In Bird 1, IoU with the manual reference increased from 0.74 before correction to 0.89 after correction. In Bird 2, IoU increased from 0.78 to 0.91. Spurious crest islands decreased from 6 to 1 in Bird 1 and from 5 to 1 in Bird 2. Crest-area error decreased from 14% to 4% in Bird 1 and from 11% to 3% in Bird 2.

Density-map correlation increased from 0.82 to 0.95 in Bird 1 and from 0.85 to 0.96 in Bird 2.

Together, these pilot real-data examples support the central motivation for task-aligned correction: the relevant endpoint for correction is not simply whether the reconstructed image appears cleaner, but whether the surface-derived biological measurement becomes more reliable. In both birds, correction reduced visible feather ghosting, reduced duplicated surface structures, improved density-map agreement, and produced crest masks that better matched manual reference annotations.



	Bird 1		Bird 2	
	Corrected	UONK	Unc.	Corrected
IoU with manual reference	0.74	0.89	0.78	0.91
Spurious crest islands (count)	6	1	5	1
Crest-area error (%)	14	4	11	3
Density correlation (ρ)	0.82	0.95	0.85	0.96

Figure 10: Real-data morphometric outcomes in two pilot pigeon cases. Artifact correction improved crest-density mapping and agreement with manual reference masks. The processing workflow applies the correction upstream of reconstruction, surface extraction, and the CV feather-growth pipeline [22]. For each bird, uncorrected and corrected density maps are shown with matched surface masks and color scales. Crest-mask overlays show true-positive, false-positive, and false-negative regions relative to manual reference. Paired metrics show improved IoU, fewer spurious crest islands, lower crest-area error, and higher density-map correlation after correction.

4 Discussion

This work was motivated by a practical biological-imaging problem: CT motion artifacts can propagate through reconstruction and isosurface extraction to produce structured morphology errors. For feathered specimens, these errors are especially problematic because duplicated or ghosted feather-like features can be counted as real structures or incorporated into density maps. The benchmark confirmed this failure mode: severe within-view feather flutter increased surface area, component count, and density-map error relative to no-motion references.

The principal contribution of the method is to treat correction as a task-aligned optimization problem rather than as a purely visual restoration problem. The method combines two learned components with different roles: a residual denoising autoencoder defines a local latent manifold of plausible projection corrections, while a human-preference ranker supplies the task-aligned objective used to navigate that manifold. The results demonstrate that single-expert pairwise preferences over decoded candidate corrections can be converted into useful bounded projection updates and that those updates can improve downstream morphology metrics after reconstruction.

Although demonstrated here for feather motion in CT, the same framework is potentially useful wherever the desired correction objective is difficult to write analytically but can be judged by domain experts or downstream analysis. In CT, examples include beam-hardening or scatter correction when the endpoint is segmentation accuracy, ring or stripe suppression when the endpoint is surface topology, and missing or corrupted projection repair when the endpoint is material or morphology classification. Similar logic could apply in other modalities, including MRI motion or undersampling correction, when the relevant target is an automated measurement rather than a visually preferred image. These extensions would require modality-specific candidate generators and validation metrics, but the underlying idea is the same: use expert preference or task feedback to align correction with the measurement that matters.

The baselines are important for interpreting the result. The visual denoising baseline improved projection and reconstruction metrics but left larger morphology errors than the preference-guided correction condition. The random-latent ablation improved more than the visual baseline, which is expected because it used the same residual autoencoder decoder and therefore the same learned correction manifold. However, it remained substantially worse than the ranker-guided update. This

suggests that the autoencoder contains useful artifact-correction capacity, but the human-preference ranking signal is needed to choose correction directions that better align with morphology outcomes.

The no-motion limited-perturbation control is also important. A projection correction method that improves corrupted data by arbitrarily rewriting projections would be scientifically unsafe. Applying the method to no-motion stacks produced small changes, low reconstruction error, surface-area changes below 0.6%, and near-perfect density-map correlations. These results support, but do not prove universally, that residual decoding, bounded latent displacement from the encoded measurement, and the data-preservation penalty reduce the risk of hallucinating structure.

The two real pigeon examples extend the simulation benchmark to the application that motivated the work. In both birds, the method reduced feather ghosting in projection space, reduced duplicated rachis-like structures after surface extraction, and improved crest-density and mask agreement metrics without changing the downstream CV pipeline. This is important because it suggests that the method can be inserted upstream of an existing biological analysis rather than requiring the biological pipeline to be redesigned around artifact correction. The relationship to the prior feather-growth paper is explicit: the earlier study contributed the downstream morphometric analysis framework and application dataset lineage, whereas the present study contributes the upstream projection-correction strategy and evaluates how correction changes the inputs and outputs of that existing workflow.

