MID: Mask-Image Distributional Divergence for Evaluating Medical Image Segmentation

Vincenzo Marcianò[1,2,3] **Xiaoming Zhang**[1,2] **Gianluca Guglielmo**[4]
Michela Antonelli[3] **Sébastien Ourselin**[3] **Maria A. Zuluaga**[1,3]

[1]EURECOM [2]Sorbonne Université [3]King's College London [4]Technische Universität Graz

{vincenzo.marciano, xiaoming.zhang, maria.zuluaga}@eurecom.fr
{michela.antonelli, sebastien.ourselin}@kcl.ac.uk gianluca.guglielmo@tugraz.at

Abstract

Current evaluation protocols for medical image segmentation remain largely anchored in sample-wise comparisons, e.g., Dice and Hausdorff distance, which require each prediction to be paired with a ground-truth mask. Yet two limitations arise in practice: they are inapplicable when predictions are made on unlabeled data; and even when annotations are available, they cannot assess whether a collection of predictions matches the distributional structure of the ground-truth population. We introduce MID (Mask-Image Distributional Divergence), a distribution-level evaluation metric for medical image segmentation. MID builds upon the observation that segmentation quality is a property of the joint image-mask distribution: a mask may be anatomically plausible in isolation yet incorrect for the image it accompanies. MID compares the joint distribution of images and masks between a reference set and a set of predictions, without requiring sample-wise correspondence. This captures both mask realism and image-mask consistency. We evaluate MID across 13 publicly available datasets spanning CT, MRI, and MRA. Under anatomically realistic corruptions across 56 organ groups, MID tracks mean Dice and Hausdorff distance without per-sample ground-truth pairing. MID also detects distributional shifts across modalities and anatomies, identifies image-mask misalignment through controlled shuffling experiments, and produces reliable model rankings across real segmentation models. Official implementation available at `github.com/erc-caravel/pytorch-mid`.

1 Introduction

Deep learning has substantially advanced automatic medical image segmentation, yielding models that range from single-anatomy specialists [18, 31, 16] to pan-organ foundation models [19, 26, 27, 40, 43]. Evaluating these models remains fundamentally tied to per-sample ground-truth (GT) annotations at test time. For instance, the Dice similarity coefficient, Hausdorff distance (HD95), and the normalised surface distance (NSD) [29] measure how accurately a predicted mask reproduces the paired GT annotation at the voxel level. Although they are effective in controlled experimental conditions, several practical challenges remain to be addressed. First, every prediction needs the corresponding GT annotation. This assumption fails in prospective deployment and for foundation models on unlabeled cohorts. Second, when GT annotations are unavailable, an alternative is to generate surrogate references using encoder-decoder refinement models [8, 14, 23, 30, 32, 41]. However, these approaches require training additional models for each target dataset, which is computationally expensive and limits scalability.

Perceptual image quality assessment has similarly moved beyond pixel-level comparison toward feature-level distances in pretrained encoder spaces. Learned Perceptual Image Patch Similarity

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

(LPIPS) [45] showed that ℓ_2 distances between VGG features of two images correlate with human perceptual judgements far better than Structural Similarity Index Measure (SSIM) or pixel norms, because pretrained encoders capture semantic structure rather than low-level coincidences. This idea extends naturally from per-sample distances to population-level comparisons. Once samples are embedded in a feature space, a dataset induces an empirical distribution, and the squared 2-Wasserstein distance W_2^2 provides a principled way to compare such distributions: it respects the geometry of the embedding space and admits a closed-form expression for Gaussian distributions [11]. The Fréchet Inception Distance (FID) [17] applied this principle to generative image evaluation: images are embedded into Inception-v3 features, Gaussian distributions are fitted to the reference and generated sets, and the squared Fréchet distance, equivalent to W_2^2 between the two fitted Gaussians, is reported as a single scalar. FID requires no per-sample correspondence between reference and candidate sets, correlates with human judgements of image quality, and has become the de facto standard in generative modelling.

Feature-space distribution matching offers a promising route toward segmentation evaluation without per-sample ground truth. However, extending this paradigm from image generation to medical image segmentation is non-trivial. First, segmentation quality is inherently defined over the *joint* distribution of images and masks: a mask may be anatomically plausible in isolation, yet clinically incorrect if it is misaligned with the anatomy in the corresponding image. Image-only distributional metrics cannot capture such failures. Second, the intensity distributions of medical images differ across modalities. CT intensities follow the calibrated Hounsfield scale, while MRI intensities have no fixed unit and change with the scanner and the sequence. A general-purpose metric should compare image-mask distributions across these settings without relying on modality-specific pipelines.

We propose MID (Mask-Image Distributional Divergence), a distribution-level metric for medical image segmentation that shifts evaluation from sample-wise mask comparison to probability-measure comparison over the joint space $\mathcal{X} \times \mathcal{M}$. Specifically, MID maps each image and its class-specific mask to a joint representation through a pretrained medical vision encoder, thereby inducing two empirical push-forward distributions: one from ground-truth image-mask pairs and one from predicted image-mask pairs. MID then fits Gaussian approximations to these embedded distributions and reports their squared 2-Wasserstein distance. By comparing distributions rather than aligned samples, MID requires no per-case correspondence between reference and candidate sets, while remaining sensitive to the image-mask dependence that defines segmentation quality.

This construction adapts feature-space distribution matching to the structured image-mask dependency inherent in segmentation. MID captures failure modes invisible to single-modality metrics: unlike image-only approaches, it detects mask-image inconsistency; unlike mask-only metrics, it identifies anatomically implausible placements even when mask shapes are realistic.

Across 56 organ groups, using five signed-distance-transform (SDT)-based corruption types at three severity levels, MID increases consistently as segmentation quality degrades. Without any per-sample correspondence, MID correlates with mean Dice (Spearman $\rho = -0.55$) and with HD95 ($\rho = 0.72$). Dice and HD95 need a ground-truth mask for each prediction. MID needs ground-truth masks only for the reference set. Per-sample metrics stay more precise when ground truth is available. MID adds a complementary distribution-level signal for candidate sets without annotations.

Our contributions are as follows: (1) we formulate segmentation evaluation without paired ground truth as distribution matching on the joint image-mask space and introduce MID; (2) we propose a diagnostic decomposition of MID into image, mask, and residual correspondence terms; and (3) across thirteen datasets, we show that MID tracks corruption severity, detects distribution shifts, identifies image-mask misalignment, and ranks models consistently with paired metrics.

2 Related Work

Per-sample segmentation metrics. The Metrics Reloaded framework [29] catalogues overlap measures (Dice, IoU), surface distances (HD95, NSD, ASSD), and detection metrics. Topology-preserving variants include clDice [37] and topologically faithful segmentation [39]. All of these metrics require per-sample GT alignment at test-time.

FID and distribution-level evaluation. FID [17] embeds images into Inception-v3 features and computes the squared Fréchet distance [11], which equals W_2^2 between the fitted Gaussians. It

correlates with human quality judgements at a scale that per-image metrics cannot. Kernel Inception Distance (KID) [5] provides an unbiased Maximum Mean Discrepancy (MMD)-based alternative. FID has been extended to joint image-caption embeddings [10] by concatenating image and conditioning features before fitting the Gaussian; however, this setting is limited to 2-D natural images and does not model structured segmentation masks or volumetric data. MID addresses these limitations by jointly encoding images and masks, supporting both 2-D and 3-D inputs, and enabling a decomposition into image, mask, and residual correspondence terms.

Medical imaging representation learning. Recent work shows that learned feature spaces for medical images organise by anatomy and generalise across datasets and modalities [34, 33]. Early self-supervised and generative approaches demonstrated this structure without label supervision, with masked autoencoders and latent diffusion models recovering clinically relevant representations in brain MRI and abdominal CT [3, 15, 34]. Scaling vision-only models, such as Rad-DINO [33], and contrastive image-text pretraining, including BiomedCLIP [46] and UniMedCLIP [22], further improved generalisation across anatomies, scanner types, and modalities. More recent large-scale medical vision encoders [36], built on SigLIP [44], produce shared feature spaces spanning CT, MRI, pathology, and fundus photography. Feature-space distances from such encoders are widely used for out-of-distribution detection, e.g., via Mahalanobis distance [25]. However, these approaches operate on images alone and do not model image-mask correspondence, a key aspect of segmentation evaluation.

