A Domain-Aligned Encoder Evaluation Framework for FID-like Quality Metrics for Synthetic 3D Medical Imaging

Anonymous authors
Paper under double-blind review

Abstract

Fréchet Inception Distance (FID) is the de facto standard for evaluating synthetic image data. For 2D natural images, FID's validity has been shown to depend strongly on the features deemed salient by the used image encoder. For 3D medical imaging, FID is either omitted due to 2D-3D incompatibility or is implemented with the Med3D image encoder as feature extractor, although this substitution has not been scrutinized. We present a framework that evaluates Med3D and other encoders along multiple criteria with controlled signals, including domain-relevant distortions and synthesis-derived fidelity and empirical-coverage signals, which ranks encoders without requiring expert annotation. Additionally, we conduct an exploratory analysis of alignment with expert visual perception, covering overall, anatomical and vessel quality as well as anatomical and pathological variation. In a use-case on Time-of-Flight Magnetic Resonance Angiography with six candidate encoders, Med3D showed no detectable response to vessel masking or to controlled changes in empirical coverage. Self-supervised ResEncL encoders pretrained on OpenMind responded to most controlled signals, with ResEncL-OM-S3D stronger than Med3D in five of six and ranking first overall. Expert-rating associations were weak for all encoders and did not separate them, reinforcing the need for systematic encoder evaluation. The framework facilitates reliable evaluation of synthetic medical imaging and is designed to transfer to other medical imaging modalities.

1 Introduction

Generative models for 3D medical images have advanced rapidly, with GAN-based (Sun et al., 2022; Aydin et al., 2024) and diffusion-based (Dorjsembe & Xiao, 2022; Pinaya et al., 2022; Khader et al., 2023; Dorjsembe et al., 2024; Wang et al., 2025) methods producing increasingly realistic volumetric Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) images. However, progress in generation has not been matched by equally well-validated methods for evaluating the resulting synthetic image distributions. Reliable quantitative evaluation remains an important open problem, particularly in medical applications where synthetic images must preserve subtle anatomical and pathological structures and reflect diversity with respect to target populations. The gold standard for synthetic image evaluation is still visual rating conducted by domain- and sub-field experts (Khader et al., 2023). Expert rating is, however, expensive and subjective, making it impractical during the development cycle of generative models.

The Fréchet Inception Distance (FID) (Heusel et al., 2017) is the de facto standard for evaluating synthetic image data. FID and related metrics, such as improved precision and recall (Kynkäänniemi et al., 2019), compare distributions of embeddings from a pretrained image encoder. Their behaviour depends critically on the representation learned by the encoder. For example, InceptionV3's embedding space is closely tied to ImageNet classifications (Kynkäänniemi et al., 2022) and may provide inadequate representations of content outside of its training distribution causing FID to disagree with human judgment (Kynkäänniemi et al., 2022; Jayasumana et al., 2024; Stein et al., 2023). Self-supervised encoders do not impose a task-specific objective and produce more consistent evaluation rankings (Morozov et al., 2020), indicating that the reliability of the metric depends on both encoder's training data and its learning objective.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

FID relies on InceptionV3 trained on natural 2D images, which cannot process 3D medical volumes. A common adaptation (Sun et al., 2022; Dorjsembe et al., 2024) replaces it with Med3D (Chen et al., 2019), a 3D ResNet-50 pretrained on heterogeneous MRI and CT segmentation datasets; we refer to FD computed in this feature space as FMD. Whether Med3D's features capture the image characteristics relevant to a given medical domain, however, has not been systematically examined. For instance, in Time-of-Flight MR Angiography (TOF-MRA), vessels only a few voxels wide can determine whether an image is anatomically plausible, and an encoder insensitive to such detail may assign favorable scores to synthetic images with implausible vasculature. This motivates our central question: how can we determine whether an encoder is suitable for evaluating synthetic 3D medical images in a particular domain?

Our contributions are:

- An evaluation framework for assessing encoders behind Fréchet-distance-based evaluation of synthetic 3D medical images based on controlled signals complemented by an exploratory analysis of alignment with expert visual perception.
- Branching Diffusion: a sampling procedure that controls the empirical coverage of finite synthetic datasets without changing the generator's marginal distribution, allowing us to measure how an encoder reacts to coverage loss in isolation.
- On TOF-MRA, we find that the field-standard Med3D encoder does not show a detectable response to vessel masking or controlled changes in empirical coverage.
- To our knowledge, the first systematic evaluation of self-supervised 3D encoders for FD-based evaluation, identifying ResEncL-OM-S3D as the most sensitive encoder across controlled signals.

2 Related Work

FID and related metrics

Several alternatives to FID have been proposed to address its limitations. Improved precision and recall (Kynkäänniemi et al., 2019) separately measure sample quality and distribution coverage, which FID conflates into a single scalar. Density and coverage (Naeem et al., 2020) offer more reliable variants, noting that the target distribution may differ from the encoder's training domain. CMMD (Jayasumana et al., 2024) replaces the Gaussianity assumption with a kernel-based comparison over CLIP embeddings (Radford et al., 2021) and is more sample-efficient. Visualizations using GradCAM (Selvaraju et al., 2017) of the features driving FID scores show that they can be manipulated without genuine quality improvement, exposing vulnerability to encoder biases (Kynkäänniemi et al., 2022).

Perturbation-Based Interpretability

Modifying an input in a controlled manner and observing the change in a model's output is an established means of probing model sensitivity. RISE (Petsiuk et al., 2018) masks image regions at random and attributes importance according to the change in class probability, and OSA-DAS (Valois et al., 2024) extends this from occlusions to general augmentations, analysing the induced changes in deep feature space. Both explain individual predictions through per-input attribution maps. Our feature sensitivity criterion adopts this principle by applying distortions of increasing magnitude and measuring whether a distribution-level metric responds accordingly, without attributing importance to individual features.

3D Medical Image Generation

Generating 3D medical images is challenging due to the high dimensionality and complexity of volumetric data, which demands substantial memory and computation. The field has evolved from GAN-based approaches (Sun et al., 2022) to diffusion models, which produce higher-quality outputs (Dorjsembe & Xiao, 2022; Khader et al., 2023; Dorjsembe et al., 2024; Xu et al., 2024). In particular, latent diffusion (Pinaya et al., 2022; Friedrich et al., 2025) enables high-resolution synthesis on feasible GPU resources, with recent work scaling to very large volumes (Guo et al., 2025; Wang et al., 2025). However, these works focus on the most common anatomical MRI sequences or Non-Contrast CT, where large-scale tissue contrast dominates the signal. TOF-MRA is more sensitive: the most relevant structures of clinical interest are fine vessels only a few voxels wide, which must be preserved faithfully. GANs have been used to generate TOF-MRA image patches (Subramaniam

et al., 2022) and to model the brain area containing the Circle of Willis (Aydin et al., 2024) as a complex and medically relevant structure, alongside diffusion-based image-to-image translation to CT Angiography (CTA) (Koch et al., 2025).

Synthetic 3D Medical Image Evaluation Whether the encoder behind FID-like metrics suits medical images has recently been questioned. A comparison of FDs from eleven ImageNet- and RadImageNet-pretrained 2D extractors across 16 StyleGAN2 models and four modalities against visual Turing tests found that medical pretraining did not improve and could compromise agreement with experts, while the self-supervised SwAV encoder was best correlated (Woodland et al., 2024). An alternative approach replaces learned features with standardized radiomic features (FRD) (Konz et al., 2026). Both works focused on 2D images (slices of volumes in FRD) and relate metrics to quality at the level of models or datasets. For 3D volumetric data, FD is typically computed with Med3D without such validation (Chen et al., 2019; Sun et al., 2022; Dorjsembe et al., 2024), while others rely on expert ratings (Khader et al., 2023), precision/recall and Dice (Dorjsembe et al., 2024), or replica detection (Aydin et al., 2025). Our framework complements these works by auditing fully volumetric encoders, including self-supervised ones, with modality-specific probes (vessel masking), a sampling-based redundancy probe, and per-image association with expert ratings.

3 Background

Encoder-Based Quality Metrics

Encoder-based quality metrics assess the similarity between a real and a synthetic dataset by comparing distributions of image features extracted by a pretrained encoder, rather than operating on pixel intensities directly. We denote an encoder by $\mathcal{E}$; given a real dataset $\mathcal{D}$ and a synthetic dataset $\mathcal{S}$, $\mathcal{E}$ maps both to embedding sets $\mathcal{E}(\mathcal{D}), \mathcal{E}(\mathcal{S}) \in \mathbb{R}^{n \times d}$, and a distributional metric is applied to compare them.

The dominant distributional metric is the Fréchet distance (squared Wasserstein-2 distance, FD), which models each embedding set as a multivariate Gaussian:

$$\mathrm{FD}(\mathcal{D}, \mathcal{S}) = \|\mu_r - \mu_s\|^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_s - 2\left(\Sigma_r \Sigma_s\right)^{1/2}\Big), \tag{1}$$

where (μ_r, Σ_r) and (μ_s, Σ_s) are the mean and covariance of the real and synthetic embeddings.

4 Methods

We propose an evaluation framework that assesses how the encoder's representation determines the behaviour of FD using controlled signals. For each criterion, a signal of known direction is induced in a controlled manner, either by image distortions or by deriving it from the training or generation process, and we test whether the metric responds accordingly. The framework comprises three criteria: the relevance of the encoder's features to the use case (C1), sensitivity to fidelity (C2), and sensitivity to empirical coverage (C3), where empirical coverage denotes the extent to which a finite synthetic dataset represents the variability of the generator's output. The criteria require no expert annotation and are aggregated into an encoder ranking (section 4.2). Responding to these signals is a necessary but not sufficient condition for clinical validity. We therefore additionally explore the alignment of per-image FD leverage with expert visual perception (section 4.3); this analysis is reported descriptively and not included in the ranking. We apply the framework to compare Med3D's representation against five alternative 3D medical image encoders (table 3).

4.1 Evaluation with Controlled Signals

4.1.1 Feature Sensitivity via Image Distortions

Feature sensitivity (C1) assesses whether the encoder has learned features relevant to the use case. Our method follows the logic of perturbation-based interpretability discussed in section 2, in a weaker form. Rather than attributing importance to individual features, we use the magnitude of an applied distortion as the controlled signal and test whether the metric correlates with it. A metric that does not respond to the selected distortions suggests that its encoder does not adequately capture the corresponding features.

We apply a distortion Φ at varying magnitudes α_i to each $x_i \in \mathcal{D} = \{x_i\}_{i=1}^N$, producing $\mathcal{S} = \{s_i = \Phi(x_i, \alpha_i)\}_{i=1}^N$. To estimate sensitivity, we compute per-image influences via jackknife sampling. For each $s_j \in \mathcal{S}$, the influence of image s_j on FD is

$$d_j = \text{FD}(\mathcal{E}(\mathcal{D}),\, \mathcal{E}(\mathcal{S} \setminus \{s_j\})) \tag{2}$$

Since FD increases as distributions diverge, we expect a negative Spearman's ρ between $\{\alpha_j\}$ and $\{d_j\}$, with stronger values indicating a better encoder. 95% percentile bootstrap CIs account for sampling uncertainty over the finite set of N images and their distortion magnitudes. However, the calculation of the jackknife influences overlaps in $N-2$ of $N-1$ distorted samples and is therefore not independent, making nominal coverage approximate. Embeddings are precomputed once and reused for each per-image FD-influence computation. The full procedure is described in algorithm 1.

4.1.2 Fidelity Sensitivity via Training Epochs

Our second criterion (C2) assesses whether the encoder is sensitive to the fidelity of synthetic images. Training a generative model for longer typically yields images of higher fidelity, so we use the training epoch as a controlled, monotone fidelity signal. We expect a negative correlation between epoch and metric value, with stronger values indicating that the encoder is more sensitive to fidelity changes. This assumes monotone improvement over training, which must be verified for the model at hand.

We denote the set of K training checkpoints as $\{C_k\}_{k=1}^K$ with corresponding epochs $\{t_k\}_{k=1}^K$. We generate M synthetic images per checkpoint C_k, forming $\mathcal{S}_k = \{s_i\}_{i=1}^M$. For each checkpoint, the influence of the synthetic data on FD is

$$d_k = \text{FD}(\mathcal{E}(\mathcal{D}),\, \mathcal{E}(\mathcal{D}' \cup \mathcal{S}_k)) \tag{3}$$

where $\mathcal{D}' \subset \mathcal{D}$ results from removing M images from $\mathcal{D} = \{x_i\}_{i=1}^N$, with $|\mathcal{D}'| = N - M$. We correlate $\{d_k\}$ with $\{t_k\}$ via Spearman's ρ, expecting a negative correlation. The procedure is repeated over R_{gen} generation seeds. The final score is the Fisher z-mean $\bar{\rho}$ across seeds, with σ_z quantifying variability across seeds and p_{strat} testing the significance of the correlation via a stratified permutation test (algorithm 3). Checkpoints are neighbors in training history, so the d_k may not be fully independent. Only M new embeddings are needed per checkpoint rather than recomputing the full dataset. The full procedure is described in algorithm 2.

4.1.3 Empirical Coverage Sensitivity via Branching Diffusion

We propose Branching Diffusion to control the empirical coverage of a diffusion generated synthetic dataset. Let $x_T \sim \mathcal{N}(0, I)$ denote the initial noise and T the total number of denoising steps, where $t = T$ corresponds to noise and $t = 0$ to the clean image. We denote by b the timestep at which we branch: we produce a prototype by denoising x_T from $t = T$ to $t = b$, then produce V variations by running as many independent denoising trajectories from $t = b$ to $t = 0$. This requires a stochastic sampler. The marginal distribution for every variation is the model's standard output distribution and independent of b:

$$p_{\theta,b}(x_0^{(v)}) = \int p_\theta(x_b)\, p_\theta(x_0^{(v)} \mid x_b)\, dx_b = p_\theta(x_0^{(v)}) \tag{4}$$

The joint distribution of the variations is thus

$$p_{\theta,b}\Big(x_0^{(1)}, \ldots, x_0^{(V)}\Big) = \int p_\theta(x_b) \prod_{v=1}^{V} p_\theta\Big(x_0^{(v)} \mid x_b\Big)\, dx_b \tag{5}$$

which depends on b due to the shared state x_b. Branching Diffusion therefore controls the empirical coverage of a finite synthetic dataset while leaving the generator's output distribution unchanged. The experiment therefore tests whether the encoder reacts to reduced coverage, not how well the generator matches the real data.

A suitable encoder should capture the image characteristics that vary between independently generated samples. With late branching, these characteristics become increasingly similar among variations of the same prototype, causing them to cluster. Consequently, a sensitive encoder is expected

to assign higher FD values to such clustered datasets than to one with broader empirical coverage of the reference distribution. A negative correlation is a necessary condition for an encoder to detect coverage deficits of a synthetic dataset with respect to real data.

For the empirical coverage sensitivity criterion (C3), we denote the set of K branching points by $\{b_k\}_{k=1}^{K}$. For each b_k, we generate P prototypes and derive V variations per prototype, forming $\mathcal{S}_k = \{s_i\}_{i=1}^{PV}$. We compute d_k following eq. (3), where $\mathcal{D}' \subset \mathcal{D}$ results from removing PV images from $\mathcal{D} = \{x_i\}_{i=1}^{N}$, with $|\mathcal{D}'| = N - PV$. Since d_k is computed against the disjoint real subset D', any variation across b_k is attributable to the coverage of the real reference distribution achieved by the synthetic sample, and not to changes in sample fidelity or model bias. We correlate $\{d_k\}$ with $\{b_k\}$ via Spearman's ρ and expect a negative correlation, with stronger values indicating greater sensitivity. The procedure is repeated over R_{gen} generation seeds. The final score is the Fisher z-mean $\bar{\rho}$ across seeds, with σ_z quantifying variability across seeds and p_{strat} testing the significance of the correlation via a stratified permutation test (algorithm 3). Only PV new embeddings are needed per branching point rather than recomputing the complete dataset. The full procedure is described in algorithm 4.