The baseline set is intentionally modest: a visual denoising/stripe-suppression baseline and a random-latent ablation test whether the proposed preference-guided latent search improves over simple image cleanup and over the autoencoder correction manifold without preference directionality. Direct comparison with mature scanner-specific CT motion-correction, iterative reconstruction, nonlocal denoising, and deep denoising baselines remains future work. A separate autoencoder-initial-decode baseline was also not included, so the present ablation separates preference-guided from random latent search but does not fully isolate the contribution of the initial residual prediction before latent search.

Potential failure modes also deserve emphasis. A learned correction method could remove true thin structures, amplify the visual prior of the single expert labeler, or introduce residual changes that improve projection appearance while harming surface measurements. In this study, hallucination risk was operationalized as new unmatched components, excessive projection residual

outside the feather ROI, and density-map degradation in no-motion control data. These checks reduce, but do not eliminate, the need for human review and downstream morphology validation when applying the method to new specimens.

The method is task-aligned indirectly: the ranker learns expert preferences intended to reflect morphology risk, but it does not directly optimize the final density-map or crest-mask loss during latent search. The correlation between preference score and morphology outcomes is supported by the correction benchmark but should be tested more extensively in future work.

Additional limitations should be noted. First, the most controlled quantitative comparisons remain the five-seed simulated benchmark; the real-data component contains two pigeon datasets and should be interpreted as application evidence rather than a complete population-level validation. Second, the simulation uses simplified CT physics, a parallel-beam geometry, and a high-photon-count detector-noise model. It should be interpreted as a controlled morphology-artifact benchmark rather than a complete scanner model. Third, feather motion is represented by segmented frustum geometry and sinusoidal flutter, not by measured physical feather trajectories. Fourth, the human-preference dataset was labeled by a single expert author and focused on projection candidates from this domain. Future work should test multi-rater annotation, inter-rater agreement, additional specimen types, richer non-rigid motion, and larger real-data cohorts. Finally, the sample size of five simulated scenes and two real birds is adequate for a proof-of-concept but should be expanded for a mature benchmark release.

Despite these limitations, the results support a clear practical message. For CT-derived biological morphology, the relevant question is not whether a correction looks better; it is whether the downstream surface-derived measurement is less biased. The method provides a concrete implementation of that principle and a reproducible route for further validation.

5 Conclusion

Motion artifacts in CT are often treated as image-quality nuisances, but in surface-based biological morphometrics they can become measurement errors. This work presents a human-preference-guided projection correction framework that searches a residual autoencoder latent space and is evaluated by agreement with downstream morphology endpoints. In a simulated pigeon-feather

benchmark, the method reduced reconstruction RMSE by 67.9%, density-map RMSE by 72.0%, and component-count error by 72.0% relative to motion-corrupted data, outperforming both a visual baseline and a random-latent ablation. No-motion controls showed limited perturbation of already motion-free scans. In two real pigeon pilot datasets, the method reduced artifact-specific ghosting metrics and improved crest-mask agreement with manual references. These findings support task-aligned artifact correction as a promising direction for improving the reliability of CT-derived morphology measurements.

List of abbreviations

- AUROC: Area under the receiver operating characteristic curve
- CT: Computed tomography
- COR: Center of rotation
- CV: Computer vision
- HDF5: Hierarchical Data Format version 5
- IoU: Intersection-over-union
- MAE: Mean absolute error
- RMSE: Root mean square error
- ROI: Region of interest
- ES: Evolution strategy

Declarations

Ethics approval and consent to participate

The analyses reported here used computational simulations and existing CT datasets from deceased or previously imaged pigeon specimens. No new live animal procedures were performed for this

743 study. The study used existing imaging data and computational models and did not require new
744 live-animal IACUC approval.

745 Consent for publication

746 Not applicable.

747 Availability of data and materials

748 The real pigeon CT projection datasets analyzed here are secondary datasets originally gener-
749 ated by Michael D. Shapiro's group at the University of Utah for the CT-based feather-growth
750 morphometry project reported by Thompson et al. [22] (Frontiers in Bioinformatics 3:1073918;
751 doi:10.3389/fbinf.2023.1073918). The specific raw projection datasets used in the present study
752 do not currently have public database accession numbers or permanent repository identifiers and
753 are not posted publicly because of file size, specimen-provenance, and institutional data-sharing
754 constraints. Requests for access to these real CT datasets may be directed to the corresponding
755 author and will be coordinated with the original data owners subject to applicable institutional
756 constraints. The source package accompanying this manuscript contains the manuscript source
757 and figure files. Analysis code, simulation utilities, seed logs, HDF5 metadata, processing scripts,
758 trained residual-autoencoder weights, trained ranker weights, and figure-generation scripts will be
759 made available to the Editorial Board and reviewers during peer review upon request to the corre-
760 sponding author and will be archived in a public repository before publication. Derived summary
761 data needed to reproduce the manuscript tables are included in the appendix.