3 Method

MID evaluates segmentation quality through the joint image-mask distribution rather than image or mask marginals alone. It is constructed in three steps. First, for each image-mask pair, the image is encoded once, and each foreground class is encoded as a separate binary mask using the same medical vision encoder for both the image and mask inputs, producing one joint vector per foreground class. Second, the resulting vectors are projected to a lower-dimensional space via PCA fitted on the reference set. Finally, the squared 2-Wasserstein distance between Gaussian approximations of the reference and candidate distributions is computed. Sections 3.1 to 3.4 formalise these steps.

3.1 Setup and Joint Embedding

Let $\mathcal{X}$ denote the space of 2-D and 3-D medical images, and let $\mathcal{M}$ denote the space of C-class segmentation masks defined on the same grid. For a segmentation mask $m \in \mathcal{M}$, let $\mathcal{C}(m) \subseteq \{1, \dots, C\}$ denote the set of foreground classes present in m. For each $c \in \mathcal{C}(m)$, let $m_c = \mathbf{1}[m = c]$ be the corresponding binary mask, so that m is represented as the collection of class-specific binary masks $\{m_c\}_{c \in \mathcal{C}(m)}$.

Let f_θ denote a medical vision encoder that maps an input image or mask rendering to a d-dimensional feature vector in $\mathbb{R}^d$. For a 2-D input, f_θ encodes the image directly. For a 3-D volume, f_θ encodes N uniformly sampled axial slices and averages the slice embeddings. Both cases give vectors in the same space $\mathbb{R}^d$. One joint distribution can thus hold 2-D and 3-D samples. The experiments in this paper use 3-D data.

Definition 1 (Per-class joint feature embedding)**.** *For an image-mask pair $(X, m) \in \mathcal{X} \times \mathcal{M}$ and foreground class $c \in \mathcal{C}(m)$, the per-class joint feature embedding is defined as*

$$\phi_c(X, m) = \pi\big([f_\theta(X) \parallel f_\theta(m_c)]\big) \in \mathbb{R}^k, \tag{1}$$

where $[\cdot \| \cdot]$ denotes the concatenation of the image and mask embeddings and $\pi : \mathbb{R}^{2d} \to \mathbb{R}^k$ is a projection.

Each $\phi_c(X, m)$ represents the pairing between image X and the class-specific binary mask m_c. For a dataset $\{(X_i, m_i)\}_{i=1}^N$, each image-mask pair contributes one such embedding for every foreground class present in the mask. The resulting collection $\{\phi_c(X_i, m_i)\}_{i=1,\dots,N;\, c \in \mathcal{C}(m_i)}$ forms the embedding set over which MID fits its Gaussian.

In practice, the image embedding is computed once per volume and reused across foreground classes. Class masks are encoded as binary masks with a shared foreground label, so their embeddings capture geometry rather than class identity. The embedding is deterministic, ensuring reproducibility without requiring fixed random seeds.

3.2 MID

Let P be a reference distribution over ground-truth image-mask pairs and Q a candidate distribution over predicted pairs. Following Section 3.1, each image-mask pair contributes one embedding per foreground class present in the mask. The empirical distributions for P and Q are obtained by pooling these embeddings before fitting Gaussian approximations. Let (μ_P, Σ_P) and (μ_Q, Σ_Q) be the empirical means and covariances of the projected embeddings $\{\phi_c(X_i, m_i)\}_{i=1,\dots,N;\, c\in\mathcal{C}(m_i)}$ under P and Q.

Definition 2 (MID)**.**

$$\begin{aligned} \mathit{MID}(P,Q) &= W_2^2\big(\mathcal{N}(\mu_P,\Sigma_P),\,\mathcal{N}(\mu_Q,\Sigma_Q)\big) \\ &= \|\mu_P-\mu_Q\|^2 \\ &\quad + \mathrm{Tr}\Big(\Sigma_P+\Sigma_Q-2\big(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2}\big)^{1/2}\Big). \end{aligned} \tag{2}$$

The closed-form expression in Equation (2) follows from Dowson & Landau [11] and corresponds to the squared 2-Wasserstein distance between Gaussian approximations of the two embedding distributions, as used in Fréchet Inception Distance (FID) [17]. P and Q need not be aligned sample by sample, as MID compares distributions rather than individual pairs.

3.3 Projection Design

The joint embedding has dimension $2d$, which is high relative to the size of typical medical imaging datasets. In this setting, reliable covariance estimation is challenging and requires dimensionality reduction. To this end, the embeddings are projected to a lower-dimensional space using PCA fitted on the reference set. This projection preserves the dominant directions of variation while improving the stability of the covariance estimates required for MID. In practice, we use $k = 128$ components, which retain more than $94\,\%$ of the variance in the reference set. An orthogonal projection cannot increase the 2-Wasserstein distance between the two Gaussian fits. Thus $\mathrm{MID}_k \le \mathrm{MID}$. The gap grows with the variance and the mean shift of both distributions in the discarded directions. For the reference set, these directions hold less than $6\,\%$ of the variance.

3.4 Sensitivity Decomposition

The joint MID score can be further analysed by separating image appearance, mask geometry, and image-mask correspondence. We compute two auxiliary Fréchet distances using the same estimator as in Equation (2): an image-only term $\mathrm{MID}_{\mathrm{img}}$ computed from image embeddings $f_\theta(X)$, and a mask-only term $\mathrm{MID}_{\mathrm{mask}}$ computed from mask embeddings $f_\theta(m_c)$. Concretely, $\mathrm{MID}(P,Q)$ is computed on joint embeddings, $\mathrm{MID}_{\mathrm{img}}(P,Q)$ on image-only embeddings, and $\mathrm{MID}_{\mathrm{mask}}(P,Q)$ on mask-only embeddings, each using the same Gaussian Fréchet estimator and reference-fitted projection protocol. The remaining contribution is attributed to image-mask correspondence and defines the residual correspondence term:

$$\mathrm{MID}_{\mathrm{corr}}(P,Q) \;:=\; \mathrm{MID}(P,Q) - \mathrm{MID}_{\mathrm{img}}(P,Q) - \mathrm{MID}_{\mathrm{mask}}(P,Q). \tag{3}$$

Unlike $\mathrm{MID}_{\mathrm{img}}$ and $\mathrm{MID}_{\mathrm{mask}}$, $\mathrm{MID}_{\mathrm{corr}}$ is a diagnostic residual rather than a standalone distance. It captures the part of the joint score not explained by image-only or mask-only shifts. Random image-mask shuffling keeps the image-only and mask-only distributions. It changes only their dependence. Shuffling therefore leaves $\mathrm{MID}_{\mathrm{img}}$ and $\mathrm{MID}_{\mathrm{mask}}$ unchanged and increases $\mathrm{MID}_{\mathrm{corr}}$. The derivation of Equation (3) is given in Appendix G.1.

4 Experiments

We evaluate MID through a series of controlled and real-world experiments. Experiment 1 assesses sensitivity to controlled, anatomically realistic segmentation corruptions. Experiment 2 examines cross-dataset distributional shift. Experiment 3 compares MID with distributional baselines, using Dice and HD95 as paired targets. Experiment 4 evaluates MID on outputs from real segmentation models. Experiment 5 isolates the role of image-mask correspondence via controlled perturbations.

Section 4.7 assesses estimator reliability as a function of sample size and derives usage guidance. We also provide additional analyses on within-group spread (Appendix D), dataset-specific projections (Appendix I), reference-set size and label quality (Appendix J), the Gaussian-mixture reference variant (Appendix K), and latent space geometry (Appendix M).