4.2 Encoder Ranking

We aggregate the three controlled-signal criteria (C1, C2, C3) into a descriptive ranking based on pairwise comparisons (Demšar, 2006). Each criterion C_k decomposes into one or more leaves, the finest granularity at which encoders are compared pairwise. For C1, each distortion type is a leaf, while C2 and C3 are single-leaf. For each leaf ℓ, all $\binom{U}{2}$ encoder pairs are evaluated (U = number of candidate encoders). A pair is tied if zero falls within the interval of the per-replicate correlation difference $\rho_i - \rho_j$ (95% percentile bootstrap CI for C1, or the observed range across generation seeds for C2, C3); tied pairs are ignored. For each non-tied pair, the encoder with the more favorable correlation wins. The per-leaf score is $s_\ell^{(i)} = \text{wins}_\ell^{(i)} - \text{losses}_\ell^{(i)}$, averaged across leaves within each criterion. The overall encoder score is:

$$S^{(i)} = \sum_{k=1}^{K} \bar{s}_{C_k}^{(i)}, \tag{6}$$

where $K = 3$ and $\bar{s}_{C_k}^{(i)}$ is encoder i's average leaf score for criterion C_k. Based on the pairwise comparisons, we also derive a partial order of encoders, resulting in significance groups displayed as compact letter display (CLD) in tables and figures. Encoders sharing a letter are not significantly different.

4.3 Exploratory Analysis via Expert Visual Rating

In addition to the controlled-signal framework, we explore whether per-image FD leverage is associated with expert-rated quality and annotated anatomical and pathological variation. Unlike controlled signals, expert ratings vary along many uncontrolled axes and are subject to inter-rater disagreement; we therefore report these analyses descriptively and exclude them from the overall ranking. The detailed rating protocol can be found in appendix A.5.

4.3.1 Perceived Visual Quality

We replace the distortion magnitude α in section 4.1.1 by the expert quality ratings, hence the correlation can be interpreted as the encoder's ability to capture quality aspects of synthetic images relevant for clinical image interpretation, namely overall, anatomical and vessel quality.

Let $\mathcal{S} = \{(s_j, y_j)\}_{j=1}^{M}$ denote the fixed set of $M = 100$ synthetic images s_j with ratings y_j on the ordinal quality aspects (Likert 1–5). We use jackknife sampling to compute the metric

$$d_j = \text{FD}(\mathcal{E}(\mathcal{D}),\, \mathcal{E}(\mathcal{D}' \cup (\mathcal{S} \setminus \{s_j\}))) \tag{7}$$

for each rated image s_j, where $\mathcal{D}' \subset \mathcal{D}$ results from removing $M - 1$ images from $\mathcal{D} = \{x_i\}_{i=1}^{N}$, with $|\mathcal{D}'| = N - M + 1$. We compute the rank correlation between y_j and d_j across the M rated images, using Kendall's τ_b to accommodate the heavy ties inherent to a five-level ordinal scale. The resulting correlation is sign-adjusted (negated) so that more negative values indicate stronger

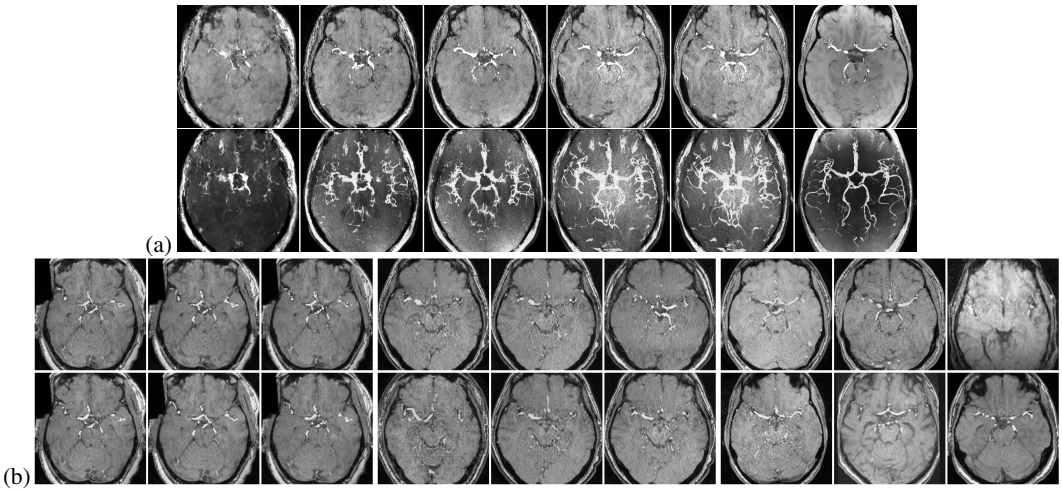


Figure 1: Synthetic TOF-MRA. (a) Monotone fidelity improvement over training epochs, top: axial slices, bottom: Maximum Intensity Projections (MIP). One sample generated from a fixed seed per checkpoint (10, 50, 100, 300, 600, 1000 epochs, left to right); more examples in fig. 6 and fig. 7. (b) Variations created from prototypes with branching point at $b = 700$, $b = 850$, and $b = 999$ (left to right), each mosaic shows six variations; more examples in fig. 8. Axial slices show arteries of the Circle of Willis.

sensitivity, consistent with the controlled-signal criteria. Because the d_j are deterministic given the fixed pool and the rated set is the entire labelled pool, uncertainty is obtained by bootstrapping over the (d_j, y_j) pairs, resampled with replacement as units. Confidence intervals are reported as 95% bootstrap percentiles. The three visual quality aspects are analysed individually. The full procedure is described in algorithm 5.

4.3.2 Association with Annotated Vessel Topology and Pathology Variations

We further explore whether per-image FD leverage is associated with anatomical and pathological variation annotated by the raters. For each synthetic image $s_j \in S = \{s_j\}_{j=1}^{M}$, we define a per-image variation load

$$v_j = \sum_{r=1}^{R} |V_{rj}|, \tag{8}$$

where V_{rj} is the set of vessel-topology and pathology variations flagged for image s_j by rater r out of R raters. v_j aggregates flags additively across raters rather than thresholding them into a consensus label, to remain well-defined even for variation classes that are individually scarce. We use the jackknife FD d_j from eq. (7), compute Kendall's τ_b between v_j and d_j, and quantify uncertainty by bootstrapping. The correlation is sign-adjusted as in section 4.3.1. The full procedure is described in algorithm 6.

5 Experiments

Our experimental results in full detail can be found in appendix C.

5.1 Experimental Setup

Data and Pre-processing We used TOF-MRA scans from seven public datasets spanning healthy and pathological populations to train a generative model (table 2). All scans were registered to a common TOF-MRA template (Aydin et al., 2024) and cropped to $256 \times 256 \times 64$ volumes centered on the Circle of Willis, resulting in 3,304 3D volumes. Further details about the datasets and pre-processing pipeline are provided in appendix A.1.

Synthetic Data We generated synthetic images by training 3D-MedDiffusion (Wang et al., 2025) from scratch on the TOF-MRA dataset described above. 3D-MedDiffusion is a latent diffusion model combining a PatchVolume autoencoder with a BiFlowNet noise estimator. Training details are provided in appendix A.2.

FD Encoders We evaluated six encoders: Med3D (Chen et al., 2019), our baseline and the most widely used encoder for FD on 3D medical images, and five encoders trained with self-supervision methods (PrimusM-OM-MAE, PrimusM-OM-VoCo, ResEncL-OM-MAE, ResEncL-OM-S3D, and ResEncL-OM-VoCo), the best-performing on downstream classification and segmentation tasks in the OpenMind benchmark (Wald et al., 2025c) with publicly available pretrained checkpoints, including two architectures (PrimusM (Wald et al., 2025a), ResEncL (Isensee et al., 2024)) and three pre-training objectives (MAE (He et al., 2022), VoCo (Wu et al., 2024), S3D (Wald et al., 2025b)). Encoders and their adaptation for feature extraction are described in more detail in appendix A.4.

Clinical ratings Expert visual ratings were created following the rating protocol detailed in appendix A.5 by four clinical raters: one medical doctor with research experience (Rater 3), one board-certified neurosurgeon (Rater 2), and two board-certified neuroradiologists (Rater 1 and 4). Ratings showed low inter-rater reliability and absolute agreement (Krippendorff's $\alpha_{\text{ord}} \leq 0.01$; ICC(2,1)$\leq$ 0.15, table 4), whereas average-measure consistency was moderate (ICC(3,4) 0.58–0.70), suggesting raters partly shared the ordering of images but used different ranges of the Likert scale (Koo & Li, 2016). We therefore removed rater-specific offsets by bias-correction before aggregation described in appendix A.6.

Experiment details and parameters For the feature sensitivity criterion (section 4.1.1), we use four distortion types: Gaussian noise, elastic deformation, random masking, and vessel masking, the latter specific to TOF-MRA (appendix B.1). We use $R = 10000$ bootstrap resamples for uncertainty estimation. For the fidelity sensitivity criterion (section 4.1.2), we use $K = 100$ checkpoints from the LDM training, saved every 10 epochs, synthesizing $M = 100$ images per checkpoint across $R_{\text{gen}} = 10$ generation seeds. Monotone improvement was verified by visual inspection (fig. 1). For the empirical coverage sensitivity criterion (section 4.1.3), we use $K = 31$ branching points $b_k \in \{999, 990, 980, \ldots, 700\}$, generating $P = 10$ prototypes with $V = 10$ variations each ($P \cdot V = 100$ images per b_k) across $R_{\text{gen}} = 10$ seeds (fig. 1). For the analysis of expert visual perception (section 4.3.1), ratings were bias-corrected and averaged (appendix A.6). We report the aggregate across all raters, the aggregate of the two neuroradiologists (Raters 1 and 4), and each rater individually. For the analysis of annotated anatomical and pathological variation (section 4.3.2), the same $M = 100$ images were used. Raters flagged vessel-topology and pathology variations per image; the per-image variation load v_j is the total flag count across all raters.

5.2 Evaluation with Controlled Signals

Feature Sensitivity (C1). Figure 2a reports feature sensitivity in terms of Spearman's ρ between magnitude of image distortions and FD metric using the candidate encoders. In general, most encoders showed strong correlation with Gaussian noise and random masking but achieved lower sensitivity towards elastic deformation and vessel masking. Med3D showed no detectable association with vessel masking ($-0.007\,[-0.040, 0.026]$; table 5). ResEncL-OM encoders registered the strongest vessel removal association (ResEncL-OM-MAE: -0.293, ResEncL-OM-S3D: -0.285, ResEncL-OM-VoCo: -0.231; table 5), while ResEncL-OM-S3D showed dominance across most distortions.

Fidelity Sensitivity (C2). Med3D and ResEncL-OM-S3D showed the strongest correlation between FD and training epochs ($\rho = -0.738, -0.699$, seed range $[-0.874, -0.247], [-0.843, -0.265]$, respectively; table 7) and were not separable (shared CLD group; table 8); PrimusM-OM-VoCo and ResEncL-OM-VoCo showed no significant association ($p_{\text{strat}} = 0.306, 0.109$; table 7).

Empirical Coverage Sensitivity (C3). ResEncL-OM-S3D, -VoCo and -MAE responded strongly to the branching point ($\rho = -0.883, -0.862, -0.725$; $p_{\text{strat}} < 10^{-4}$, fig. 2c; table 9), whereas Med3D (0.027, $p_{\text{strat}} = 0.666$; table 9) and both PrimusM encoders did not. Two-dimensional PCA projections of the encoder embeddings in fig. 9 and fig. 10 illustrate that for ResEncL-OM-S3D, variations of a prototype cluster under late branching (b=700) but already exhibit larger spread starting from earlier branching (b=850, 999), while Med3D embeddings change little with b.

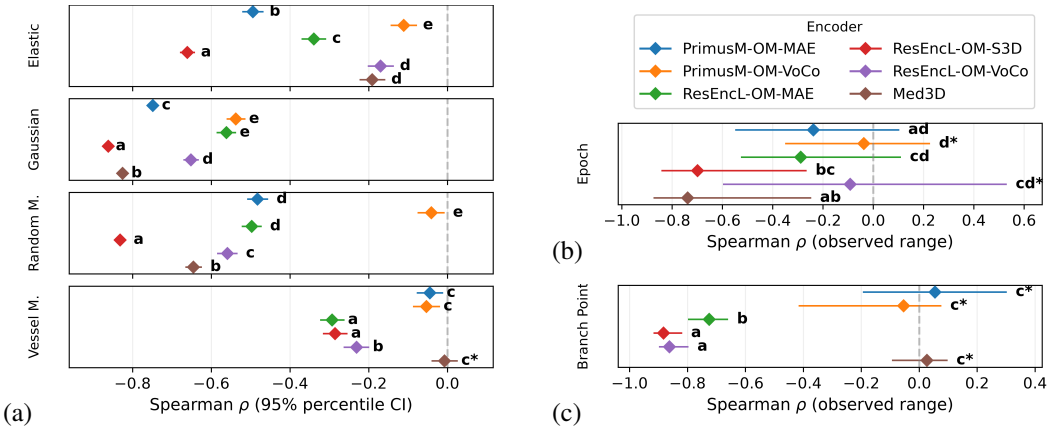


Figure 2: (a) Feature sensitivity (C1), 95% percentile bootstrap CIs. (b) Fidelity sensitivity (C2) and (c) Empirical coverage sensitivity (C3), observed range across 10 generation seeds; more negative values indicate stronger sensitivity. $*$ marks no significant correlation: (a) bootstrap CI containing zero, (b, c) $p_{strat} \geq 0.05$ in stratified permutation test. CLD labels are shown per encoder.

Table 1: Encoder ranking over the controlled-signal criteria. C1: feature sensitivity. C2: fidelity sensitivity. C3: empirical coverage sensitivity. Total: sum of per-criterion scores. Parenthetical counts ($n_{\text{sig}}/n_{\text{total}}$): statistically significant leaves per criterion (i.e. 95% percentile bootstrap CI excludes zero for C1; $p_{\text{strat}} < 0.05$ for C2, C3).

Encoder	C1	C2	C3	Total	Rank
ResEncL-OM-S3D	4.75 (4/4)	2 (1/1)	4 (1/1)	10.75	1
ResEncL-OM-VoCo	−0.25 (4/4)	−1 (0/1)	4 (1/1)	2.75	2
Med3D	0.25 (3/4)	3 (1/1)	−3 (0/1)	0.25	3
ResEncL-OM-MAE	−0.25 (4/4)	−1 (1/1)	1 (1/1)	−0.25	4
PrimusM-OM-MAE	−0.25 (4/4)	−1 (1/1)	−3 (0/1)	−4.25	5
PrimusM-OM-VoCo	−4.25 (4/4)	−2 (0/1)	−3 (0/1)	−9.25	6

5.3 Encoder Ranking

Table 1 aggregates the three controlled-signal criteria following the method in section 4.2. ResEncL-OM-S3D ranked first overall (10.75), with the highest scores on feature (C1 $= 4.75$) and empirical coverage sensitivity (C3 $= 4.0$). The field-standard Med3D ranked only third (0.25), with a negative score on empirical coverage sensitivity (C3 $= -3.0$), while the two PrimusM encoders ranked last.

5.4 Expert Visual Rating

Perceived visual quality. Associations between bias-corrected ratings and per-image FD leverage were weak for all encoders ($|\tau_b| \leq 0.26$; fig. 3a,b, table 12), and pairwise comparisons did not separate any encoders (single CLD group on every axis; table 13). Individual significance depended on rater aggregation: with all raters, every encoder except ResEncL-OM-VoCo had intervals excluding zero on all three aspects; with Raters 1 and 4, only ResEncL-OM-S3D did ($\tau_b = -0.178, -0.209, -0.166$; table 12). Expert-rated quality therefore did not discriminate between encoders in our study.

Vessel topology and pathology variations. τ_b between per-image variation load and FD leverage was weak and inconsistent in sign (-0.099 to 0.131, fig. 3c, table 14), with all intervals covering zero; we did not detect an association for any encoder. Because variation load counts findings per image rather than set-level coverage, we regard this analysis as exploratory.