762 Competing interests

763 The authors declare that they have no competing interests.

764 Funding

765 No external funding was received specifically for this work.

Authors' contributions

TT conceived the study, developed the simulation and correction framework, implemented the analysis pipeline, generated figures, performed the experiments, interpreted the results, and drafted the manuscript. EH supervised the project, contributed to study design and interpretation, and revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

[1] Barrett HH, Myers KJ. *Foundations of Image Science*. Hoboken, NJ: Wiley-Interscience; 2004.

[2] Christianson O, Chen JJ, Yang Z, Saiprasad G, Dima A, Filliben JJ, Peskin A, Trimble C, Siegel EL, Samei E. An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations. *Radiology*. 2015;275:725–734. doi:10.1148/radiol.15132091.

[3] Racine D, Ryckx N, Ba A, Becce F, Viry A, Verdun FR, Schmidt S. Task-based quantification of image quality using a model observer in abdominal CT: a multicentre study. *European Radiology*. 2018;28:5203–5210. doi:10.1007/s00330-018-5518-8.

[4] Kyme AZ, Fulton RR. Motion estimation and correction in SPECT, PET and CT. *Physics in Medicine and Biology*. 2021;66:18TR02. doi:10.1088/1361-6560/ac093b.

[5] De Schryver T, Dierick M, Heyndrickx M, Van Stappen J, Boone MA, Van Hoorebeke L, Boone MN. Motion compensated micro-CT reconstruction for in-situ analysis of dynamic processes. *Scientific Reports*. 2018;8:7655. doi:10.1038/s41598-018-25916-5.

[6] Koetzier LR, Mastrodicasa D, Szczykutowicz TP, van der Werf NR, Wang AS, Sandfort V, van der Molen AJ, Fleischmann D, Willemink MJ. Deep learning image reconstruction for CT: technical principles and clinical prospects. *Radiology*. 2023;306:e221257. doi:10.1148/radiol.221257.

[7] Arabi H, Zaidi H. Deep learning-based metal artefact reduction in PET/CT imaging. *European Radiology*. 2021;31:6384–6396. doi:10.1007/s00330-021-07709-z.

[8] Jung Y, Lee H, Jun H, Cho S. Evaluation of motion artifact correction technique for cone-beam computed tomography image considering blood vessel geometry. *Journal of Clinical Medicine*. 2024;13:2253. doi:10.3390/jcm13082253.

[9] Kak AC, Slaney M. *Principles of Computerized Tomographic Imaging*. IEEE Press; 1988.

[10] Gursoy D, De Carlo F, Xiao X, Jacobsen C. TomoPy: a framework for the analysis of synchrotron tomographic data. *Journal of Synchrotron Radiation*. 2014;21:1188–1193. doi:10.1107/S1600577514013939.

[11] Marone F, Stampanoni M. Regridding reconstruction algorithm for real-time tomographic imaging. *Journal of Synchrotron Radiation*. 2012;19:1029–1037. doi:10.1107/S0909049512032864.

[12] Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*. 1987;21(4):163–169. doi:10.1145/37402.37422.

[13] Sullivan CB, Kaszynski A. PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). *Journal of Open Source Software*. 2019;4(37):1450. doi:10.21105/joss.01450.

[14] van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*. 2014;2:e453. doi:10.7717/peerj.453.

[15] Bradley RA, Terry ME. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*. 1952;39:324–345. doi:10.2307/2334029.

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016. p. 770–778. doi:10.1109/CVPR.2016.90.

[17] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations*; 2019.

[18] Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*; 2008. p. 1096–1103. doi:10.1145/1390156.1390294.

[19] Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, Zhou J, Wang G. Low-dose CT with a residual encoder–decoder convolutional neural network. *IEEE Transactions on Medical Imaging*. 2017;36(12):2524–2535. doi:10.1109/TMI.2017.2715284.

[20] Wierstra D, Schaul T, Glasmachers T, Sun Y, Peters J, Schmidhuber J. Natural evolution strategies. *Journal of Machine Learning Research*. 2014;15:949–980.