4.1 Setup

Datasets. We use thirteen publicly available datasets spanning CT, MRI, and magnetic resonance angiography (MRA): AbdomenCT-1K [28], ACDC [4], AMOS CT and MRI [21], IXI MRA with VesselVerse annotations [6, 12], six Medical Segmentation Decathlon tasks (Task03 Liver, Task05 Prostate, Task06 Lung, Task07 Pancreas, Task09 Spleen, Task10 Colon) [1], and TotalSegmentator v2 CT and MRI [42]. Dataset-specific foreground classes and sample counts are reported in Appendix A.

Reference distribution. Throughout the paper we count multimodal datasets by modality (so AMOS contributes AMOS CT and AMOS MRI, and TotalSegmentator v2 contributes TotalSeg CT and TotalSeg MRI), giving the 13 datasets listed above. Unless otherwise stated, the reference distribution is built from clean pairs of images and ground-truth masks in the reference split of a 10-dataset reference pool: ACDC, IXI MRA (VesselVerse), the six MSD tasks (Task03 Liver, Task05 Prostate, Task06 Lung, Task07 Pancreas, Task09 Spleen, Task10 Colon), and TotalSegmentator v2 CT and MRI. The remaining three datasets, AbdomenCT-1K, AMOS CT, and AMOS MRI, are excluded from the reference pool and serve only as candidate sets.

Candidate distributions. Candidate distributions are experiment-specific and are compared against the fixed reference distribution. We construct candidate distributions from two sources. First, controlled synthetic corruptions are applied to ground-truth masks to produce candidate masks with known degradation type and severity. Following controlled experimental designs in [2, 20], we use five signed-distance-transform (SDT) based corruption types: elastic boundary deformation, smooth erosion, smooth dilation, blob-shaped false positives, and partial missing slices, each at mild, moderate, and severe levels (Appendix C). Second, we use predictions from real segmentation models. SwinUNETR [16] and MedSAM [26] are kept frozen and used directly for inference: SwinUNETR is applied to abdominal CT datasets, while MedSAM is applied across non-vascular datasets using synthetic prompts derived from ground-truth masks. SegResNet [31] is trained on all non-vascular datasets and evaluated at 50, 150, and 300 epochs, yielding predictions at increasing quality levels. For vascular segmentation on IXI MRA, we use model predictions released with VesselVerse: nnUNet [18], A2V [13], and SPOCKMIP [35].

Baselines. We compare MID against five distributional baselines. **FID-Inception** computes standard FID on Inception-v3 [17] features of axial mid-slice renderings, a 2-D proxy that discards volumetric information. **FID-img** computes the Fréchet distance on image features from f_θ only, without masks. **FID-mask** computes the Fréchet distance on mask features from f_θ only, without images. **Mahalanobis** computes the mean squared Mahalanobis distance from a reference Gaussian in the joint feature space. **KID** [5] computes the unbiased Kernel Inception Distance on MedSigLIP mask features. **Paired reference targets** are mean Dice and HD95 over paired predictions and ground-truth masks.

For the main experiments, we instantiate f_θ with MedSigLIP [36]. For 3-D volumes, we take the ℓ_2-normalised pooled MedSigLIP embedding of 16 uniformly sampled axial slices and average the slice embeddings; for MAISI-based encoders (Appendix N), each 2-D axial slice is encoded as a 3-D input with depth 1. Alternative encoder choices are evaluated in Appendix N. Appendix B lists all preprocessing and estimator settings. We pre-extract MedSigLIP features for all datasets, yielding 17 631 per-class image-mask feature representations (of which 11 223 belong to the 10-dataset reference pool and 6408 to the held-out AMOS CT, AMOS MRI, and AbdomenCT-1K candidate sets). Unless otherwise stated, MID and the distribution-based baselines are computed using the same encoder f_θ, with feature inputs selected according to each metric. We use incremental PCA with $k = 128$ components, fitted on the reference set. Covariance is estimated via Ledoit–Wolf shrinkage [24]. All code was implemented in Python 3.11 using PyTorch 2.11 and CUDA 12.4. All experiments use a fixed random seed. Feature extraction was run on NVIDIA A100 80 GB GPUs, and all metric computations on CPU nodes.

Table 1: Spearman correlation between MID and Dice by corruption type, over 56 organ groups $\times$ 3 severities (168 cells per row). 95 % bootstrap CIs use 1000 replicates.

Corruption type	ρ vs. Dice	95 % CI
Smooth erosion	-0.65	$[-0.71, -0.58]$
Smooth dilation	-0.62	$[-0.68, -0.55]$
Elastic deformation	-0.58	$[-0.64, -0.51]$
Blob false positives	-0.45	$[-0.52, -0.38]$
Partial missing slices	-0.42	$[-0.49, -0.34]$
All (840 cells)	-0.55	$[-0.61, -0.49]$

Table 2: MID between TotalSeg CT train (reference, 1105 cases, 6630 vectors) and candidate datasets, ordered by score. Mahalanobis distance (Mah.) shown for comparison.

Candidate	Mod.	MID	Mah.
TotalSeg CT (held-out)	CT	0.011	11.2
TotalSeg MRI	MRI	0.278	34.7
Prostate MRI	MRI	0.932	48.8
ACDC cardiac	MRI	0.949	46.6

4.2 Response to Controlled Segmentation Corruptions

Experiment 1 assesses whether MID responds monotonically to controlled, anatomically realistic mask corruptions of known severity. For each combination of dataset, foreground class, corruption type, and severity level, the candidate distribution is built from the evaluation split by replacing ground-truth masks with corrupted masks. MID is computed against the global reference distribution defined in Section 4.1, without sample-wise pairing between reference and candidate embeddings. Dice and HD95 are computed on the same evaluation set and serve as paired reference-quality indicators. The quantitative analysis is performed over all such combinations in the full corruption sweep.

Figure 1 shows three representative cases from the sweep, selected to span different datasets, anatomies, and degradation patterns, and one real model at three training stages. In each panel, MID follows the change in Dice and HD95 without using per-case ground truth at inference time. MID is normalised by the panel-wise maximum for visualisation only; all reported correlations use raw values. Across the full 840-cell sweep, MID tracks both paired-reference indicators in the expected direction: it is negatively correlated with mean Dice (Spearman $\rho = -0.55$, $p < 0.001$), since lower MID and higher Dice indicate better segmentations, and positively correlated with HD95 ($\rho = 0.72$, $p < 0.001$). Within each foreground class, MID increases monotonically with corruption severity across all five corruption types. Disaggregating by corruption type provides direct evidence that shape-altering corruptions yield stronger correlations with Dice than additive false-positive or missing-slice errors (Table 1). Smooth erosion, smooth dilation, and elastic boundary deformation reach ρ between -0.58 and -0.65, whereas blob false positives and partial missing slices reach -0.45 and -0.42. This indicates that MID responds not only to voxel-level disagreement but also to the type of anatomical perturbation.

The corruption sweep also reveals that the scale of MID varies substantially across anatomies, spanning roughly an order of magnitude between compact structures (e.g., adrenal glands, vessels) and large organs (e.g., liver, lung lobes). As a result, absolute MID values should be interpreted relative to anatomy-specific reference ranges, rather than compared directly across anatomies. Additional analyses, including the per-anatomy reference index (Appendix D.2, Figure 7) and clean-reference variability (Appendix D.1), are reported in the appendix, and the sample-size behaviour of the estimator in Section 4.7.

4.3 Distributional Shift Detection

MID compares a candidate distribution of image-mask pairs to a fixed reference distribution. It should therefore increase when the candidate set differs systematically from the reference, whether through changes in anatomy, modality, acquisition quality, or predicted mask geometry. In Experiment 2, we test two complementary modes of such distributional shift: cross-dataset shifts in anatomy or modality, and within-dataset shifts induced by image-acquisition artefacts.