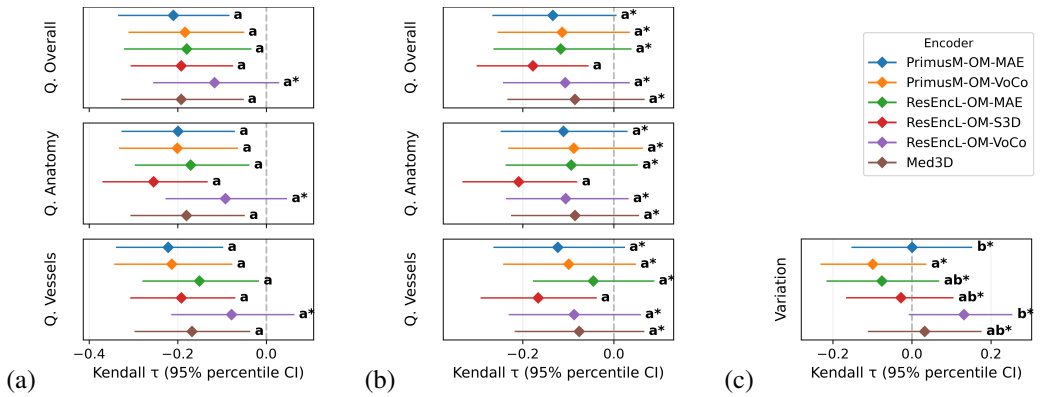


Figure 3: Exploratory analysis via expert rating. (a,b) Perceived visual quality aggregated across all raters (a) and aggregated from Raters 1 and 4 (b). (c) Vessel topology and pathology variations. 95% percentile bootstrap CIs shown. More negative values indicate stronger agreement. $*$ marks no significant correlation i.e. bootstrap CI containing zero. CLD labels are shown per encoder.

6 Conclusion

Discussion. ResEncL-OM-S3D ranked first or tied-first on C1 and on C3, and was tied with Med3D on C2, i.e. stronger than Med3D in five of six controlled signals. Sensitivity to global perturbations did not imply sensitivity to vessels: Med3D responded strongly to Gaussian noise but showed no detectable association with vessel masking. ResEncL encoders separated from PrimusM encoders on vessel masking and branching, although architecture is confounded with embedding dimension (320 vs. 864) and pooling. Modality-specific signals such as vessel masking are inexpensive and require no annotation, and could therefore precede costly expert evaluation. Expert perception ultimately defines the quality of synthetic medical images but is difficult to measure. Absolute inter-rater agreement was near floor on all three quality axes, and consistency was moderate only for the four-rater average; raters also differed markedly in scale use, with the least experienced rater scoring on average 1.0 and 2.67 points higher than the two neuroradiologists. Judging vascular realism in TOF-MRA is intrinsically demanding, as plausibility hinges on vessels only a few voxels wide within highly variable cerebrovascular anatomy (Yang et al., 2025). Weak, aggregation-dependent associations with reader evaluation have also been reported for FID-like metrics on 2D medical images (Woodland et al., 2024; Konz et al., 2026). Finally, angiographic data are scarce in encoder pretraining (0.43% of OpenMind; none for Med3D). Although ResEncL-OM encoders responded to vessel masking (which alters far fewer voxels than random masking), encoder pretraining on angiographic data may improve reliability of FID-like evaluation of TOF-MRA.

Limitations. We studied a single modality and generative model. Our controlled signals test necessary rather than sufficient conditions: fidelity improvement over training was verified only visually, and branching varies empirical coverage rather than the generator's output diversity. To keep covariance estimates stable with $M = 100$, synthetic sets were mixed into the real reference set, probing local rather than conventional real-versus-synthetic FD sensitivity. Radiomic alternatives to encoder-based FD (e.g., FRD) were not included. The expert visual assessment (100 images, four raters, low agreement) had limited power to separate encoders.

Conclusion. We proposed an evaluation framework that assesses encoders for FD-based evaluation of synthetic 3D medical images using controlled signals. Across six encoder-FD pipelines on TOF-MRA, Med3D did not respond to vessel masking or to changes in empirical coverage, whereas ResEncL-based representations did, with ResEncL-OM-S3D ranking first. An additional, exploratory analysis found weak associations with expert-rated quality for all encoders that did not establish between-encoder superiority, and association with annotated anatomical and pathological variation remained inconclusive. These findings support testing encoders on domain-specific controlled signals before adopting them for synthetic-image evaluation. Our framework is designed to transfer to other medical imaging modalities and could facilitate reliable evaluation of synthetic medical imaging preceding costly expert-based validation.

AI USE STATEMENT

In this work, we used generative AI tools (Claude (Anthropic), GLM 5.2 (Z.AI) via OpenCode) to provide feedback on the research methodology and on the interpretation of results during manuscript preparation, including critical review of the statistical analysis, the framing of the encoder ranking and the expert-rating analysis, and robustness checks of the reported rankings.

We used AI coding assistants to create and edit software code for the experiments and analyses.

We did not use generative AI tools to generate synthetic data, formulate hypotheses, clean or reformat datasets, or analyse the expert ratings; other tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools to draft parts of the initial manuscript, to improve language and readability, and to identify relevant literature. All AI-assisted content was reviewed by the authors: suggested methodological changes and interpretations were assessed against our own results, all reported numbers were verified against the experiment outputs, all cited works were checked against the original sources and AI-based literature search was complemented with tools searching indexed scientific research (e.g. Google Scholar). We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All TOF-MRA data are publicly available (table 2); preprocessing, registration to a public template (Aydin et al., 2024), and filtering criteria are described in Appendix A.1. The generative model is the publicly available 3D-MedDiffusion (Wang et al., 2025), trained with the configuration given in Appendix A.2. All six encoders use publicly released pretrained checkpoints; their adaptation for feature extraction is detailed in Appendix A.4. Distortion parameterizations are specified in Appendix B.1, and pseudocode for all sensitivity estimates, the stratified permutation test, and the expert-rating analyses is given in Algorithms 1–6. Numbers of checkpoints, branching points, generation seeds, and bootstrap resamples are reported in Section 5.1, and the rating protocol and bias correction in Appendices A.5 and A.6. Code for all experiments and analyses is provided as anonymized supplementary material.

ETHICS STATEMENT

This work uses only publicly available, de-identified TOF-MRA datasets (CASILAB, ICBM, IXI, Lausanne, OASIS-3, RSNA Intracranial Aneurysm Detection, TopCoW), accessed and used in accordance with their respective licenses and data use agreements; no new patient data were acquired. All real and synthetic images were stored and processed exclusively on institutional systems, including during the expert rating study. The rating study involved four clinicians who contributed professional quality judgments of synthetic images as study collaborators; no personal, health, or behavioural data about the raters were collected, and no patient data were involved. Synthetic volumes covered only a restricted region around the Circle of Willis, limiting re-identification risk. The authors declare no competing interests.

REFERENCES

Orhun Utku Aydin, Adam Hilbert, Alexander Koch, Felix Lohrke, Jana Rieger, Satoru Tanioka, and Dietmar Frey. Generative modeling of the Circle of Willis using 3D-StyleGAN. *NeuroImage*, 304:120936, December 2024. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2024.120936.

Orhun Utku Aydin, Alexander Koch, Adam Hilbert, Jana Rieger, Felix Lohrke, Fujimaro Ishida, Satoru Tanioka, and Dietmar Frey. RELICT: A Replica Detection Framework for Medical Image Generation, February 2025.

Imperial College London Biomedical Image Analysis Group. IXI Dataset.

Elizabeth Bullitt, Donglin Zeng, Guido Gerig, Stephen Aylward, Sarang Joshi, J. Keith Smith, Weili Lin, and Matthew G. Ewend. Vessel Tortuosity and Brain Tumor Malignancy: A Blinded Study1.

Academic Radiology, 12(10):1232–1240, October 2005. ISSN 1076-6332, 1878-4046. doi: 10.1016/j.acra.2005.05.027.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer Learning for 3D Medical Image Analysis, July 2019.

J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, December 2006.

Tommaso Di Noto, Guillaume Marie, Sebastien Tourbier, Yasser Alemán-Gómez, Oscar Esteban, Guillaume Saliou, Meritxell Bach Cuadra, Patric Hagmann, and Jonas Richiardi. Towards Automated Brain Aneurysm Detection in TOF-MRA: Open Data, Weak Labels, and Anatomical Knowledge. *Neuroinformatics*, 21(1):21–34, January 2023. ISSN 1559-0089. doi: 10.1007/s12021-022-09597-0.

Zolnamar Dorjsembe and Furen Xiao. Three-Dimensional Medical Image Synthesis with Denoising Diffusion Probabilistic Models. 2022.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, July 2024. ISSN 2168-2208. doi: 10.1109/JBHI.2024.3385504.

Paul Friedrich, Julia Wolleb, Florentin Bieder, Alicia Durrer, and Philippe C. Cattin. WDM: 3D Wavelet Diffusion Models for High-Resolution Medical Image Synthesis. In Anirban Mukhopadhyay, Ilkay Oksuz, Sandy Engelhardt, Dorit Mehrof, and Yixuan Yuan (eds.), *Deep Generative Models*, pp. 11–21, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72744-3. doi: 10.1007/978-3-031-72744-3_2.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. MAISI: Medical AI for Synthetic Imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4430–4441, February 2025. doi: 10.1109/WACV61041.2025.00435.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

M. Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Neural Information Processing Systems*, June 2017.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation, July 2024.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.00889.

Jeff Rudie, Evan Calabrese, Robyn Ball, Peter Chang, Rennie Chen, Errol Colak, Maria Correia de Verdier, Luciano Prevedello, Tyler Richards, Rachit Saluja, Greg Zaharchuk, Jason Sho, and Maryam Vazirabad. RSNA 2025 Intracranial Aneurysm Detection, 2025.

Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation, January 2023.

Alexander Koch, Orhun Utku Aydin, Adam Hilbert, Jana Rieger, Satoru Tanioka, Fujimaro Ishida, and Dietmar Frey. Cross-modality image synthesis from TOF-MRA to CTA using diffusion-based models. *Medical Image Analysis*, 105:103722, October 2025. ISSN 1361-8415. doi: 10.1016/j.media.2025.103722.

Nicholas Konz, Richard Osuala, Preeti Verma, Yuwen Chen, Hanxue Gu, Haoyu Dong, Yaqian Chen, Andrew Marshall, Lidia Garrucho, Kaisar Kushibar, Daniel M. Lang, Gene S. Kim, Lars J. Grimm, John M. Lewin, James S. Duncan, Julia A. Schnabel, Oliver Diaz, Karim Lekadir, and Maciej A. Mazurowski. Fréchet radiomic distance (FRD): A versatile metric for comparing medical imaging datasets. *Medical Image Analysis*, 110:103943, May 2026. ISSN 1361-8415. doi: 10.1016/j.media.2026.103943.

Terry K. Koo and Mae Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163, June 2016. ISSN 1556-3707. doi: 10.1016/j.jcm.2016.02.012.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models, October 2019.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Fréchet Inception Distance. 2022. doi: 10.48550/ARXIV.2203.06026.

Pamela J. LaMontagne, Tammie LS Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease, December 2019. ISSN 1901-4902.

J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 356(1412):1293–1322, August 2001. ISSN 0962-8436. doi: 10.1098/rstb.2001.0915.

Stanislav Morozov, Andrey Voynov, and Artem Babenko. On Self-Supervised Image Representations for GAN Evaluation. In *International Conference on Learning Representations*, October 2020.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable Fidelity and Diversity Metrics for Generative Models, June 2020.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, September 2018.

Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain Imaging Generation with Latent Diffusion Models, September 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. https://arxiv.org/abs/2306.04675v2, June 2023.

Pooja Subramaniam, Tabea Kossen, Kerstin Ritter, Anja Hennemuth, Kristian Hildebrand, Adam Hilbert, Jan Sobesky, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey, and Vince I. Madai. Generating 3D TOF-MRA volumes and segmentation labels using generative adversarial networks. *Medical Image Analysis*, 78:102396, May 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102396.

Li Sun, Junxiang Chen, Yanwu Xu, Mingming Gong, Ke Yu, and Kayhan Batmanghelich. Hierarchical Amortized GAN for 3D High Resolution Medical Image Synthesis. *IEEE journal of biomedical and health informatics*, 26(8):3966, August 2022. doi: 10.1109/JBHI.2022.3172976.

Pedro H. V. Valois, Koichiro Niinuma, and Kazuhiro Fukui. Occlusion Sensitivity Analysis With Augmentation Subspace Perturbation in Deep Feature Space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4829–4838, 2024.

Tassilo Wald, Saikat Roy, Fabian Isensee, Constantin Ulrich, Sebastian Ziegler, Dasha Trofimova, Raphael Stock, Michael Baumgartner, Gregor Köhler, and Klaus Maier-Hein. Primus: Enforcing Attention Usage for 3D Medical Image Segmentation, March 2025a.

Tassilo Wald, Constantin Ulrich, Stanislav Lukyanenko, Andrei Goncharov, Alberto Paderno, Maximilian Miller, Leander Maerkisch, Paul F. Jäger, and Klaus Maier-Hein. Revisiting MAE pretraining for 3D medical image segmentation, April 2025b.

Tassilo Wald, Constantin Ulrich, Jonathan Suprijadi, Sebastian Ziegler, Michal Nohel, Robin Peretzke, Gregor Köhler, and Klaus H. Maier-Hein. An OpenMind for 3D medical vision self-supervised learning, April 2025c.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3D MedDiffusion: A 3D Medical Latent Diffusion Model for Controllable and High-quality Medical Image Generation, November 2025.

Bastian Wittmann, Yannick Wattenberg, Tamaz Amiranashvili, Suprosanna Shit, and Bjoern Menze. vesselFM: A Foundation Model for Universal 3D Blood Vessel Segmentation, March 2025.

McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P. Yung, Ankit B. Patel, and Kristy K. Brock. Feature Extraction for Generative Medical Imaging Evaluation: New Evidence Against an Evolving Trend. volume 15012, pp. 87–97. 2024. doi: 10.1007/978-3-031-72390-2_9.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. VoCo: A Simple-yet-Effective Volume Contrastive Learning Framework for 3D Medical Image Analysis, April 2024.

Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afrooz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich. MedSyn: Text-Guided Anatomy-Aware Synthesis of High-Fidelity 3-D CT Images. *IEEE Transactions on Medical Imaging*, 43 (10):3648–3660, October 2024. ISSN 1558-254X. doi: 10.1109/TMI.2024.3415032.

Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C. Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, Suprosanna Shit, Houjing Huang, Chinmay Prabhakar, Ezequiel de la Rosa, Bastian Wittmann, Diana Waldmannstetter, Florian Kofler, Fernando Navarro, Martin Menten, Ivan Ezhov, Daniel Rueckert, Iris N. Vos, Ynte M. Ruigrok, Birgitta K. Velthuis, Hugo J. Kuijf, Pengcheng Shi, Wei Liu, Ting Ma, Maximilian R. Rokuss, Yannick Kirchhoff, Fabian Isensee, Klaus Maier-Hein, Chengcheng Zhu, Huilin Zhao, Philippe Bijlenga, Julien Hämmerli, Catherine Wurster, Laura Westphal, Jeroen Bisschop, Elisa Colombo, Hakim Baazaoui, Hannah-Lea Handelsmann, Andrew Makmur, James Hallinan, Amrish Soundararajan, Bene Wiestler, Jan S. Kirschke, Roland Wiest, Emmanuel Montagnon, Laurent Letourneau-Guillon, Kwanseok Oh, Dahye Lee, Adam Hilbert, Orhun Utku Aydin, Dimitrios Rallios, Jana Rieger, Satoru Tanioka, Alexander Koch, Dietmar Frey, Abdul Qayyum, Moona Mazher, Steven Niederer, Nico Disch, Julius Holzschuh, Dominic LaBella, Francesco Galati, Daniele Falcetta, Maria A. Zuluaga, Chaolong Lin, Haoran Zhao, Zehan Zhang, Minghui Zhang, Xin You, Hanxiao Zhang, Guang-Zhong Yang, Yun Gu, Sinyoung Ra, Jongyun Hwang, Hyunjin Park, Junqiang Chen, Marek Wodzinski, Henning Müller, Nesrin Mansouri, Florent Autrusseau, Cansu Yalçin, Rachika E. Hamadache, Clara Lisazo, Joaquim Salvi, Adrià

Casamitjana, Xavier Lladó, Uma Maria Lal-Trehan Estrada, Valeriia Abramova, Luca Giancardo, Arnau Oliver, Paula Casademunt, Adrian Galdran, Matteo Delucchi, Jialu Liu, Haibin Huang, Yue Cui, Zehang Lin, Yusheng Liu, Shunzhi Zhu, Tatsat R. Patel, Adnan H. Siddiqui, Vincent M. Tutino, Maysam Orouskhani, Huayu Wang, Mahmud Mossa-Basha, Yuki Sato, Sven Hirsch, Susanne Wegener, and Bjoern Menze. Benchmarking the CoW with the TopCoW Challenge: Topology-Aware Anatomical Segmentation of the Circle of Willis for CTA and MRA, July 2025.