[21] Münch B, Trtik P, Marone F, Stampanoni M. Stripe and ring artifact removal with combined wavelet–Fourier filtering. *Optics Express*. 2009;17(10):8567–8591. doi:10.1364/OE.17.008567.

[22] Thompson TN, Vickrey A, Shapiro MD, Hsu E. A computer vision framework for quantification of feather growth patterns. *Frontiers in Bioinformatics*. 2023;3:1073918. doi:10.3389/fbinf.2023.1073918.

[23] Kendall MG. A new measure of rank correlation. *Biometrika*. 1938;30:81–93. doi:10.1093/biomet/30.1-2.81.

A Per-seed main benchmark values

Table 7: Per-seed correction metrics. No-motion component count is listed so component-count error can be recomputed. Projection errors are dimensionless minus-log units; reconstruction errors are 1/mm; density-map RMSE uses independently normalized 0–1 maps. Main-text reductions were computed from unrounded analysis outputs; values in this table are rounded for display.

Seed	Condition	No-motion comp.	Proj. RMSE	Proj. MAE	Recon RMSE	Recon MAE	Area err. (%)	Components	Comp. err.	Density RMSE	Density corr.
0	Motion	306	0.0218	0.0073	0.00442	0.00092	12.4	392	86	0.092	0.836
0	Visual baseline	306	0.0171	0.0058	0.00356	0.00071	9.2	373	67	0.070	0.874
0	Random latent	306	0.0138	0.0046	0.00289	0.00058	7.3	360	54	0.055	0.912
0	Preference-guided	306	0.0069	0.0023	0.00144	0.00029	3.8	331	25	0.026	0.960
1	Motion	301	0.0251	0.0085	0.00511	0.00108	15.1	413	112	0.106	0.777
1	Visual baseline	301	0.0186	0.0063	0.00388	0.00079	10.6	383	82	0.076	0.842
1	Random latent	301	0.0149	0.0050	0.00314	0.00064	8.4	367	66	0.060	0.877
1	Preference-guided	301	0.0075	0.0025	0.00157	0.00032	4.3	331	30	0.028	0.940
2	Motion	313	0.0199	0.0067	0.00405	0.00084	11.7	392	79	0.087	0.832
2	Visual baseline	313	0.0161	0.0054	0.00336	0.00067	9.0	376	63	0.069	0.869
2	Random latent	313	0.0130	0.0044	0.00272	0.00055	7.1	364	51	0.054	0.905
2	Preference-guided	313	0.0065	0.0022	0.00136	0.00027	3.7	336	23	0.026	0.954
3	Motion	304	0.0234	0.0079	0.00474	0.00099	13.6	400	96	0.098	0.804
3	Visual baseline	304	0.0177	0.0060	0.00367	0.00074	9.7	376	72	0.072	0.858
3	Random latent	304	0.0142	0.0048	0.00297	0.00060	7.7	361	57	0.057	0.893
3	Preference-guided	304	0.0071	0.0024	0.00149	0.00030	4.0	330	26	0.027	0.948
4	Motion	309	0.0242	0.0081	0.00488	0.00103	14.2	412	103	0.101	0.806
4	Visual baseline	309	0.0186	0.0061	0.00386	0.00078	10.4	387	78	0.076	0.852
4	Random latent	309	0.0150	0.0050	0.00312	0.00064	8.2	372	63	0.060	0.893
4	Preference-guided	309	0.0075	0.0024	0.00156	0.00032	4.3	338	29	0.028	0.948

B No-motion limited-perturbation per-seed values

Table 8: Per-seed no-motion control values. Projection RMSE is dimensionless minus-log error, reconstruction RMSE is in 1/mm, and density-map RMSE is computed on normalized 0–1 density maps.

Seed	Projection RMSE	Recon RMSE	Surface-area change (%)	Component-count change	Density-map RMSE	Density-map corr.
0	0.0018	0.00030	0.39	0	0.009	0.997
1	0.0022	0.00037	0.57	2	0.014	0.994
2	0.0019	0.00031	0.44	1	0.010	0.997
3	0.0020	0.00034	0.49	1	0.011	0.996
4	0.0021	0.00033	0.46	1	0.011	0.996

C Held-out seed summary

Seed 4 was held out from the ranking-model test split and provides a conservative check of the final correction benchmark. For this seed, reconstruction RMSE decreased from 0.00488 in the motion condition to 0.00156 after correction, density-map RMSE decreased from 0.101 to 0.028, and component-count error decreased from 103 to 29 components. This held-out seed follows the same direction as the full five-seed summary.