Cross-dataset shift. We fix a single reference distribution from TotalSegmentator CT train (1105 cases, 6630 latent vectors across six abdominal organs: spleen, left kidney, right kidney, liver, stomach, and pancreas) and measure MID against four candidate sets spanning in-distribution, modality-shift, and anatomy-shift settings, with the corresponding scores reported in Table 2. The in-distribution control is the TotalSegmentator CT held-out test split (123 cases, 738 vectors), drawn from the same

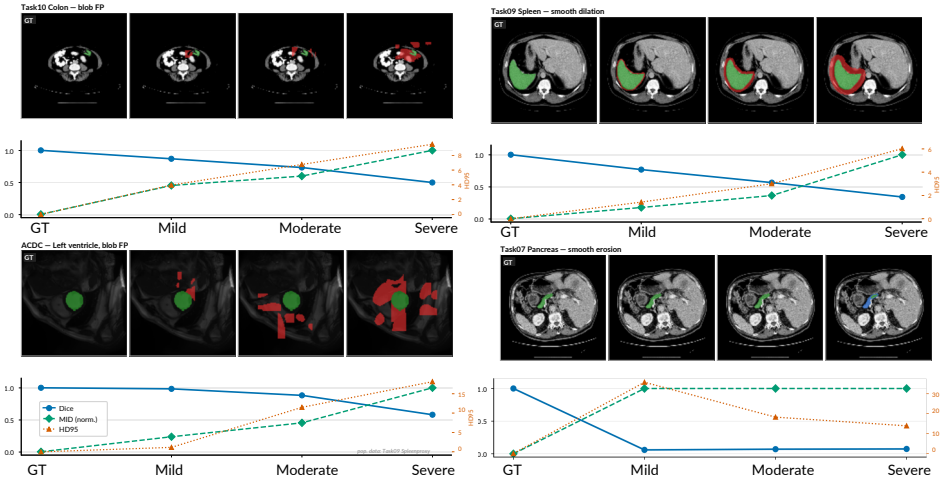

Figure 1: Representative examples from the controlled corruption sweep (Experiment 1) and from a real segmentation model. The four panels show Task10 Colon with blob-shaped false positives, Task09 Spleen with smooth dilation, ACDC left ventricle with blob-shaped false positives, and Task07 Pancreas with SegResNet predictions after 50, 150, and 300 training epochs. Overlays show the ground-truth mask and the corrupted or predicted masks. Line plots show MID, Dice, and HD95 across severity levels or training epochs. In the ACDC panel, the line plots use Task09 Spleen population data as a proxy. MID is normalised by the panel-wise maximum for visualisation only; all reported correlations use raw values.

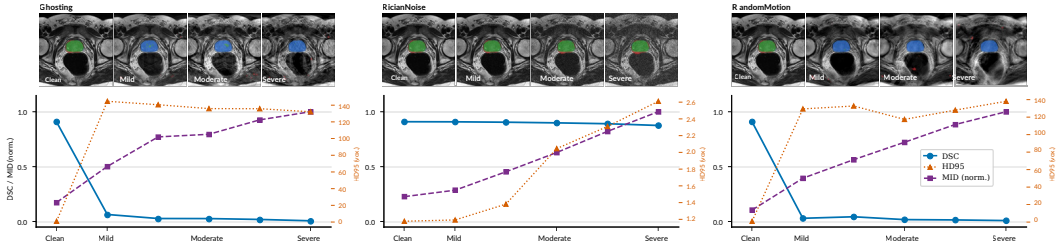

Figure 2: ROOD-MRI prostate on Task05_Prostate using SegResNet-300. For each artefact type, representative T2 slices across severity levels are shown above, with population metrics below.

dataset and acquisition protocol as the reference. The resulting $\mathrm{MID} = 0.011$ provides an empirical finite-sample floor: even a held-out split from the same distribution yields a small non-zero score. All shifted candidates are interpreted relative to this baseline. TotalSegmentator MRI increases MID to 0.278, a $25\times$ increase over the floor, reflecting a modality shift with the same abdominal anatomy. Prostate MRI and ACDC cardiac MRI produce larger shifts ($\mathrm{MID} = 0.932$ and 0.949), approximately $85\times$ above the floor, as anatomy, modality, and acquisition protocol all differ from the reference.

Within-dataset shift via image artefacts. We next apply ROOD-MRI [7], a robustness benchmark for controlled MRI acquisition artefacts, to Task05_Prostate. Three artefact types (Ghosting, Rician noise, Random Motion) are applied at three severity levels (mild, moderate, severe). The reference distribution is pre-computed from clean T2 images with ground-truth masks; candidate distributions are built from corrupted images paired with SegResNet-300 predictions, without using ground-truth masks at inference time. Additional setup details are provided in Appendix E.2.

Figure 2 shows that MID tracks the degradation monotonically across all three artefact types and all severity levels, mirroring the Dice and HD95 trends without requiring any per-case ground truth at inference time. This setting differs from the controlled corruption experiment (Section 4.2), where image quality is fixed and only masks are corrupted. Here, both image appearance and predicted masks shift with artefact severity, and MID captures the combined distributional change as a single divergence from the clean reference.

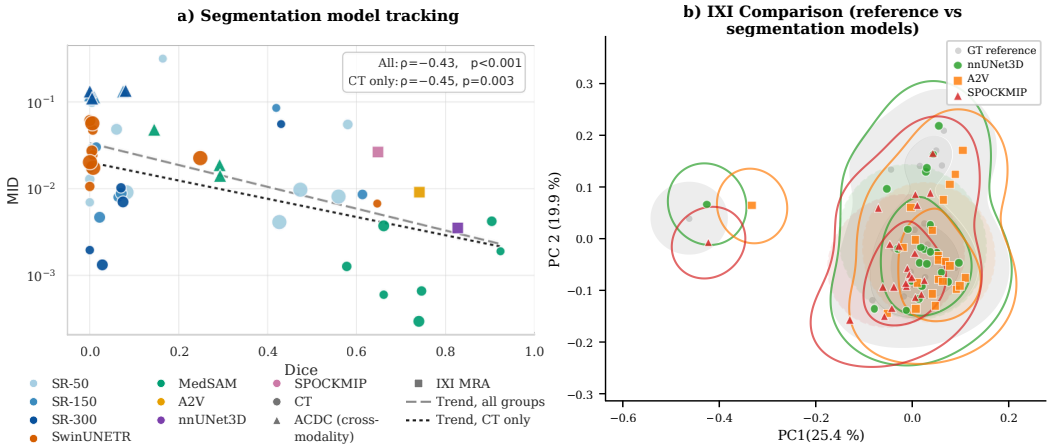

Figure 3: **a)** MID vs. Dice across real segmentation model outputs. Points are coloured by model; y-axis is log-scaled. **b)** Ground-truth and predicted-mask distributions for three IXI MRA vessel segmentation methods. The axes are the first two principal components of the ground-truth reference. Lines are iso-density contours. Legend values are MID in the full $k = 128$ space.

4.4 Metric Comparison

We compare scalar MID against distributional baselines and per-sample metrics on the same experimental setup as Experiment 1 (56 organ groups $\times$ 5 corruption types $\times$ 3 severities $=$ 840 cells). Standard FID applied to Inception-v3 features of axial mid-slice renderings achieves $\rho = -0.38$ with Dice. FID-mask reaches $\rho = -0.47$. FID-img is essentially zero ($\rho = -0.03$) because images are held fixed across the sweep and only masks are corrupted, confirming that mask information is needed to track segmentation quality. Mahalanobis distance from the joint reference Gaussian reaches $\rho = -0.42$. Joint MID reaches $\rho = -0.55$ with Dice, with an absolute gain of 0.08 over FID-mask (overlapping CIs), and $\rho = 0.72$ with HD95 without any per-sample pairing. Full Spearman correlations with 95 % bootstrap CIs (1000 replicates) for all metrics are in Table 6 (Appendix F). FID-mask, the closest distributional baseline, operates on mask features alone and cannot capture image-mask agreement. A within-organ paired-cell analysis removes the between-organ scale effects. Under this analysis, MID improves on FID-mask at $p < 0.001$ (Appendix L).