A Dataset, Generative Model and Exploratory Visual Rating Details

A.1 Original Data and Preprocessing

We used TOF-MRA scans from seven datasets as listed in table 2. For preprocessing, we registered all scans on the custom TOF-MRA template from (Aydin et al., 2024) resulting in scans of $288 \times 320 \times 199$ (width $\times$ height $\times$ depth) and voxel spacing of $0.62 \times 0.62 \times 0.62$. To create the training dataset for the generative models, volumes of $256 \times 256 \times 64$ were cropped from the images, centered on width and height, and the image slices were selected so that the Circle of Willis is contained. We then filtered out volumes that contained less than 60% of voxels with intensities above 0 and manually inspected a limited number of samples to ensure the quality of the training data, resulting in 3,304 3D volumes.

Dataset	Scans	Population	Reference
CASILAB	109	healthy	(Bullitt et al., 2005)
ICBM	191	healthy	(Mazziotta et al., 2001)
IXI	568	healthy	(Biomedical Image Analysis Group)
LAUSANNE	292	healthy, brain aneurysm	(Di Noto et al., 2023)
OASIS-3	784	neuro-degenerative disease	(LaMontagne et al., 2019)
RSNA	1230	brain aneurysm	(Jeff Rudie et al., 2025)
TOPCOW2024	130	stroke- related neurological disorder	(Yang et al., 2025)

Table 2: Source datasets with final count of scans used for training.

A.2 Generative Model and Training

We trained 3D-MedDiffusion (Wang et al., 2025) from scratch on the TOF-MRA dataset described in appendix A.1, using the default configuration with $8\times$ downsampling. The PatchVolume autoencoder was trained in two stages: Stage 1 (patch-wise encoding, 4 GPUs) and Stage 2 (volume-wise decoding, 1 GPU), after which training images were encoded to latents for the BiFlowNet noise estimator. The BiFlowNet was trained for 1,000 epochs on a single GPU with checkpoints saved every 10 epochs, yielding 100 checkpoints used for the fidelity sensitivity criterion (section 4.1.2).

A.3 Synthetic Image Results

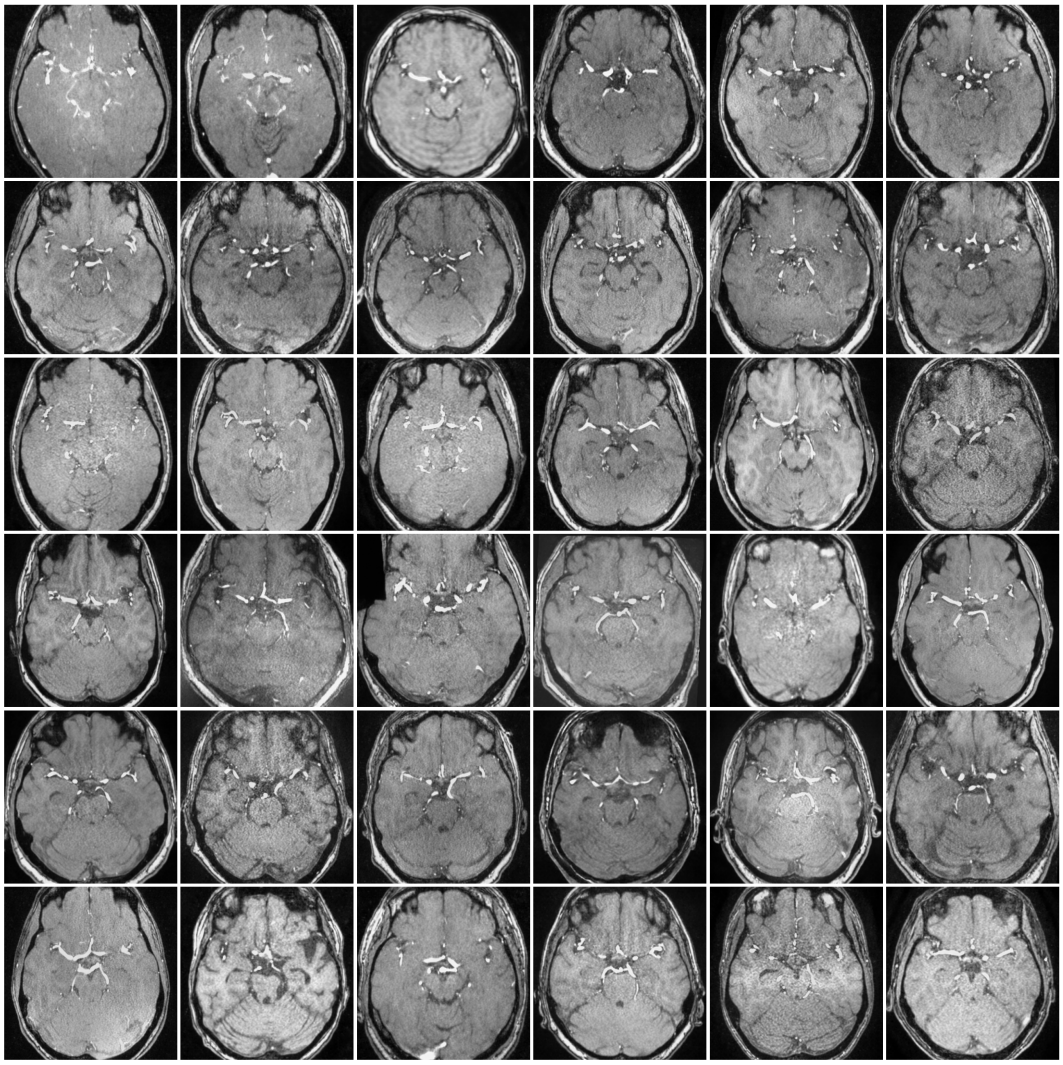


Figure 4: Example synthetic TOF-MRA image slices generated with the generative model described at appendix A.2. Axial slices are extracted from the skull base to show arteries of the Circle of Willis.

A.4 Encoder Adaptation for Feature Extraction

Med3D (Chen et al., 2019) is a 3D ResNet-50 trained under supervision on 3DSeg-8, a heterogeneous collection of eight 3D medical segmentation datasets spanning MRI and CT across multiple organs and pathologies. Prior work has used Med3D as a feature encoder for FD by replacing the segmentation head with global average pooling over the deepest layer (`layer4`), yielding a $d = 2048$ embedding (Sun et al., 2022; Dorjsembe et al., 2024); we follow this procedure. Images are preprocessed and resized to the input dimensions used in the Med3D paper.

The remaining five encoders were pretrained with self-supervised methods on the OpenMind benchmark (Wald et al., 2025c), a large-scale dataset of 114k brain MRI volumes. The PrimusM-S3D combination is not offered by OpenMind. Pretrained checkpoints are publicly available and were downloaded from the project's HuggingFace page.[1] Images are preprocessed and resized to the patch size used in the OpenMind benchmark. We use the encoder parts of these models and adapt them for feature extraction as described below.

[1] `https://huggingface.co/datasets/MIC-DKFZ/OpenMind`

Table 3: Overview of the six evaluated encoders.

Encoder	Architecture	Training	d
Med3D	3D ResNet-50	Supervised (3DSeg-8)	2048
PrimusM-OM-MAE	PrimusM	MAE	864
PrimusM-OM-VoCo	PrimusM	VoCo	864
ResEncL-OM-MAE	ResEncL	MAE	320
ResEncL-OM-S3D	ResEncL	S3D	320
ResEncL-OM-VoCo	ResEncL	VoCo	320

For the ResEncL architecture (Isensee et al., 2024), a residual-encoder U-Net, we use the output of the deepest encoder layer and apply global average pooling over the spatial axes of the resulting feature map with $C = 320$ channels, yielding a $d = 320$ embedding.

For the PrimusM architecture (Wald et al., 2025a), a Transformer with thin convolutional patch embedding and decoding, we use the $N = 8000$ tokens (each of dimension $d_{\text{tok}} = 864$) produced by the Transformer encoder before the final convolutional decoder layer, and apply global average pooling over the token dimension, yielding a $d = 864$ embedding.

No normalization is applied to the embeddings before computing the FD. Table 3 provides an overview of the six evaluated encoders, their architectures, training objectives, and embedding dimensions.

A.5 Visual Rating Protocol

The visual rating protocol consisted of three main parts and was conducted blinded across raters. First, raters were shown a set of real TOF-MRA image data from the training set of the generative model. This aimed to set expectations about general image quality and to mimic the comparative nature in which the FID-like metrics evaluate the fidelity of synthetic data relative to real data. Second, raters inspected synthetic images in four distinct aspects:

- Overall quality, realism, 3D continuity (Likert scale 1-5, the higher the better)
- Anatomical correspondence, realism and biological plausibility (Likert scale 1-5, the higher the better)
- Vessel continuity, representation, realism and biological plausibility (Likert scale 1-5, the higher the better)
- Prevalence of artifacts (Present? If yes, realistic or unusual for TOF-MRA?)

Third, for each synthetic image presented, raters were asked to select specific anatomical variations of vessel topology, pathology if present and further comment on any additional visual element important to be noted. We did not include any direct control on generating anatomical or pathological variations, however, we anticipated potential generation of such heterogeneity, since the training set of our generative models contained multiple pathologies (such as aneurysms, potential arterial calcification and stenosis) and the cerebral vascular tree exhibits a great, natural biological variability. We included the following variations:

- **Vessel topology** - Select (multiple) applicable variation from the list and assess anatomical correspondence and realism specifically to each variation selected (Likert scale 1-5, the higher the better)
  - Fetal-type PCA (FTP): a prominent PComA, with a hypoplastic (”partial fetal”) or absent (”full fetal”) ipsilateral P1
  - Hypoplastic or absent A1: both ACAs supplied by one ICA across the AComA
  - AComA variants: absence, duplication/fenestration of the AComA
  - A2 variants: ”azygos” (fused A2s), third A2
  - MCA variations: missing or short M1, trifurcation up to pentafurcation
- **Pathology** - Select (multiple) applicable variation from the list and assess anatomical correspondence and realism specifically to each variation selected (Likert scale 1-5, the higher the better)
  - Aneurysm
  - Arterial stenosis, vasculitis, occlusion

  - Ischemia
- **Other comment** - Free text field

Table 4: Inter-rater agreement for the three rating axes. α ord.: Krippendorff's α for ordinal data. ICC(2,1): single-rater absolute agreement (four raters). ICC(3,1): single-rater consistency (four raters). ICC(2,4): four-rater average absolute agreement. ICC(3,4): four-rater average consistency. ICC(2,2) R1+R4: two-rater average absolute agreement for the Rater 1+4 pair. ICC(3,2) R1+R4: two-rater average consistency for the Rater 1+4 pair. 95% percentile bootstrap confidence intervals shown in brackets.

Axis	α ord.	ICC(2,1)	ICC(3,1)	ICC(2,4)	ICC(3,4)	ICC(2,2) R1+R4	ICC(3,2) R1+R4
Quality Overall	0.008	0.149 [0.020, 0.310]	0.367 [0.260, 0.480]	0.412 [0.060, 0.640]	0.699 [0.590, 0.790]	0.233 [−0.170, 0.530]	0.520 [0.290, 0.680]
Quality Anatomy	−0.079	0.092 [0.000, 0.210]	0.256 [0.160, 0.370]	0.288 [0.010, 0.510]	0.579 [0.430, 0.700]	0.110 [−0.120, 0.320]	0.217 [−0.160, 0.470]
Quality Vessels	−0.036	0.124 [0.010, 0.270]	0.328 [0.230, 0.440]	0.362 [0.030, 0.590]	0.662 [0.540, 0.760]	0.220 [−0.150, 0.490]	0.419 [0.140, 0.610]

Following visual rating, results of the second rating step were used for exploratory analysis of visual quality (overall, anatomical and vessel quality), while the third rating step was used to test association between metric behaviour and annotated topological and pathological variations. Inter-rater agreements are shown in table 4.

A.6 Bias Correction

Let $y_{r,i}$ denote the raw Likert rating assigned by rater $r \in \{1, \dots, R\}$ to synthetic image s_i, $i \in \{1, \dots, N\}$ for a given quality axis of rating, where $R = 4$. For each rater r, we compute the sample mean and sample standard deviation across all rated images:

$$\bar{y}_r = \frac{1}{N} \sum_{i=1}^{N} y_{r,i}, \tag{9}$$

$$\sigma_r = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_{r,i} - \bar{y}_r)^2}. \tag{10}$$

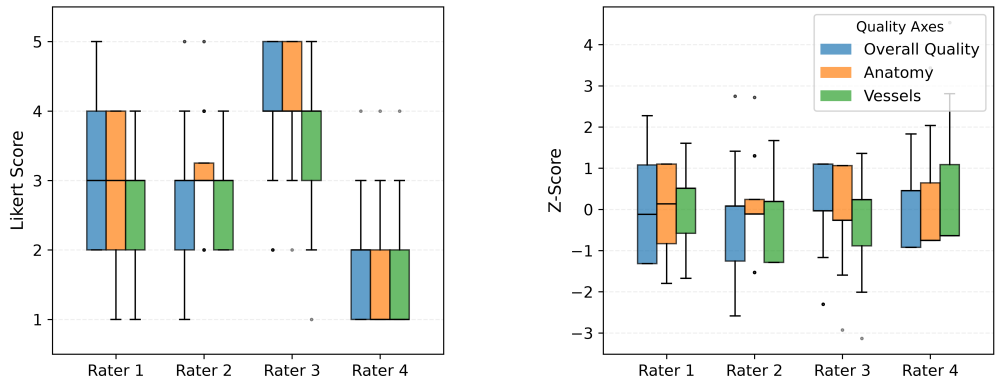
Each rating is then z-score normalized:

$$\tilde{y}_{r,i} = \frac{y_{r,i} - \bar{y}_r}{\sigma_r}. \tag{11}$$

Finally, the bias-corrected score for image s_i on the given rating axis is obtained by averaging the normalized ratings across raters:

$$\hat{y}_i = \frac{1}{R} \sum_{r=1}^{R} \tilde{y}_{r,i}. \tag{12}$$

Figure 5 illustrates the per-rater rating distributions before and after bias correction, showing the removal of systematic inter-rater offsets while preserving relative ranking.



(a) **(b)**

Figure 5: Per-rater rating distributions before and after bias correction. (a) Raw Likert scores (1–5) across the three quality axes. (b) Z-score normalized scores per rater, showing the removal of systematic inter-rater bias.

B Method Details and Algorithms

This appendix provides additional information on the methods introduced in section 4, such as pseudocode.

B.1 Feature Sensitivity via Image Distortions

We apply four distortion types, each parameterized by a magnitude α sampled uniformly from $[0, 1]$. For each image x_i, a distortion $\Phi(x_i, \alpha_i)$ produces a distorted variant s_i. The sampled α is scaled differently for each distortion type.