4.5 Real Segmentation Models

We evaluate MID on real model predictions in two settings: heterogeneous model tracking against the global reference distribution, and vessel-model ranking on IXI MRA. For the first setting, we use SegResNet at three training stages (SR-50, SR-150, SR-300), frozen SwinUNETR and MedSAM predictions, on Task03 Liver, Task06 Lung, Task07 Pancreas, Task09 Spleen, Task10 Colon, and ACDC. For each group of model, dataset and foreground class, the candidate distribution is built from disjoint evaluation images paired with model predictions and compared against the global reference pool. Dice and HD95 are computed on the same cases as paired-reference quality indicators.

Figure 3a shows MID against Dice for these groups. On CT datasets, MID decreases as SegResNet training progresses (SR-50 $\rightarrow$ SR-150 $\rightarrow$ SR-300), and frozen models cluster at lower MID and higher Dice. Across CT groups, MID correlates negatively with Dice (Spearman $\rho = -0.45$); all groups together, with ACDC and IXI, give $\rho = -0.43$ ($p < 0.001$). ACDC lies away from the CT trend. The global reference pool is mostly CT, and cardiac MRI is a small part of it. The elevated MID of ACDC thus mixes segmentation quality with this modality imbalance. For IXI MRA, we use VesselVerse predictions from nnUNet, A2V, and SPOCKMIP [12]. Figure 3b shows the ground-truth distribution and the three predicted-mask distributions. MID ranks the three vessel segmentation methods in the same order as Dice: nnUNet has the lowest MID (0.0035) and highest Dice (0.83), followed by A2V (0.0091, Dice 0.74) and SPOCKMIP (0.026, Dice 0.65). HD95 agrees for the two methods with reported values: 4.6 mm for nnUNet and 20.7 mm for SPOCKMIP. MID thus recovers the model ranking without per-case annotations at inference time. Additional details in Appendix H.



Figure 4: MID decomposition across corruption types on TotalSegmentator v2 CT. **(a)** Stacked bars show the image, mask, and residual correspondence terms for each corruption type and severity level (m = mild, M = moderate, S = severe). **(b)** Component trends as severity increases.

4.6 Image-Mask Correspondence

We use the diagnostic decomposition from Section 3.4 to test whether MID separates image appearance, mask geometry, and image-mask correspondence. The main correspondence test uses two controlled TotalSegmentator CT settings. Scenario A corrupts masks while preserving image-mask pairing. Scenario B randomly shuffles image-mask pairs, preserving image-only and mask-only distributions but disrupting correspondence. Both use the same TotalSegmentator CT reference distribution and $n = 300$ feature vectors per organ. Figure 4 visualises the per-corruption component breakdown for Scenario A. Under pair shuffling, $\mathrm{MID}_{\mathrm{img}}$ and $\mathrm{MID}_{\mathrm{mask}}$ keep their unshuffled values. The residual correspondence term carries the increase. This isolates image-mask dependence: the marginal feature distributions are unchanged, but the joint distribution is not. Thus, $\mathrm{MID}_{\mathrm{corr}}$ detects broken image-mask correspondence even when image-only and mask-only statistics are preserved. As an attribution check, Scenario C measures cross-modality shift between TotalSegmentator CT and AMOS MRI over six shared abdominal organs: spleen, left kidney, right kidney, liver, stomach, and pancreas. Here, $\mathrm{MID}_{\mathrm{img}}$ is the largest term. This result agrees with an image appearance shift between modalities. Additional decomposition details are in Appendix G.

4.7 Estimator Reliability and Usage Guidance

MID relies on Gaussian statistics estimated in a projected feature space, where reliability is driven by the number of samples available per group. We quantify that dependence and convert it into usage guidance for typical low-sample datasets in the medical imaging domain.

Sample size and sample control. Table 3 reports the mean and coefficient of variation (CV) of MID as a function of candidate set size n, over 50 bootstrap replicates per n in a representative cross-dataset setting (TotalSeg CT reference vs. TotalSeg MRI candidate), with the reference set fixed at $n = 500$. With this smaller reference, MID approaches 0.30 at large n. Table 2 reports 0.278 for the full reference. The mean decreases monotonically with n, consistent with the finite-sample positive bias of the Fréchet estimator, and approaches its large-n value for $n \gtrsim 250$ (0.34 at $n = 250$, 0.30 at $n = 1000$); the CV decreases roughly as $n^{-1/2}$, as expected for a moment-based estimator. Comparisons are therefore most reliable when candidate sets are of similar size, or when the protocol fixes the candidate size across methods. The complementary question of how small the *reference* set may be, and how tolerant MID is to mislabelled reference cases, is addressed in Appendix J.

We test whether sample-count variation drives the comparison between MID and the baselines in Section 4.4. For each organ, we subsample all candidate cells to the smallest available count and compute the correlations again. The comparison between MID and the baselines keeps the same conclusions. Sample count thus does not drive the comparison.

Usage guidance. We recommend the following operating points. Use MID as a standalone score with at least 250 representations per group. Between 100 and 250 representations, subsample all candidate sets to the smallest size before comparison. Below 100, use MID for exploratory analysis

Table 3: Sample-size sensitivity of MID in a cross-dataset setting (TotalSeg CT reference vs. TotalSeg MRI candidate; MID ≈ 0.30 at large n). The reference set is fixed at $n = 500$. Mean and coefficient of variation (CV) are computed over 50 bootstrap replicates. The decreasing mean reflects the finite-sample positive bias of the squared 2-Wasserstein distance between Gaussian fits.

Statistic	Candidate set size n						
	10	25	50	100	250	500	1000
Mean MID	1.85	0.92	0.58	0.42	0.34	0.31	0.30
CV	32%	20%	14%	10%	6.3%	4.5%	3.2%

or for comparisons at a fixed candidate size. The same threshold applies to the reference set, where the bias falls below $25\,\%$ at 250 representations (Appendix J). In Section 4.5, the IXI comparison uses one fixed candidate size for the three methods. The other real-model groups have unequal sizes, and some have fewer than 100 representations. Their trend is exploratory. To rank models or to track degradation under a fixed reference, use the default single-Gaussian estimator. To calibrate absolute values across heterogeneous modes, use a mixture estimator (Appendix K). Interpret MID against an anatomy-specific reference range (Appendix D.2).

5 Discussion

The experiments show that MID provides a consistent distribution-level quality signal across controlled corruptions, real model outputs, and cross-dataset shifts. Its correlation with Dice and HD95 under controlled corruptions ($\rho = -0.55$ and 0.72) holds without requiring per-sample GT pairing, confirming the metric's practical utility in deployment settings where annotations are unavailable. The response to distributional shift is graded. MID is 0.011 for same-distribution held-out data, 0.278 for a modality shift, and 0.932 to 0.949 for combined anatomy and modality shifts. The main limitation is the Gaussian approximation: when the goal is within-anatomy monitoring, anatomy-specific references may be preferable. The anatomy-dependence of the absolute scale is a related consequence: MID values for the liver and for the adrenal gland are not directly comparable, and the metric should be interpreted against anatomy-specific reference ranges (Appendix D.2). Moreover, we highlight the following behavior: a pooled CT and MRI reference is bimodal, therefore in a synthetic bimodal setting, a two-component mixture estimator gives nearly the same ranking of candidates as the single Gaussian (Appendix K). It can inherit demographic, scanner, or acquisition biases from that pool. Interpret absolute values with care for a heterogeneous reference. MID depends on the quality of the underlying encoder. The choice of MedSigLIP is motivated empirically and by its large-scale medical pretraining, but the framework is encoder-agnostic: as stronger medical vision encoders become available, MID can adopt them without changes to the metric definition or estimation procedure. Computationally, feature extraction is the bottleneck. Once features are pre-extracted (approximately 11 seconds per volume on a single GPU), fitting PCA and computing MID takes under one minute on a CPU node for the full 11 223-vector reference set, making the metric practical for routine evaluation. A conditional variant that conditions the reference distribution on anatomy class would allow cross-anatomy comparison without restricting the reference set. Extending the validation to inherently 2-D modalities (fundus photography, histopathology, X-ray) is a natural next step.