Gaussian Noise. Additive Gaussian noise with standard deviation $\sigma = (\alpha \cdot 0.5) \cdot \mathrm{std}(x_i)$.

Elastic Deformation. A random displacement field on a $7 \times 7 \times 7$ control-point grid is interpolated with cubic splines (order $= 3$) using the `elasticdeform` library. Per-axis displacement magnitudes are $\alpha \cdot 10$ voxels, scaled by $1/8$ along the thin axis to account for anisotropic voxel spacing.

Random Masking. The volume is divided into a regular grid of non-overlapping 3D patches of size $16 \times 16 \times 8$ voxels. A fraction $\alpha \cdot 0.1$ of patches is randomly selected and set to zero.

Vessel Masking. Vessels are segmented using vesselFM (Wittmann et al., 2025). From a random seed voxel on the vessel mask, a connected region is grown via geodesic dilation (6-connectivity, constrained to the mask) until reaching a voxel limit of $100 + \alpha \cdot 4900$, then dilated by 2 additional layers unconstrained to cover vessel edges not fully captured by the segmentation. The selected voxels are set to zero. This distortion targets the clinically most relevant structures in TOF-MRA, where vessels only a few voxels wide can determine whether an image is anatomically plausible.

Algorithm 1 Distortion Sensitivity Estimation

Require: Original dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, encoder $\mathcal{E}$, bootstrap repetitions R, distortion Φ
Ensure: ρ, 95% confidence interval $[\rho^{\text{lo}},\ \rho^{\text{hi}}]$

$\mathcal{S} \leftarrow \{s_i = \Phi(x_i, \alpha_i) \mid x_i \in \mathcal{D},\ \alpha_i \sim \text{Unif}\}$ ▷ perturbed dataset

for $j = 1$ to N **do** ▷ jackknife sampling over all images
  $d_j \leftarrow \text{FD}(\mathcal{E}(\mathcal{D}),\ \mathcal{E}(\mathcal{S} \setminus \{s_j\}))$ ▷ single image influence
end for
$\rho \leftarrow \text{Spearman}\big((\alpha_j)_{j=1}^{N},\ (d_j)_{j=1}^{N}\big)$

for $r = 1$ to R **do**
  Draw $I_r \leftarrow i_1, \ldots, i_N \sim \text{Unif}\{1, \ldots, N\}$ with replacement
  $\rho_r \leftarrow \text{Spearman}((\alpha_j)_{j \in I_r},\ (d_j)_{j \in I_r})$
end for
$[\rho^{\text{lo}},\ \rho^{\text{hi}}] \leftarrow \text{Percentile}\big((\rho_r)_{r=1}^{R},\ [2.5,\ 97.5]\big)$

return ρ, $[\rho^{\text{lo}},\ \rho^{\text{hi}}]$

B.2 Fidelity Sensitivity via Training Epochs

Algorithm 2 Fidelity Sensitivity Estimation

Require: Original dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, checkpoints $\{C_k\}_{k=1}^{K}$ with epochs $\{t_k\}$, encoder $\mathcal{E}$, repetitions R_{gen}, sample size M
Ensure: $\bar{\rho}$, σ_z, observed range $[\rho^{\min}, \rho^{\max}]$, p_{strat}

for $r = 1$ to R_{gen} **do**
  for $k = 1$ to K **do** ▷ generate synthetic sets for checkpoints
    $\mathcal{S}_k^{(r)} \leftarrow \{s_i \sim C_k\}_{i=1}^{M}$
  end for
end for

for $r = 1$ to R_{gen} **do**
  $\mathcal{D}' \subseteq \mathcal{D}, |\mathcal{D}'| = N - M$
  for $k = 1$ to K **do**
    $d_k^{(r)} \leftarrow \text{FD}\Big(\mathcal{E}(\mathcal{D}),\ \mathcal{E}\Big(\mathcal{D}' \cup \mathcal{S}_k^{(r)}\Big)\Big)$
  end for
  $\rho_r \leftarrow \text{Spearman}\Big((t_k)_{k=1}^{K},\ (d_k^{(r)})_{k=1}^{K}\Big)$
  $z_r \leftarrow \text{artanh}\big(\text{clip}(\rho_r,\ -1 + \varepsilon,\ 1 - \varepsilon)\big)$
end for

$\bar{\rho} \leftarrow \tanh\big(\frac{1}{R_{\text{gen}}} \sum_{r=1}^{R_{gen}} z_r\big)$
$\sigma_z \leftarrow \text{SD}\left((z_r)_{r=1}^{R_{gen}}\right)$
$\rho^{\min} \leftarrow \min_r \rho_r, \quad \rho^{\max} \leftarrow \max_r \rho_r$
$p_{\text{strat}} \leftarrow \text{StratifiedPermutationTest}\big(\{(t_k, d_k^{(r)})\}_{r=1}^{R_{\text{gen}}},\ \bar{\rho}\big)$ ▷ algorithm 3

return $\bar{\rho},\ \sigma_z,\ [\rho^{\min}, \rho^{\max}],\ p_{\text{strat}}$

Algorithm 3 Stratified Permutation Test

Require: Per-seed paired observations $\{(x^{(r)}, d^{(r)})\}_{r=1}^{R_{\text{gen}}}$, observed statistic T_{obs}, permutations N_{perm}
Ensure: p-value p_{strat}

for $i = 1$ to N_{perm} **do**
  for $r = 1$ to R_{gen} **do**
    $\tilde{d}^{(r)} \leftarrow \text{Permute}\big(d^{(r)}\big)$
    $\tilde{\rho}_r \leftarrow \text{Spearman}\big(x^{(r)},\ \tilde{d}^{(r)}\big)$
  end for
  $T_i \leftarrow \tanh\big(\frac{1}{R_{\text{gen}}} \sum_{r=1}^{R_{\text{gen}}} \text{arctanh}\big(\text{clip}(\tilde{\rho}_r,\ -1{+}\varepsilon,\ 1{-}\varepsilon)\big)\big)$
end for
$p_{\text{strat}} \leftarrow \frac{|\{i : T_i \le T_{\text{obs}}\}| + 1}{N_{\text{perm}} + 1}$ ▷ one-sided p-value
return p_{strat}

B.3 Empirical Coverage Sensitivity via Branching Diffusion

Algorithm 4 Empirical Coverage Sensitivity Estimation

Require: Original dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, diffusion model with T steps, branching points $\mathcal{B} = \{b_k\}_{k=1}^{K}$, number of prototypes P, number of variations V, encoder $\mathcal{E}$, repetitions R_{gen}
Ensure: $\bar{\rho}$, σ_z, observed range $[\rho^{\min}, \rho^{\max}]$, p_{strat}

for $r = 1$ to R_{gen} **do**
  for $k = 1$ to K **do** ▷ synthetic datasets with empirical coverage controlled by branching point
    $\mathcal{S}_k^{(r)} \leftarrow \emptyset$
    for $p = 1$ to P **do** ▷ create P prototypes
      Sample $z_p \sim \mathcal{N}(0, I)$
      $\text{proto}_p \leftarrow \text{Denoise}(z_p,\ T - b_k)$
      for $i = 1$ to V **do** ▷ for each prototype, create V variations
        $s_i \leftarrow \text{Denoise}(\text{proto}_p,\ b_k)$
        $\mathcal{S}_k^{(r)} \leftarrow \mathcal{S}_k^{(r)} \cup \{s_i\}$
      end for
    end for
  end for
end for

for $r = 1$ to R_{gen} **do**
  $\mathcal{D}' \subseteq \mathcal{D}, |\mathcal{D}'| = N - PV$
  for $k = 1$ to K **do**
    $d_k^{(r)} \leftarrow \text{FD}\Big(\mathcal{E}(\mathcal{D}),\ \mathcal{E}\Big(\mathcal{D}' \cup \mathcal{S}_k^{(r)}\Big)\Big)$
  end for
  $\rho_r \leftarrow \text{Spearman}\Big((b_k)_{k=1}^{K},\ (d_k^{(r)})_{k=1}^{K}\Big)$
  $z_r \leftarrow \text{artanh}\big(\text{clip}(\rho_r,\ -1+\varepsilon,\ 1-\varepsilon)\big)$
end for

$\bar{\rho} \leftarrow \tanh\big(\frac{1}{R_{\text{gen}}} \sum_{r=1}^{R_{gen}} z_r\big)$
$\sigma_z \leftarrow \text{SD}\left((z_r)_{r=1}^{R_{gen}}\right)$
$\rho^{\min} \leftarrow \min_r \rho_r, \quad \rho^{\max} \leftarrow \max_r \rho_r$
$p_{\text{strat}} \leftarrow \text{StratifiedPermutationTest}\big(\{(b_k, d_k^{(r)})\}_{r=1}^{R_{\text{gen}}},\ \bar{\rho}\big)$ ▷ algorithm 3

return $\bar{\rho},\ \sigma_z,\ [\rho^{\min}, \rho^{\max}],\ p_{\text{strat}}$

B.4 Analysis via Expert Visual Perception

Algorithm 5 Perceived Visual Quality

Require: Original dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, rated synthetic dataset $\mathcal{S} = \{(s_j, \hat{y}_j)\}_{j=1}^{M}$, encoder $\mathcal{E}$, bootstrap repetitions R
Ensure: τ_b, 95% confidence interval $[\tau_b^{\text{lo}},\ \tau_b^{\text{hi}}]$

$\mathcal{D}' \subseteq \mathcal{D}, |\mathcal{D}'| = N - M + 1$

for $j = 1$ to M **do**
    $d_j \leftarrow \text{FD}(\mathcal{E}(\mathcal{D}),\ \mathcal{E}(\mathcal{D}' \cup (\mathcal{S} \setminus \{s_j\})))$
end for
$\tau_b \leftarrow \text{KendallTauB}\big((\hat{y}_j)_{j=1}^{M},\ (d_j)_{j=1}^{M}\big)$
$\tau_b \leftarrow -\tau_b$ ▷ sign-adjust for comparability

for $r = 1$ to R **do**
    Draw $I_r \leftarrow i_1, \ldots, i_M \sim \text{Unif}\{1, \ldots, M\}$ with replacement
    $(\tau_b)_r \leftarrow \text{KendallTauB}((\hat{y}_j)_{j \in I_r},\ (d_j)_{j \in I_r})$
    $(\tau_b)_r \leftarrow -(\tau_b)_r$
end for
$[\tau_b^{\text{lo}},\ \tau_b^{\text{hi}}] \leftarrow \text{Percentile}\big((\tau_b)_r)_{r=1}^{R},\ [2.5,\ 97.5]\big)$

return τ_b, $[\tau_b^{\text{lo}},\ \tau_b^{\text{hi}}]$

B.5 Association with Annotated Vessel Topology and Pathology Variations

Algorithm 6 Vessel Topology and Pathology Variation

Require: Real dataset $D = \{x_i\}_{i=1}^{N}$, synthetic dataset $S = \{s_j, v_j\}_{j=1}^{M}$ with v_j being the variation load of s_j, encoder E, bootstrap repetitions R
Ensure: τ_b, 95% confidence interval $[\tau_b^{\text{lo}},\ \tau_b^{\text{hi}}]$

$\mathcal{D}' \subseteq \mathcal{D}, |\mathcal{D}'| = N - M + 1$

for $j = 1$ to M **do**
    $d_j \leftarrow \text{FD}\,(\mathcal{E}(D),\ \mathcal{E}(D' \cup (\mathcal{S} \setminus \{s_j\})))$
end for
$\tau \leftarrow \text{KendallTauB}\big((v_j)_{j=1}^{M},\ (d_j)_{j=1}^{M}\big)$
$\tau \leftarrow -\tau$ ▷ sign-adjust for comparability

for $r = 1$ to R **do**
    Draw $I_r \leftarrow i_1, \ldots, i_M \sim \text{Unif}\{1, \ldots, M\}$ with replacement
    $(\tau_b)_r \leftarrow \text{KendallTauB}\big((v_j)_{j \in I_r},\ (d_j)_{j \in I_r}\big)$
    $(\tau_b)_r \leftarrow -(\tau_b)_r$
end for
$[\tau_b^{\text{lo}},\ \tau_b^{\text{hi}}] \leftarrow \text{Percentile}\big((\tau_b)_r)_{r=1}^{R},\ [2.5,\ 97.5]\big)$
return $\tau,\ [\tau^{\text{lo}}, \tau^{\text{hi}}]$

C Detailed Results

This appendix provides detailed numerical results complementing the combined forest plot in fig. 2 and in fig. 3.

C.1 Feature Sensitivity via Image Distortions

Table 5: Distortion sensitivity (Spearman's ρ and Kendall's τ_b between distortion magnitude and FD, leave-one-out). More negative values indicate stronger sensitivity. 95% percentile bootstrap confidence intervals shown in brackets.

Encoder		Elastic Deformation	Gaussian Noise	Random Masking	Vessel Masking
Med3D	ρ	-0.191 [-0.223, -0.158]	-0.825 [-0.837, -0.812]	-0.645 [-0.666, -0.624]	-0.007 [-0.040, 0.026]
	τ_b	-0.128 [-0.150, -0.105]	-0.627 [-0.639, -0.615]	-0.461 [-0.478, -0.444]	-0.005 [-0.027, 0.017]
PrimusM-OM-MAE	ρ	-0.494 [-0.521, -0.467]	-0.749 [-0.762, -0.735]	-0.483 [-0.509, -0.455]	-0.045 [-0.078, -0.011]
	τ_b	-0.342 [-0.362, -0.323]	-0.540 [-0.552, -0.527]	-0.333 [-0.352, -0.313]	-0.030 [-0.052, -0.007]
PrimusM-OM-VoCo	ρ	-0.111 [-0.144, -0.077]	-0.538 [-0.561, -0.513]	-0.041 [-0.076, -0.007]	-0.053 [-0.087, -0.019]
	τ_b	-0.074 [-0.096, -0.051]	-0.371 [-0.389, -0.353]	-0.027 [-0.051, -0.004]	-0.036 [-0.059, -0.013]
ResEncL-OM-MAE	ρ	-0.339 [-0.370, -0.308]	-0.562 [-0.587, -0.537]	-0.497 [-0.523, -0.472]	-0.293 [-0.323, -0.261]
	τ_b	-0.231 [-0.252, -0.209]	-0.395 [-0.414, -0.377]	-0.343 [-0.361, -0.324]	-0.197 [-0.218, -0.176]
ResEncL-OM-S3D	ρ	-0.661 [-0.680, -0.641]	-0.862 [-0.871, -0.852]	-0.832 [-0.844, -0.818]	-0.285 [-0.316, -0.254]
	τ_b	-0.465 [-0.481, -0.450]	-0.670 [-0.681, -0.659]	-0.641 [-0.654, -0.628]	-0.193 [-0.215, -0.172]
ResEncL-OM-VoCo	ρ	-0.170 [-0.202, -0.136]	-0.652 [-0.671, -0.631]	-0.559 [-0.585, -0.533]	-0.231 [-0.264, -0.199]
	τ_b	-0.114 [-0.136, -0.091]	-0.459 [-0.475, -0.444]	-0.395 [-0.415, -0.375]	-0.155 [-0.177, -0.133]

Table 6: Pairwise bootstrap difference CIs (Spearman's $\rho_i - \rho_j$) underlying CLD groups and ranking scores for distortion sensitivity. Cell (i, j): 95% percentile CI of $\rho_i - \rho_j$ across bootstrap replicates. CLD letters and ranking scores shown below each matrix.