6 Conclusion

We have introduced MID, a distribution-level evaluation metric for medical image segmentation. MID embeds image-mask pairs jointly through a pretrained medical vision encoder, fits Gaussian approximations of the reference and candidate distributions, and reports their squared 2-Wasserstein distance. The underlying divergence is non-negative. The estimator fits PCA on the reference and is therefore asymmetric in P and Q. Appendix O reports a symmetric union-PCA variant and a bound on the asymmetry gap. A diagnostic decomposition into image, mask, and residual correspondence terms is available for failure-mode analysis. Across 13 datasets spanning CT, MRI, and MRA, MID tracks per-sample geometric metrics under controlled corruptions, detects cross-dataset and cross-modality distributional shifts, and segmentation methods in the same order as Dice without per-case annotations at inference time.

Acknowledgements

Funded by the European Union (ERC, CARAVEL, 101171357). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work has been supported by the French government, through the France 2030 plan managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001, and by the ANR-BMBF project TRAIN (ANR-22-FAI1-0003-02). This work was granted access to the HPC resources of IDRIS under the allocation 2026-A0200617550 made by GENCI. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC Development Access call (EHPC-DEV-2026D01-008).

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

[2] Guilherme Aresta, Hrvoje Bogunović, et al. FAZ segmentation quality assessment in OCTA via denoising autoencoders and segmentation uncertainty estimation. In *Medical Imaging with Deep Learning, Short Papers*, 2025.

[3] Cosmin I Bercea, Michael Neumayr, Daniel Rueckert, and Julia A Schnabel. Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML) Workshop*, 2023.

[4] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[5] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[6] Biomedical Image Analysis Group, Imperial College London. IXI dataset – information eXtraction from Images. `https://brain-development.org/ixi-dataset/`, 2006. EPSRC project GR/S21533/02. License: CC BY-SA 3.0.

[7] Lyndon Boone, Mahdi Biparva, Parisa Mojiri Forooshani, Joel Ramirez, Mario Masellis, Robert Bartha, Sean Symons, Stephen Strother, Sandra E Black, Chris Heyn, et al. ROOD-MRI: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI. *NeuroImage*, 278:120289, 2023.

[8] Matias Cosarinsky, Ramiro Billot, Lucas Mansilla, Gabriel Jimenez, Nicolás Gaggion, Guanghui Fu, Tom Tirer, and Enzo Ferrante. ConfIC-RCA: Statistically grounded efficient estimation of segmentation quality. *IEEE Transactions on Medical Imaging*, 2026. Early Access.

[9] Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[10] Terrance DeVries, Adriana Romero, Joelle Pineau, Graham W. Taylor, and Ross Goroshin. On the evaluation of conditional GANs. *arXiv preprint arXiv:1907.08175*, 2019.

[11] D.C. Dowson and B.V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[12] Daniele Falcetta, Vincenzo Marcianò, Kaiyuan Yang, Jon Cleary, Loïc Legris, Massimiliano Domenico Rizzaro, Ioannis Pitsiorlas, Hava Chaptoukaev, Benjamin Lemasson, Bjoern Menze, and Maria A. Zuluaga. VesselVerse: A Dataset and Collaborative Framework for Vessel Annotation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 655–665, September 2025.

[13] Francesco Galati, Daniele Falcetta, Rosa Cortese, Barbara Casolla, Ferran Prados, Ninon Burgos, and Maria A Zuluaga. A2V: A semi-supervised domain adaptation framework for brain vessel segmentation via two-phase training angiography-to-venography translation. *arXiv preprint arXiv:2309.06075*, 2023.

[14] Francesco Galati and Maria A. Zuluaga. Efficient model monitoring for quality control in cardiac image segmentation. In *Proceedings of Functional Imaging and Modeling of the Heart*, 2021.

[15] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. MAISI: Medical AI for synthetic imaging. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441, 2025.

[16] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brain Lesion: Toward Natural Science for Brain Tumor Segmentation (BrainLes 2021)*, volume 12962 of *Lecture Notes in Computer Science*, pages 272–284, 2021.

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

[18] Fabian Isensee, Paul Friedrich Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: Self-adapting framework for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[19] Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al. nnInteractive: Redefining 3D promptable segmentation. *arXiv preprint arXiv:2503.08373*, 2025.

[20] Hana Jebril, Thomas Pinetz, and Hrvoje Bogunović. Shape prior for quality assessment in OCTA via denoising autoencoders at the segmentation level. *IEEE Access*, 2025.

[21] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

[22] Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. UniMed-CLIP: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024.

[23] Agostina J Larrazabal, César Martínez, Ben Glocker, and Enzo Ferrante. Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders. *IEEE Transactions on Medical Imaging*, 39(12):3813–3820, 2020.

[24] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[25] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[26] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[27] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. MedSAM2: Segment anything in 3D medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

[28] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. AbdomenCT-1K: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022.

[29] Lena Maier-Hein et al. Metrics Reloaded: Recommendations for image analysis validation. *Nature Methods*, 21:195–212, 2024.

[30] Vincenzo Marcianò, Hava Chaptoukaev, Virginia Fernandez, M Jorge Cardoso, Sébastien Ourselin, Michela Antonelli, and Maria A Zuluaga. Diffusion-based quality control of medical image segmentations across organs. *arXiv preprint arXiv:2511.09588*, 2025.

[31] Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop*, 2018.

[32] Nathan Painchaud, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalande, and Pierre-Marc Jodoin. Cardiac segmentation with strong anatomical guarantees. *IEEE Transactions on Medical Imaging*, 39(11):3703–3713, 2020.

[33] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

[34] Walter HL 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. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshops*, pages 117–126. Springer, 2022.

[35] Chethan Radhakrishna, Karthikesh Varma Chintalapati, Sri Chandana Hudukula Ram Kumar, Raviteja Sutrave, Hendrik Mattern, Oliver Speck, Andreas Nürnberger, and Soumick Chatterjee. SPOCKMIP: Segmentation of vessels in MRAs with enhanced continuity using maximum intensity projection as loss. *arXiv preprint arXiv:2407.08655*, 2024.

[36] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[37] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. clDice – a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16560–16569, 2021.

[38] Gilbert W. Stewart. Perturbation theory for the singular value decomposition. In Richard J. Vaccaro, editor, *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, pages 99–109. Elsevier, 1991.

[39] Nico Stucki, Johannes C Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning*, pages 32698–32727. PMLR, 2023.

[40] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Qiao Yu. SAM-Med3D. *arXiv preprint arXiv:2310.15161*, 2023.

[41] Shuo Wang, Giacomo Tarroni, Chen Qin, Yuanhan Mo, Chengliang Dai, Chen Chen, Ben Glocker, Yike Guo, Daniel Rueckert, and Wenjia Bai. Deep generative model-based quality control for cardiac MRI segmentation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020.

[42] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[43] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. ScribblePrompt: fast and flexible interactive segmentation for any biomedical image. In *European Conference on Computer Vision (ECCV)*, volume 15098, pages 3–19, 2024.

[44] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023.

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[46] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

A Dataset and Implementation Details

This appendix summarises the datasets and implementation details used throughout the experiments. Table 4 reports the modality, number of foreground structures, and number of pre-extracted per-class image-mask feature representations for each dataset.

Feature representations are extracted once with a frozen MedSigLIP encoder and reused for MID and all distributional baselines. For each image-mask pair, the image embedding is computed once and shared across foreground classes, whereas each class-specific binary mask is embedded separately. Thus, N_{repr} denotes the number of per-class image-mask representations, not the number of subjects. Unless otherwise stated, PCA is fitted on the reference split with $k = 128$ components, and covariance matrices are estimated using Ledoit–Wolf shrinkage. The 56 foreground classes used in the main corruption sweep (Section 4.2) are the subset for which all five corruption types produced valid non-empty masks across the evaluation split; the explicit list of these classes is provided in the released metadata accompanying the code.. By dataset, the 56 groups are 38 TotalSegmentator CT organs, 6 TotalSegmentator MRI organs, 3 ACDC structures, 2 Task03 Liver classes, 2 Task05 Prostate zones, and one class each from Task06 Lung, Task07 Pancreas, Task09 Spleen, Task10 Colon, and IXI MRA.

Table 4: Datasets used in the experiments. Multimodal datasets are listed once per modality, giving 13 entries in total. Datasets marked $^\bullet$ are held out from the reference pool and used only as candidate sets; the remaining 10 datasets form the reference pool described in Section 4.1. *For TotalSegmentator, Structures and N_{repr} count the six abdominal organs of Section 4.3. Features also exist for 44 TotalSegmentator CT organs. Of these, 38 give valid corrupted masks and enter the corruption sweep. ‡Peripheral zone (PZ) and transition zone (TZ).

Dataset	Modality	Structures	N_{repr}
AbdomenCT-1K$^\bullet$ [28]	CT	4	2888
ACDC [4]	MRI	3	900
AMOS CT$^\bullet$ [21]	CT	15	3000
AMOS MRI$^\bullet$	MRI	13	520
IXI MRA (VesselVerse [6, 12])	MRA	1	49
Task03 Liver (Decathlon [1])	CT	2	262
Task05 Prostate (Decathlon)	MRI	$2^\ddagger$	64
Task06 Lung (Decathlon)	CT	1	63
Task07 Pancreas (Decathlon)	CT	2	562
Task09 Spleen (Decathlon)	CT	1	41
Task10 Colon (Decathlon)	CT	1	126
TotalSegmentator v2 CT [42]	CT	6^*	7368
TotalSegmentator v2 MRI	MRI	6^*	1788
Total			17631

B Preprocessing and Practical Implementation

MID requires no training. Applying it to new data reduces to the fixed preprocessing sequence in Table 5. The image and its label map share one voxel grid. Volumes are reoriented to a canonical axis order before slicing, which proceeds along the third axis. Predictions from a model that resampled internally are mapped back onto the original image grid. Each volume is reduced to 16 uniform axial slices, with identical indices for image and mask. Each foreground class is binarised and rendered as a white foreground. Absent classes are dropped. The image is encoded once per case and reused across its classes. Image and mask vectors are concatenated in that order, since the decomposition of Section 3.4 reads the first half as the image block. The projection is fitted on the reference set alone, which makes the estimator reference-conditioned (Appendix O). Comparable MID values require identical settings in both distributions.

Table 5: Preprocessing and estimator settings used throughout this paper.

Setting	Value
Encoder	MedSigLIP (`google/medsiglip-448`), frozen
Embedding	pooled output, ℓ_2-normalised, $d = 1152$
Joint vector	$[\,\text{image} \parallel \text{mask}\,]$, $2d = 2304$
Axial slices per volume	$N = 16$, uniform; all slices if $D \leq N$
2-D inputs	identical path with $N = 1$
CT rendering	three HU windows $[-1024, 1024]$, $[-135, 215]$, $[0, 80]$ as RGB
MRI rendering	per-slice min-max to 8-bit, replicated to 3 channels
Mask rendering	one binary mask per foreground class, white on black
Projection	PCA, $k = 128$, fitted on the reference set only
Covariance	Ledoit–Wolf shrinkage
Reference size	≥ 250 representations per group
Candidate size	≥ 250 standalone; equal across compared candidates

C Controlled Corruption Protocol

This appendix describes the controlled mask-corruption protocol used in Experiment 1. The goal is to create candidate segmentation distributions with known degradation type and severity while keeping the image distribution fixed. This allows us to test whether MID responds monotonically to segmentation quality degradation without using per-sample pairing during metric computation.

We apply five signed-distance-transform (SDT)-based corruption types: elastic boundary deformation, smooth erosion, smooth dilation, blob-shaped false positives, and partial missing slices. Each corruption is applied at three severity levels: mild, moderate, and severe. These corruptions are designed to mimic common segmentation model failures, including boundary bias, over-segmentation, false-positive regions, and incomplete volumetric predictions.

Figure 5 illustrates the five corruption types on representative organs. Green denotes preserved true-positive regions, red denotes false positives, and blue denotes false negatives.

D Additional Results for Controlled Corruptions

This appendix provides additional analyses for Experiment 1. The main text reports the overall correlation between MID and paired-reference metrics under controlled mask corruptions. Here we further examine clean-reference variability, anatomy-specific calibration ranges, and the limitation of comparing absolute MID values across anatomies.

D.1 Clean-Reference Variability

We first estimate the natural variability of MID when both reference and candidate samples are clean. For each dataset-organ group, we split the clean ground-truth image-mask pairs into two disjoint subsets and compute MID between them. This clean-vs-clean setting provides a null baseline for interpreting non-zero MID values caused by finite-sample estimation and anatomical variability.

Figure 6 reports MID for clean-vs-clean comparisons within each dataset-organ group. Each comparison uses repeated random half-splits of the clean feature files of the group. Groups with fewer than 20 samples are excluded. Dice is 1.0 for all groups by construction. The observed spread reflects natural anatomical variability and finite-sample estimation noise.

D.2 Per-Anatomy Reference Index

Because MID is computed in a learned feature space, its absolute scale depends on the anatomy being evaluated. Large organs and compact structures have different baseline variability and different responses to the same corruption type. Figure 7 gives a per-anatomy reference index. For each structure, it shows the clean-vs-clean MID and its spread over random half-splits of the clean pairs.

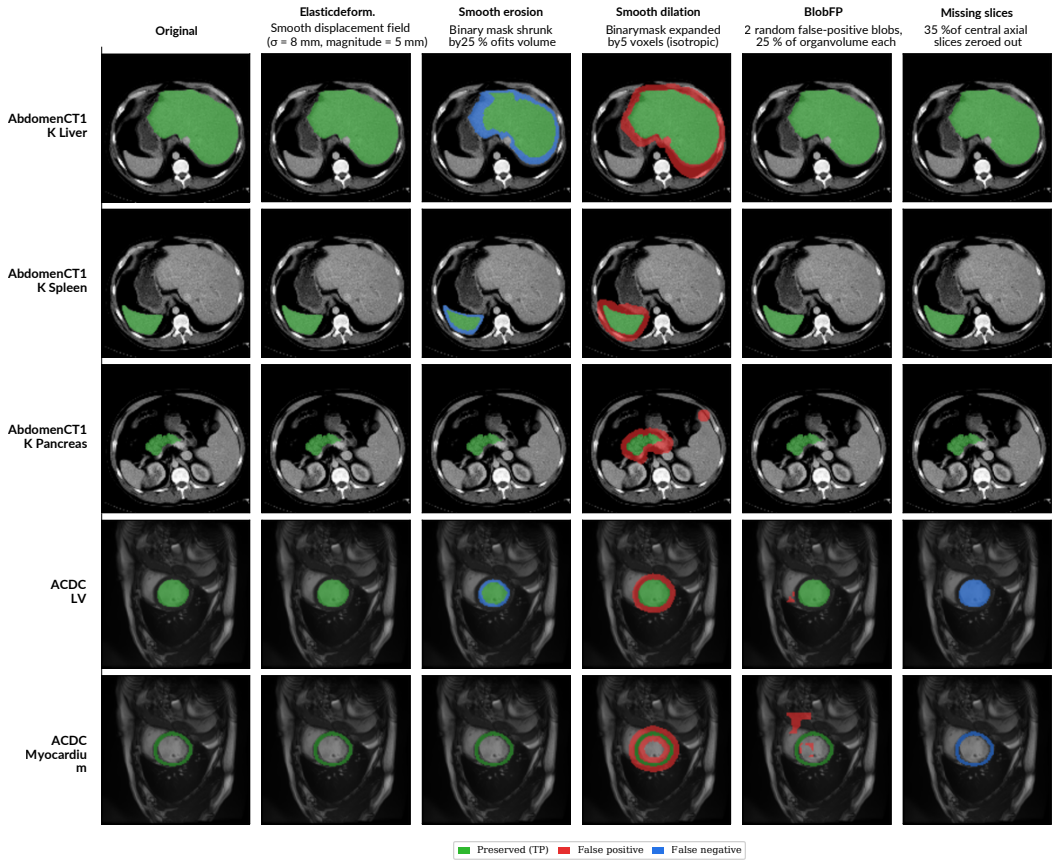

Figure 5: Anatomically realistic SDT-based mask corruptions applied to five organs. Columns show the original mask and five corruption types: elastic deformation, smooth erosion, smooth dilation, blob-shaped false positives, and partial missing slices. Overlay colours: green = preserved region, red = false positive, blue = false negative.