Gaussian Noise

	Med3D	PrimusM-OM-MAE	PrimusM-OM-VoCo	ResEncL-OM-MAE	ResEncL-OM-S3D	ResEncL-OM-VoCo
Med3D	—	[-0.095, -0.058]	[-0.316, -0.261]	[-0.291, -0.236]	[0.021, 0.052]	[-0.197, -0.151]
PrimusM-OM-MAE	[0.058, 0.095]	—	[-0.239, -0.183]	[-0.215, -0.158]	[0.096, 0.130]	[-0.121, -0.073]
PrimusM-OM-VoCo	[0.261, 0.316]	[0.183, 0.239]	—	[-0.010, 0.059]	[0.298, 0.351]	[0.083, 0.145]
ResEncL-OM-MAE	[0.236, 0.291]	[0.158, 0.215]	[-0.059, 0.010]	—	[0.274, 0.327]	[0.058, 0.122]
ResEncL-OM-S3D	[-0.052, -0.021]	[-0.130, -0.096]	[-0.351, -0.298]	[-0.327, -0.274]	—	[-0.233, -0.188]
ResEncL-OM-VoCo	[0.151, 0.197]	[0.073, 0.121]	[-0.145, -0.083]	[-0.122, -0.058]	[0.188, 0.233]	—
CLD	b	c	e	e	a	d
Wins	4	3	0	0	5	2
Losses	1	2	4	4	0	3
Score	3	1	-4	-4	5	-1

Elastic Deformation

	Med3D	PrimusM-OM-MAE	PrimusM-OM-VoCo	ResEncL-OM-MAE	ResEncL-OM-S3D	ResEncL-OM-VoCo
Med3D	—	[0.260, 0.346]	[-0.127, -0.034]	[0.103, 0.194]	[0.432, 0.507]	[-0.068, 0.024]
PrimusM-OM-MAE	[-0.346, -0.260]	—	[-0.428, -0.339]	[-0.197, -0.113]	[0.133, 0.199]	[-0.368, -0.281]
PrimusM-OM-VoCo	[0.034, 0.127]	[0.339, 0.428]	—	[0.183, 0.276]	[0.510, 0.589]	[0.012, 0.106]
ResEncL-OM-MAE	[-0.194, -0.103]	[0.113, 0.197]	[-0.276, -0.183]	—	[0.285, 0.358]	[-0.216, -0.125]
ResEncL-OM-S3D	[-0.507, -0.432]	[-0.199, -0.133]	[-0.589, -0.510]	[-0.358, -0.285]	—	[-0.529, -0.452]
ResEncL-OM-VoCo	[-0.024, 0.068]	[0.281, 0.368]	[-0.106, -0.012]	[0.125, 0.216]	[0.452, 0.529]	—
CLD	d	b	e	c	a	d
Wins	1	4	0	3	5	1
Losses	3	1	5	2	0	3
Score	-2	3	-5	1	5	-2

Random Masking

	Med3D	PrimusM-OM-MAE	PrimusM-OM-VoCo	ResEncL-OM-MAE	ResEncL-OM-S3D	ResEncL-OM-VoCo
Med3D	—	[-0.197, -0.129]	[-0.644, -0.564]	[-0.181, -0.116]	[0.162, 0.211]	[-0.119, -0.052]
PrimusM-OM-MAE	[0.129, 0.197]	—	[-0.484, -0.397]	[-0.022, 0.051]	[0.320, 0.379]	[0.038, 0.115]
PrimusM-OM-VoCo	[0.564, 0.644]	[0.397, 0.484]	—	[0.414, 0.499]	[0.754, 0.827]	[0.474, 0.562]
ResEncL-OM-MAE	[0.116, 0.181]	[-0.051, 0.022]	[-0.499, -0.414]	—	[0.306, 0.363]	[0.025, 0.098]
ResEncL-OM-S3D	[-0.211, -0.162]	[-0.379, -0.320]	[-0.827, -0.754]	[-0.363, -0.306]	—	[-0.302, -0.243]
ResEncL-OM-VoCo	[0.052, 0.119]	[-0.115, -0.038]	[-0.562, -0.474]	[-0.098, -0.025]	[0.243, 0.302]	—
CLD	b	d	e	d	a	c
Wins	4	1	0	1	5	3
Losses	1	3	5	3	0	2
Score	3	-2	-5	-2	5	1

Vessel Masking

	Med3D	PrimusM-OM-MAE	PrimusM-OM-VoCo	ResEncL-OM-MAE	ResEncL-OM-S3D	ResEncL-OM-VoCo
Med3D	—	[-0.010, 0.085]	[-0.002, 0.094]	[0.240, 0.330]	[0.233, 0.324]	[0.178, 0.271]
PrimusM-OM-MAE	[-0.085, 0.010]	—	[-0.040, 0.056]	[0.203, 0.294]	[0.196, 0.286]	[0.139, 0.234]
PrimusM-OM-VoCo	[-0.094, 0.002]	[-0.056, 0.040]	—	[0.195, 0.286]	[0.185, 0.279]	[0.131, 0.225]
ResEncL-OM-MAE	[-0.330, -0.240]	[-0.294, -0.203]	[-0.286, -0.195]	—	[-0.051, 0.037]	[-0.106, -0.018]
ResEncL-OM-S3D	[-0.324, -0.233]	[-0.286, -0.196]	[-0.279, -0.185]	[-0.037, 0.051]	—	[-0.100, -0.008]
ResEncL-OM-VoCo	[-0.271, -0.178]	[-0.234, -0.139]	[-0.225, -0.131]	[0.018, 0.106]	[0.008, 0.100]	—
CLD	c	c	c	a	a	b
Wins	0	0	0	4	4	3
Losses	3	3	3	0	0	2
Score	-3	-3	-3	4	4	1
Avg. Score	0.25	-0.25	-4.25	-0.25	4.75	-0.25

C.2 Fidelity Sensitivity via Training Epochs

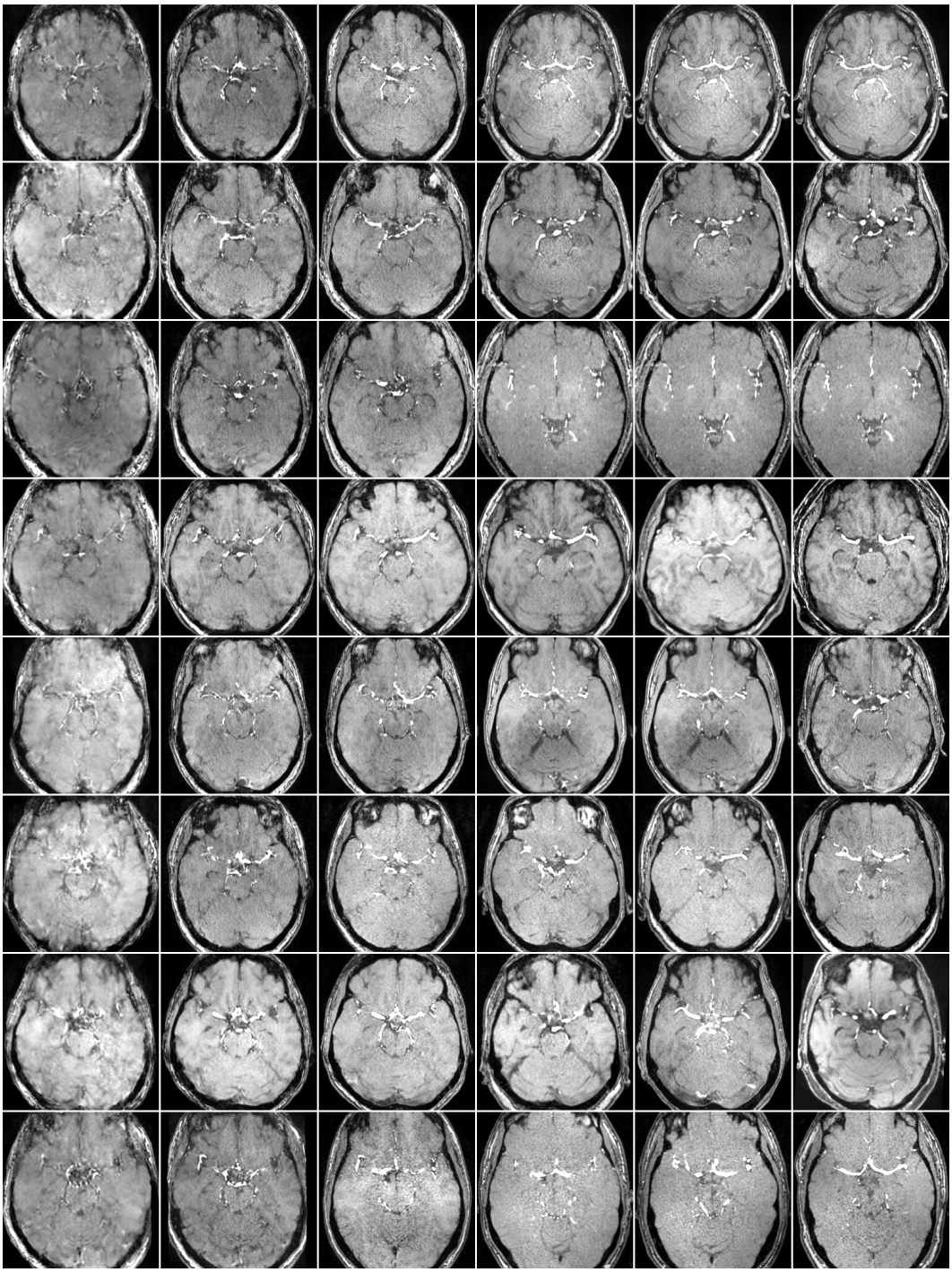


Figure 6: Additional samples generated from fixed seeds across training checkpoints (10, 50 100, 300, 600, 1000 epochs), left to right. Axial slices are extracted from the skull base to show arteries of the Circle of Willis.

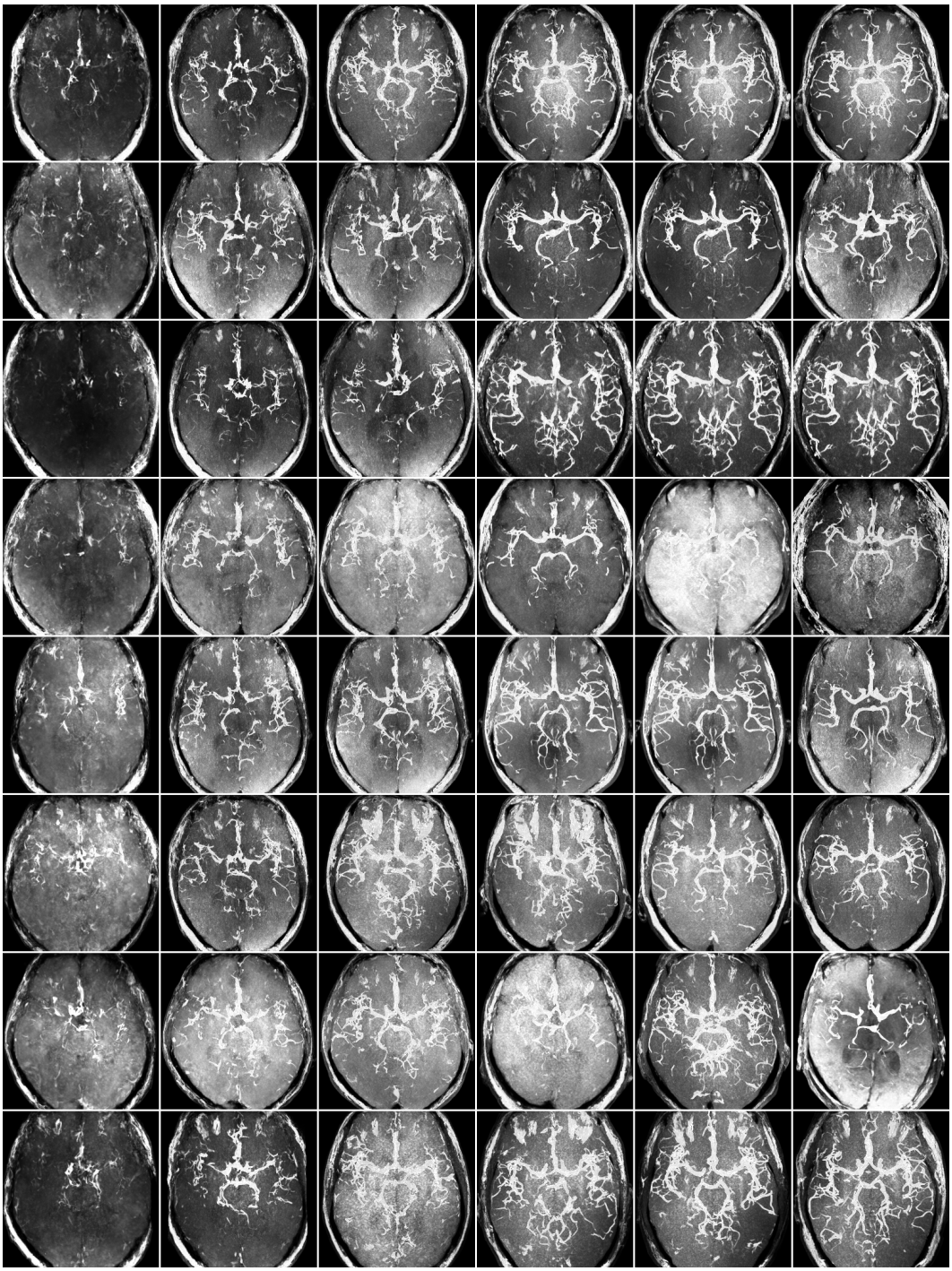


Figure 7: Maximum Intensity Projection (MIP) applied on examples in fig. 6, generated from fixed seeds across training checkpoints (10, 50 100, 300, 600, 1000 epochs), left to right.

Table 7: Fidelity sensitivity (Spearman's $\bar{\rho}$ between training epoch of the checkpoint and FD). $\bar{\rho}$: Fisher z-mean across 10 seeds, projected back to ρ space. σ_z: standard deviation in Fisher z space. Range: min and max across 10 seeds. p_{strat}: one-sided p-value from the stratified permutation test.

Encoder	$\bar{\rho}$	σ_z	**Range**	p_{strat}
Med3D	-0.738	0.367	$[-0.874 \ldots -0.247]$	$<10^{-4}$
PrimusM-OM-MAE	-0.238	0.218	$[-0.549 \ldots 0.104]$	$<10^{-4}$
PrimusM-OM-VoCo	-0.037	0.161	$[-0.351 \ldots 0.226]$	0.306
ResEncL-OM-MAE	-0.288	0.209	$[-0.526 \ldots 0.110]$	$<10^{-4}$
ResEncL-OM-S3D	-0.699	0.251	$[-0.843 \ldots -0.265]$	$<10^{-4}$
ResEncL-OM-VoCo	-0.092	0.411	$[-0.597 \ldots 0.531]$	0.109

Table 8: Pairwise seed differences (observed range of $\rho_i - \rho_j$ across 10 seeds) underlying CLD groups and ranking scores for fidelity sensitivity. Cell (i, j): $[\min(\rho_i - \rho_j), \max(\rho_i - \rho_j)]$ across seeds. CLD letters and ranking scores shown below the matrix.

Fidelity Evaluation						
	Med3D	**PrimusM-OM-MAE**	**PrimusM-OM-VoCo**	**ResEncL-OM-MAE**	**ResEncL-OM-S3D**	**ResEncL-OM-VoCo**
Med3D	—	[−0.913, 0.040]	[−1.065, −0.183]	[−0.755, −0.031]	[−0.609, 0.516]	[−1.213, −0.222]
PrimusM-OM-MAE	[−0.040, 0.913]	—	[−0.432, 0.114]	[−0.466, 0.483]	[0.116, 0.778]	[−0.909, 0.400]
PrimusM-OM-VoCo	[0.183, 1.065]	[−0.114, 0.432]	—	[−0.151, 0.568]	[0.108, 0.952]	[−0.882, 0.587]
ResEncL-OM-MAE	[0.031, 0.755]	[−0.483, 0.466]	[−0.568, 0.151]	—	[−0.261, 0.873]	[−1.034, 0.281]
ResEncL-OM-S3D	[−0.516, 0.609]	[−0.778, −0.116]	[−0.952, −0.108]	[−0.873, 0.261]	—	[−1.312, 0.026]
ResEncL-OM-VoCo	[0.222, 1.213]	[−0.400, 0.909]	[−0.587, 0.882]	[−0.281, 1.034]	[−0.026, 1.312]	—
CLD	ab	ad	d	cd	bc	cd
Wins	3	0	0	0	2	0
Losses	0	1	2	1	0	1
Score	3	-1	-2	-1	2	-1

C.3 Empirical Coverage Sensitivity via Branching Diffusion

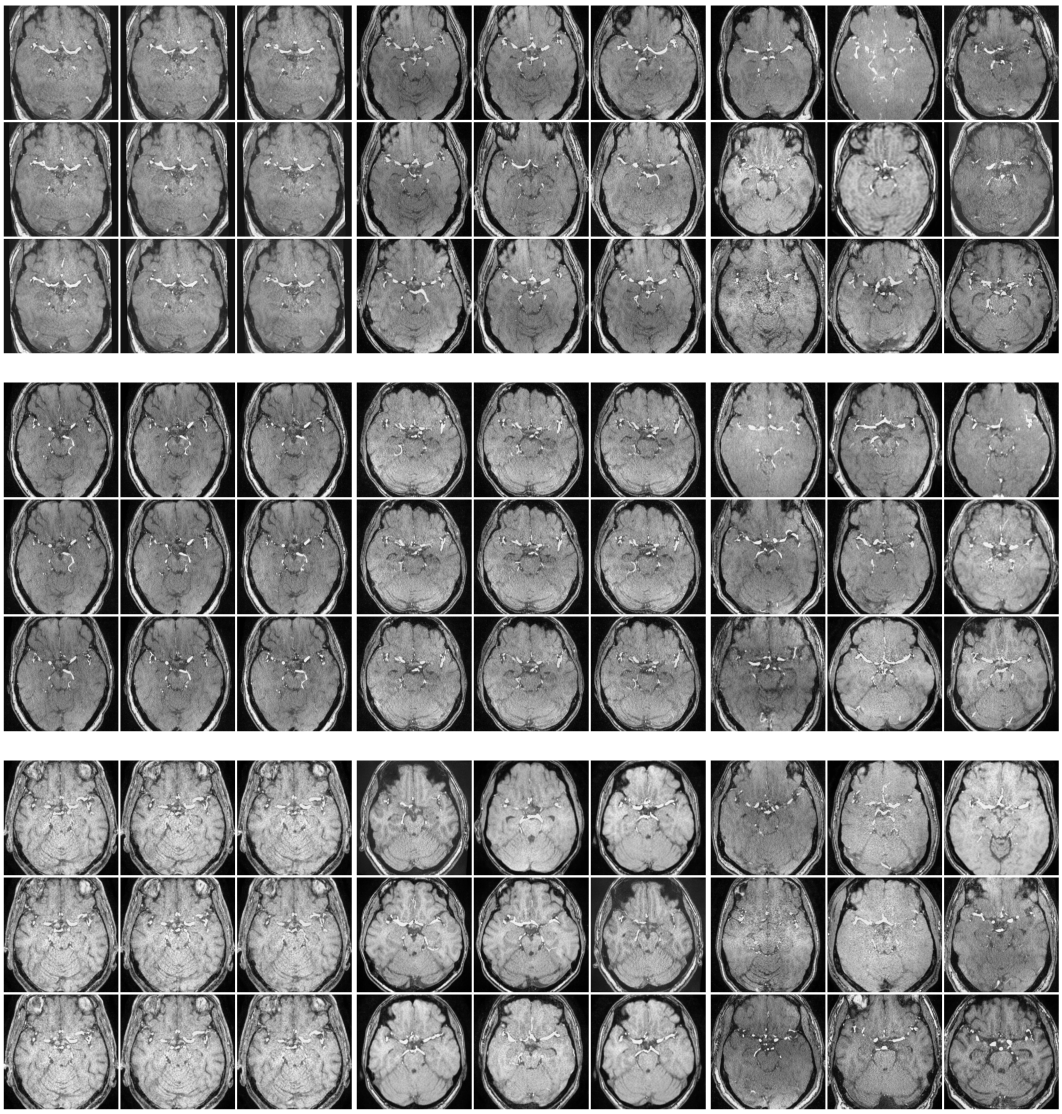


Figure 8: Results generated with Branching Diffusion. Branching points from left to right (b=700, b=850, b=999) with 9 derived variations each. Rows show results for three different initializations of randomness.

Table 9: Empirical coverage sensitivity (Spearman's $\bar{\rho}$ between branching point b and FD). Same columns and reporting convention as table 7. Bold = most negative $\bar{\rho}$.

Encoder	$\bar{\rho}$	σ_z	**Range**	p_{strat}
Med3D	0.027	0.066	$[-0.094 \ldots 0.098]$	0.666
PrimusM-OM-MAE	0.054	0.166	$[-0.194 \ldots 0.302]$	0.819
PrimusM-OM-VoCo	-0.054	0.193	$[-0.416 \ldots 0.077]$	0.178
ResEncL-OM-MAE	-0.725	0.102	$[-0.798 \ldots -0.660]$	$<10^{-4}$
ResEncL-OM-S3D	-0.883	0.16	$[-0.917 \ldots -0.818]$	$<10^{-4}$
ResEncL-OM-VoCo	-0.862	0.128	$[-0.899 \ldots -0.796]$	$<10^{-4}$

Table 10: Pairwise seed differences (observed range of $\rho_i - \rho_j$ across 10 seeds) underlying CLD groups and ranking scores for empirical coverage sensitivity. Cell (i, j): $[\min(\rho_i - \rho_j), \max(\rho_i - \rho_j)]$ across seeds. CLD letters and ranking scores shown below the matrix.

Empirical Coverage

	Med3D	**PrimusM-OM-MAE**	**PrimusM-OM-VoCo**	**ResEncL-OM-MAE**	**ResEncL-OM-S3D**	**ResEncL-OM-VoCo**
Med3D	—	[−0.220, 0.279]	[−0.081, 0.514]	[0.639, 0.857]	[0.763, 1.013]	[0.764, 0.961]
PrimusM-OM-MAE	[−0.279, 0.220]	—	[−0.121, 0.235]	[0.514, 1.077]	[0.723, 1.212]	[0.679, 1.122]
PrimusM-OM-VoCo	[−0.514, 0.081]	[−0.235, 0.121]	—	[0.279, 0.863]	[0.499, 0.986]	[0.444, 0.944]
ResEncL-OM-MAE	[−0.857, −0.639]	[−1.077, −0.514]	[−0.863, −0.279]	—	[0.054, 0.233]	[0.045, 0.239]
ResEncL-OM-S3D	[−1.013, −0.763]	[−1.212, −0.723]	[−0.986, −0.499]	[−0.233, −0.054]	—	[−0.112, 0.059]
ResEncL-OM-VoCo	[−0.961, −0.764]	[−1.122, −0.679]	[−0.944, −0.444]	[−0.239, −0.045]	[−0.059, 0.112]	—
CLD	c	c	c	b	a	a
Wins	0	0	0	3	4	4
Losses	3	3	3	2	0	0
Score	-3	-3	-3	1	4	4

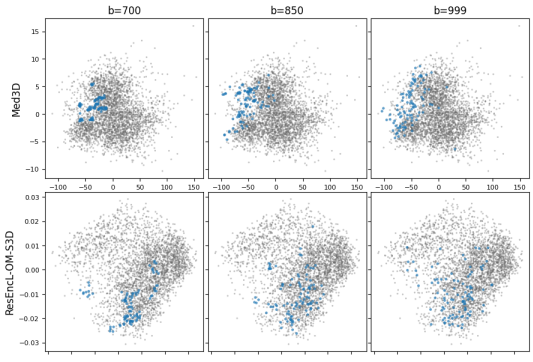


Figure 9: PCA embeddings of original (gray) and synthetic (blue) data, for Med3D (top) and ResEncL-OM-S3D (bottom). Synthetic data was created with Branching Diffusion, with $P = 10$ and $V = 10$ and progressively earlier branching ($b = 700$, $b = 850$ and $b = 999$).

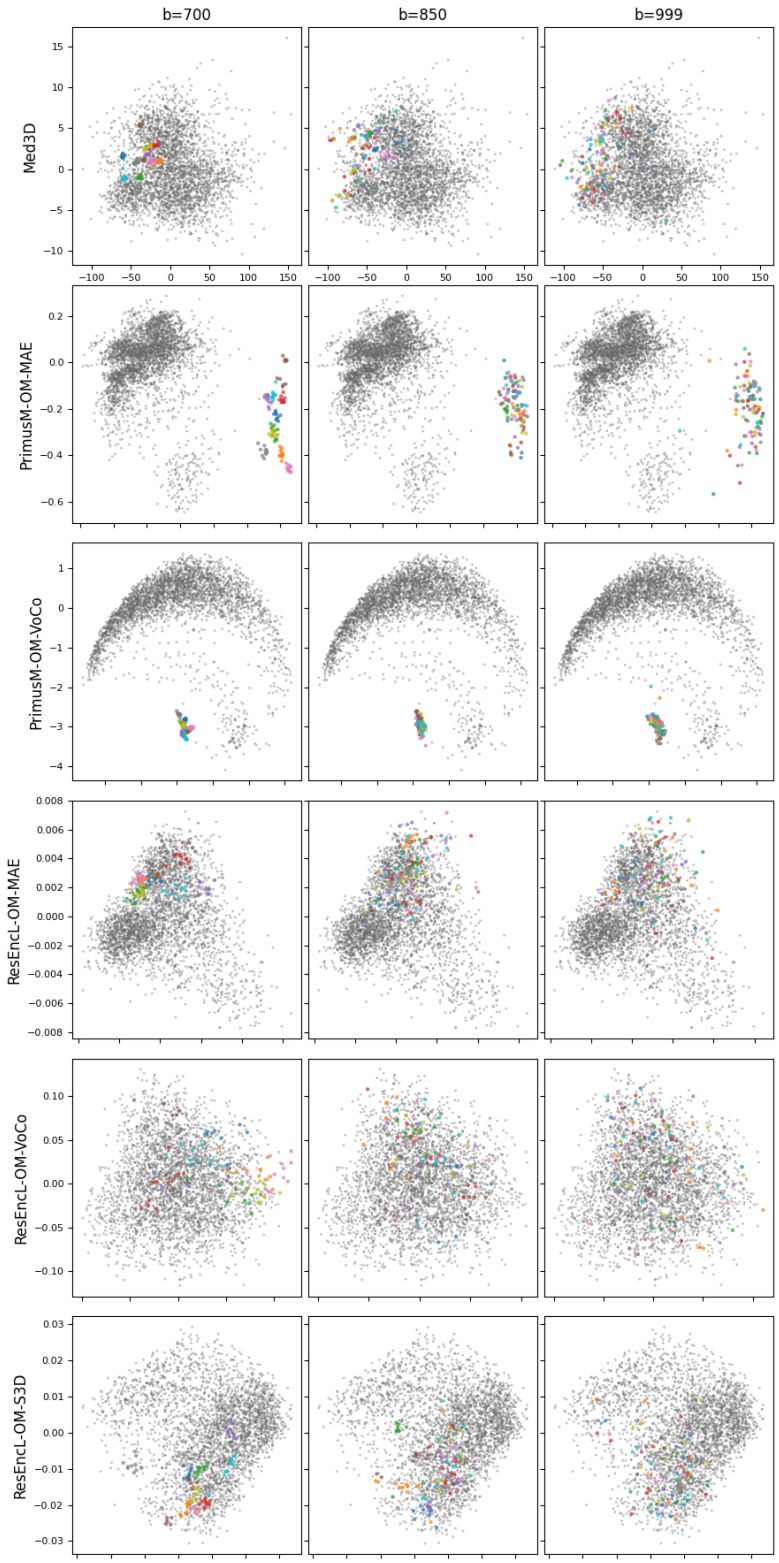


Figure 10: PCA embeddings of original (gray) and synthetic (colored by prototype) data for each encoder. Synthetic data was created with Branching Diffusion, with $P = 10$ and $V = 10$ and progressively earlier branching ($b = 700$, $b = 850$ and $b = 999$).

C.4 Exploratory Analysis via Expert Visual Rating

C.4.1 Quality Evaluation by Visual Perception

Table 11: Pairwise inter-rater agreement for all rater pairs. ICC(2,2): two-rater average absolute agreement. ICC(3,2): two-rater average consistency. 95% percentile bootstrap confidence intervals shown in brackets.

Pair	**Axis**	**ICC(2,2)** [95% CI]	**ICC(3,2)** [95% CI]
R1+R2	Quality Overall	0.652 [0.480, 0.770]	0.659 [0.490, 0.770]
	Quality Anatomy	0.450 [0.190, 0.630]	0.459 [0.200, 0.640]
	Quality Vessels	0.543 [0.310, 0.700]	0.575 [0.370, 0.710]
R1+R3	Quality Overall	0.431 [−0.110, 0.690]	0.606 [0.410, 0.730]
	Quality Anatomy	0.181 [−0.130, 0.430]	0.343 [0.020, 0.560]
	Quality Vessels	0.362 [−0.190, 0.660]	0.606 [0.410, 0.740]
R1+R4	Quality Overall	0.233 [−0.170, 0.530]	0.520 [0.290, 0.680]
	Quality Anatomy	0.110 [−0.120, 0.320]	0.217 [−0.160, 0.470]
	Quality Vessels	0.220 [−0.150, 0.490]	0.419 [0.140, 0.610]
R2+R3	Quality Overall	0.410 [−0.200, 0.700]	0.654 [0.490, 0.770]
	Quality Anatomy	0.363 [−0.200, 0.680]	0.651 [0.480, 0.760]
	Quality Vessels	0.429 [−0.150, 0.700]	0.627 [0.450, 0.750]
R2+R4	Quality Overall	0.190 [−0.150, 0.460]	0.413 [0.130, 0.610]
	Quality Anatomy	0.151 [−0.130, 0.410]	0.429 [0.150, 0.620]
	Quality Vessels	0.134 [−0.120, 0.380]	0.431 [0.150, 0.620]
R3+R4	Quality Overall	0.057 [−0.070, 0.210]	0.268 [−0.090, 0.510]
	Quality Anatomy	0.059 [−0.060, 0.220]	0.372 [0.070, 0.580]
	Quality Vessels	0.016 [−0.050, 0.100]	0.094 [−0.350, 0.390]

Table 12: Correlation between bias-corrected expert quality ratings and per-image influence on FD across encoders (Spearman's ρ and Kendall's τ_b, leave-one-out). More negative values indicate stronger agreement. 95% percentile bootstrap confidence intervals shown in brackets.

		Quality Overall		**Quality Anatomy**		**Quality Vessels**	
Rater	**Encoder**	ρ	τ_b	ρ	τ_b	ρ	τ_b
Raters 1+4	Med3D	−0.114 [−0.314, 0.095]	−0.085 [−0.234, 0.068]	−0.120 [−0.313, 0.076]	−0.085 [−0.226, 0.055]	−0.105 [−0.297, 0.091]	−0.076 [−0.218, 0.067]
	PrimusM-OM-MAE	−0.185 [−0.368, 0.009]	−0.134 [−0.267, 0.006]	−0.148 [−0.342, 0.051]	−0.110 [−0.249, 0.030]	−0.176 [−0.364, 0.026]	−0.123 [−0.265, 0.025]
	PrimusM-OM-VoCo	−0.161 [−0.352, 0.042]	−0.114 [−0.256, 0.035]	−0.125 [−0.325, 0.086]	−0.088 [−0.232, 0.064]	−0.138 [−0.335, 0.066]	−0.099 [−0.243, 0.049]
	ResEncL-OM-MAE	−0.157 [−0.350, 0.049]	−0.117 [−0.265, 0.039]	−0.134 [−0.327, 0.065]	−0.094 [−0.238, 0.053]	−0.062 [−0.247, 0.130]	−0.045 [−0.178, 0.089]
	ResEncL-OM-S3D	−0.257 [−0.425, −0.084]	−0.178 [−0.301, −0.055]	−0.296 [−0.463, −0.115]	−0.209 [−0.333, −0.080]	−0.226 [−0.398, −0.045]	−0.166 [−0.293, −0.038]
	ResEncL-OM-VoCo	−0.147 [−0.335, 0.050]	−0.107 [−0.244, 0.035]	−0.150 [−0.334, 0.043]	−0.106 [−0.237, 0.032]	−0.122 [−0.316, 0.080]	−0.087 [−0.231, 0.059]
Aggregated	Med3D	−0.278 [−0.462, −0.078]	−0.192 [−0.327, −0.051]	−0.264 [−0.442, −0.075]	−0.180 [−0.307, −0.049]	−0.247 [−0.428, −0.058]	−0.168 [−0.297, −0.036]
	PrimusM-OM-MAE	−0.305 [−0.479, −0.122]	−0.210 [−0.335, −0.083]	−0.292 [−0.475, −0.104]	−0.199 [−0.327, −0.071]	−0.319 [−0.483, −0.139]	−0.222 [−0.339, −0.097]
	PrimusM-OM-VoCo	−0.264 [−0.444, −0.071]	−0.184 [−0.311, −0.049]	−0.285 [−0.471, −0.088]	−0.201 [−0.332, −0.064]	−0.302 [−0.481, −0.108]	−0.213 [−0.343, −0.076]
	ResEncL-OM-MAE	−0.257 [−0.445, −0.058]	−0.179 [−0.321, −0.034]	−0.247 [−0.425, −0.059]	−0.171 [−0.297, −0.038]	−0.218 [−0.398, −0.025]	−0.151 [−0.280, −0.016]
	ResEncL-OM-S3D	−0.283 [−0.445, −0.111]	−0.192 [−0.306, −0.075]	−0.368 [−0.529, −0.195]	−0.254 [−0.370, −0.132]	−0.278 [−0.444, −0.100]	−0.192 [−0.307, −0.070]
	ResEncL-OM-VoCo	−0.171 [−0.367, 0.035]	−0.117 [−0.255, 0.029]	−0.135 [−0.329, 0.064]	−0.092 [−0.227, 0.047]	−0.118 [−0.314, 0.083]	−0.078 [−0.215, 0.064]
Rater 1	Med3D	−0.071 [−0.265, 0.122]	−0.058 [−0.211, 0.092]	−0.035 [−0.231, 0.163]	−0.027 [−0.175, 0.122]	−0.072 [−0.266, 0.124]	−0.055 [−0.206, 0.092]
	PrimusM-OM-MAE	−0.095 [−0.286, 0.102]	−0.073 [−0.223, 0.081]	0.059 [−0.142, 0.258]	0.043 [−0.110, 0.193]	−0.088 [−0.290, 0.113]	−0.068 [−0.222, 0.086]
	PrimusM-OM-VoCo	−0.119 [−0.321, 0.089]	−0.095 [−0.255, 0.072]	0.010 [−0.201, 0.214]	0.009 [−0.153, 0.168]	−0.025 [−0.232, 0.184]	−0.019 [−0.181, 0.140]
	ResEncL-OM-MAE	−0.162 [−0.358, 0.040]	−0.128 [−0.285, 0.033]	−0.019 [−0.208, 0.174]	−0.015 [−0.159, 0.130]	−0.079 [−0.267, 0.113]	−0.065 [−0.207, 0.081]
	ResEncL-OM-S3D	−0.168 [−0.347, 0.018]	−0.127 [−0.265, 0.014]	−0.155 [−0.333, 0.035]	−0.117 [−0.253, 0.025]	−0.115 [−0.309, 0.082]	−0.090 [−0.238, 0.059]
	ResEncL-OM-VoCo	−0.168 [−0.360, 0.032]	−0.133 [−0.284, 0.023]	−0.086 [−0.284, 0.113]	−0.068 [−0.219, 0.084]	−0.098 [−0.291, 0.099]	−0.076 [−0.224, 0.074]
Rater 2	Med3D	−0.283 [−0.460, −0.090]	−0.217 [−0.355, −0.066]	−0.198 [−0.361, −0.024]	−0.152 [−0.277, −0.019]	−0.218 [−0.405, −0.021]	−0.173 [−0.324, −0.018]
	PrimusM-OM-MAE	−0.321 [−0.496, −0.133]	−0.254 [−0.394, −0.106]	−0.196 [−0.370, −0.010]	−0.149 [−0.285, −0.006]	−0.332 [−0.501, −0.147]	−0.264 [−0.403, −0.116]
	PrimusM-OM-VoCo	−0.287 [−0.464, −0.091]	−0.226 [−0.369, −0.071]	−0.237 [−0.424, −0.033]	−0.189 [−0.339, −0.026]	−0.365 [−0.542, −0.174]	−0.294 [−0.440, −0.139]
	ResEncL-OM-MAE	−0.242 [−0.428, −0.049]	−0.192 [−0.344, −0.037]	−0.118 [−0.294, 0.067]	−0.092 [−0.231, 0.051]	−0.229 [−0.416, −0.032]	−0.184 [−0.336, −0.026]
	ResEncL-OM-S3D	−0.241 [−0.405, −0.063]	−0.186 [−0.318, −0.050]	−0.241 [−0.407, −0.065]	−0.186 [−0.318, −0.049]	−0.260 [−0.432, −0.077]	−0.203 [−0.340, −0.060]
	ResEncL-OM-VoCo	−0.101 [−0.305, 0.105]	−0.080 [−0.245, 0.084]	0.003 [−0.211, 0.216]	0.003 [−0.167, 0.174]	−0.089 [−0.295, 0.120]	−0.065 [−0.230, 0.103]
Rater 3	Med3D	−0.383 [−0.565, −0.187]	−0.306 [−0.457, −0.148]	−0.311 [−0.485, −0.124]	−0.242 [−0.382, −0.093]	−0.280 [−0.464, −0.084]	−0.221 [−0.369, −0.068]
	PrimusM-OM-MAE	−0.268 [−0.460, −0.068]	−0.215 [−0.369, −0.057]	−0.309 [−0.486, −0.120]	−0.245 [−0.388, −0.097]	−0.225 [−0.408, −0.032]	−0.178 [−0.322, −0.029]
	PrimusM-OM-VoCo	−0.234 [−0.421, −0.039]	−0.185 [−0.333, −0.031]	−0.306 [−0.484, −0.118]	−0.242 [−0.385, −0.092]	−0.200 [−0.375, −0.013]	−0.153 [−0.288, −0.011]
	ResEncL-OM-MAE	−0.309 [−0.490, −0.112]	−0.247 [−0.397, −0.088]	−0.325 [−0.488, −0.144]	−0.255 [−0.386, −0.112]	−0.220 [−0.402, −0.027]	−0.175 [−0.321, −0.025]
	ResEncL-OM-S3D	−0.146 [−0.341, 0.052]	−0.114 [−0.265, 0.039]	−0.267 [−0.436, −0.095]	−0.205 [−0.340, −0.073]	−0.178 [−0.346, −0.004]	−0.130 [−0.256, 0.001]
	ResEncL-OM-VoCo	−0.100 [−0.291, 0.093]	−0.080 [−0.228, 0.068]	−0.119 [−0.319, 0.084]	−0.092 [−0.250, 0.069]	−0.047 [−0.244, 0.151]	−0.040 [−0.191, 0.113]
Rater 4	Med3D	−0.111 [−0.306, 0.095]	−0.087 [−0.242, 0.077]	−0.153 [−0.338, 0.044]	−0.121 [−0.270, 0.036]	−0.089 [−0.280, 0.107]	−0.072 [−0.228, 0.088]
	PrimusM-OM-MAE	−0.211 [−0.390, −0.019]	−0.164 [−0.308, −0.015]	−0.263 [−0.448, −0.069]	−0.212 [−0.361, −0.054]	−0.229 [−0.416, −0.033]	−0.186 [−0.339, −0.025]
	PrimusM-OM-VoCo	−0.157 [−0.351, 0.043]	−0.127 [−0.282, 0.033]	−0.178 [−0.376, 0.037]	−0.144 [−0.306, 0.029]	−0.232 [−0.417, −0.035]	−0.191 [−0.345, −0.030]
	ResEncL-OM-MAE	−0.120 [−0.307, 0.079]	−0.095 [−0.246, 0.063]	−0.168 [−0.354, 0.028]	−0.132 [−0.285, 0.026]	0.003 [−0.199, 0.205]	0.002 [−0.163, 0.168]
	ResEncL-OM-S3D	−0.240 [−0.425, −0.046]	−0.192 [−0.343, −0.037]	−0.273 [−0.459, −0.079]	−0.221 [−0.373, −0.063]	−0.257 [−0.428, −0.075]	−0.209 [−0.349, −0.061]
	ResEncL-OM-VoCo	−0.085 [−0.268, 0.106]	−0.066 [−0.209, 0.082]	−0.140 [−0.326, 0.051]	−0.109 [−0.256, 0.042]	−0.086 [−0.270, 0.102]	−0.070 [−0.218, 0.082]

Table 13: Pairwise bootstrap difference CIs (Kendall's $\tau_i - \tau_j$) underlying CLD groups for expert visual quality ratings (Rater 1+4, bias-corrected). Cell (i, j): 95% percentile CI of $\tau_i - \tau_j$ across bootstrap replicates. CLD letters shown below each matrix.

Quality Overall

	Med3D	**PrimusM-OM-MAE**	**PrimusM-OM-VoCo**	**ResEncL-OM-MAE**	**ResEncL-OM-S3D**	**ResEncL-OM-VoCo**
Med3D	—	[−0.078, 0.182]	[−0.144, 0.200]	[−0.138, 0.208]	[−0.064, 0.251]	[−0.178, 0.227]
PrimusM-OM-MAE	[−0.182, 0.078]	—	[−0.115, 0.074]	[−0.167, 0.131]	[−0.103, 0.195]	[−0.193, 0.141]
PrimusM-OM-VoCo	[−0.200, 0.144]	[−0.074, 0.115]	—	[−0.163, 0.170]	[−0.097, 0.233]	[−0.171, 0.158]
ResEncL-OM-MAE	[−0.208, 0.138]	[−0.131, 0.167]	[−0.170, 0.163]	—	[−0.129, 0.255]	[−0.220, 0.203]
ResEncL-OM-S3D	[−0.251, 0.064]	[−0.195, 0.103]	[−0.233, 0.097]	[−0.255, 0.129]	—	[−0.257, 0.109]
ResEncL-OM-VoCo	[−0.227, 0.178]	[−0.141, 0.193]	[−0.158, 0.171]	[−0.203, 0.220]	[−0.109, 0.257]	—
CLD	a	a	a	a	a	a

Quality Anatomy

	Med3D	**PrimusM-OM-MAE**	**PrimusM-OM-VoCo**	**ResEncL-OM-MAE**	**ResEncL-OM-S3D**	**ResEncL-OM-VoCo**
Med3D	—	[−0.098, 0.149]	[−0.162, 0.166]	[−0.152, 0.172]	[−0.038, 0.285]	[−0.172, 0.208]
PrimusM-OM-MAE	[−0.149, 0.098]	—	[−0.117, 0.068]	[−0.169, 0.131]	[−0.060, 0.260]	[−0.170, 0.160]
PrimusM-OM-VoCo	[−0.166, 0.162]	[−0.068, 0.117]	—	[−0.158, 0.171]	[−0.053, 0.298]	[−0.147, 0.180]
ResEncL-OM-MAE	[−0.172, 0.152]	[−0.131, 0.169]	[−0.171, 0.158]	—	[−0.081, 0.313]	[−0.184, 0.214]
ResEncL-OM-S3D	[−0.285, 0.038]	[−0.260, 0.060]	[−0.298, 0.053]	[−0.313, 0.081]	—	[−0.278, 0.071]
ResEncL-OM-VoCo	[−0.208, 0.172]	[−0.160, 0.170]	[−0.180, 0.147]	[−0.214, 0.184]	[−0.071, 0.278]	—
CLD	a	a	a	a	a	a

Quality Vessels

	Med3D	**PrimusM-OM-MAE**	**PrimusM-OM-VoCo**	**ResEncL-OM-MAE**	**ResEncL-OM-S3D**	**ResEncL-OM-VoCo**
Med3D	—	[−0.095, 0.192]	[−0.162, 0.207]	[−0.186, 0.132]	[−0.077, 0.258]	[−0.192, 0.212]
PrimusM-OM-MAE	[−0.192, 0.095]	—	[−0.115, 0.064]	[−0.225, 0.070]	[−0.097, 0.186]	[−0.203, 0.131]
PrimusM-OM-VoCo	[−0.207, 0.162]	[−0.064, 0.115]	—	[−0.213, 0.107]	[−0.078, 0.216]	[−0.186, 0.161]
ResEncL-OM-MAE	[−0.132, 0.186]	[−0.070, 0.225]	[−0.107, 0.213]	—	[−0.058, 0.293]	[−0.160, 0.243]
ResEncL-OM-S3D	[−0.258, 0.077]	[−0.186, 0.097]	[−0.216, 0.078]	[−0.293, 0.058]	—	[−0.259, 0.100]
ResEncL-OM-VoCo	[−0.212, 0.192]	[−0.131, 0.203]	[−0.161, 0.186]	[−0.243, 0.160]	[−0.100, 0.259]	—
CLD	a	a	a	a	a	a

C.4.2 Association with Annotated Vessel Topology and Pathology Variations

Table 14: Association analysis between metric behaviour and annotated vessel topology and pathology variations (Spearman's ρ and Kendall's τ_b between variation load and FD). More negative values indicate stronger association. 95% percentile bootstrap confidence intervals shown in brackets.

Encoder	ρ	τ_b
Med3D	$0.037\,[-0.164,\ 0.235]$	$0.032\,[-0.112,\ 0.176]$
PrimusM-OM-MAE	$0.004\,[-0.204,\ 0.211]$	$0.001\,[-0.153,\ 0.152]$
PrimusM-OM-VoCo	$-0.136\,[-0.319,\ 0.059]$	$-0.099\,[-0.232,\ 0.036]$
ResEncL-OM-MAE	$-0.105\,[-0.295,\ 0.097]$	$-0.076\,[-0.216,\ 0.069]$
ResEncL-OM-S3D	$-0.044\,[-0.233,\ 0.141]$	$-0.028\,[-0.167,\ 0.105]$
ResEncL-OM-VoCo	$0.180\,[-0.011,\ 0.346]$	$0.131\,[-0.008,\ 0.254]$

Table 15: Pairwise bootstrap difference CIs (Kendall's $\tau_{b,i} - \tau_{b,j}$) underlying CLD groups for annotated vessel and pathology variations. Cell (i, j): 95% percentile CI of $\tau_{b,i} - \tau_{b,j}$ across bootstrap replicates. CLD letters shown below each matrix.

Variation Load

	Med3D	**PrimusM-OM-MAE**	**PrimusM-OM-VoCo**	**ResEncL-OM-MAE**	**ResEncL-OM-S3D**	**ResEncL-OM-VoCo**
Med3D	—	[−0.115, 0.186]	[−0.039, 0.298]	[−0.038, 0.243]	[−0.121, 0.256]	[−0.287, 0.095]
PrimusM-OM-MAE	[−0.186, 0.115]	—	[0.018, 0.178]	[−0.092, 0.225]	[−0.138, 0.191]	[−0.315, 0.052]
PrimusM-OM-VoCo	[−0.298, 0.039]	[−0.178, −0.018]	—	[−0.188, 0.138]	[−0.228, 0.087]	[−0.399, −0.058]
ResEncL-OM-MAE	[−0.243, 0.038]	[−0.225, 0.092]	[−0.138, 0.188]	—	[−0.233, 0.157]	[−0.406, 0.008]
ResEncL-OM-S3D	[−0.256, 0.121]	[−0.191, 0.138]	[−0.087, 0.228]	[−0.157, 0.233]	—	[−0.336, 0.024]
ResEncL-OM-VoCo	[−0.095, 0.287]	[−0.052, 0.315]	[0.058, 0.399]	[−0.008, 0.406]	[−0.024, 0.336]	—
CLD	ab	b	a	ab	ab	b