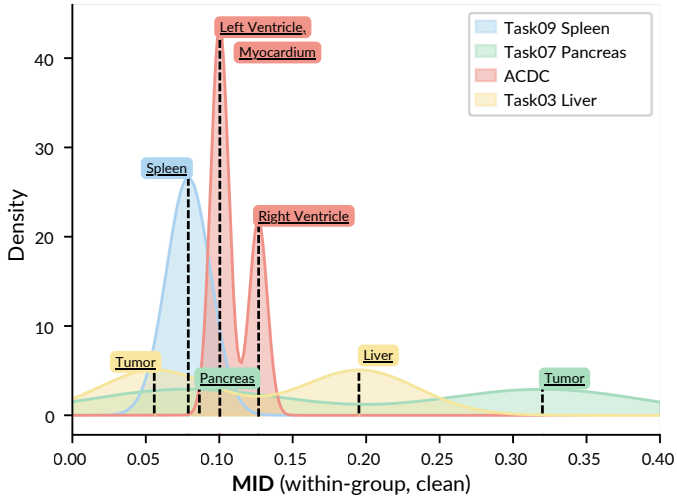

Figure 6: Within-group MID distribution across datasets. Each curve shows the kernel density estimate of clean-vs-clean MID values over organ groups within a dataset.

Overall, the per-anatomy reference index shows that absolute MID values can vary by roughly an order of magnitude across anatomies. This supports the recommendation in Section 5: MID should be interpreted within an anatomy or relative to an anatomy-specific reference range.

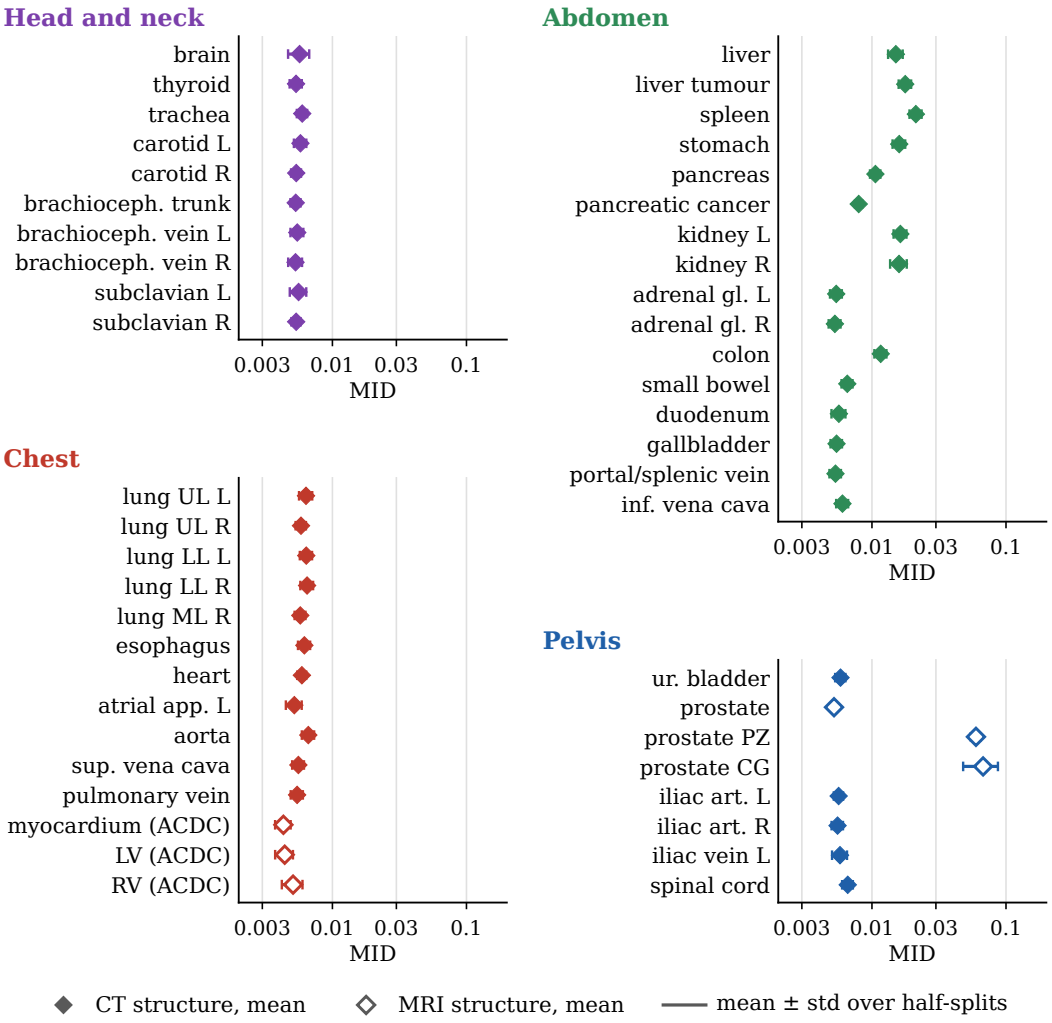

Figure 7: Per-anatomy MID reference index for 48 structures in four anatomical regions. The structures come from TotalSegmentator and the overlapping MSD, AbdomenCT-1K and ACDC annotations. For each structure, the diamond marks the mean clean-vs-clean MID over repeated random half-splits of its clean pairs (Appendix D.1). The bar spans the mean $\pm$ one standard deviation. Solid bars mark MRI structures. Dashed bars mark CT structures. The MID axis uses a log scale.

E Distributional Shift and ROOD-MRI Details

This appendix provides additional details for the distributional-shift experiments in Section 4.3. We consider two complementary settings: cross-dataset shifts, where anatomy or modality differs from the reference population, and within-dataset image-quality shifts induced by MRI artefacts. These analyses test whether MID responds to changes in the joint image-mask distribution beyond synthetic mask corruptions.

E.1 Cross-Dataset Shift Setup

For the cross-dataset experiment, the reference distribution is built from TotalSegmentator CT training cases over six abdominal organs shared across candidate datasets. Candidate distributions are constructed from held-out TotalSegmentator CT, TotalSegmentator MRI, Task05 Prostate MRI, and ACDC cardiac MRI. This design creates an ordered sequence of shifts: same dataset and modality, modality shift, and combined anatomy and modality shift.

The in-distribution held-out TotalSegmentator CT split provides the finite-sample floor for MID. Candidate datasets with different modality or anatomy are expected to produce progressively larger scores. This setup is used in Table 2 in the main text.

E.2 ROOD-MRI Prostate Robustness

This subsection gives the full setup for the within-dataset artefact experiment. The reference distribution is computed once from clean Task05_Prostate T2 images paired with ground-truth masks. For each artefact type and severity level, the candidate distribution is built from corrupted T2 images paired with SegResNet predictions on those corrupted images. Ground-truth masks are used only to compute Dice and HD95 for analysis, not to compute MID at inference time.

Setup. Dataset: Task05_Prostate [1], consisting of T2-weighted prostate MRI with peripheral-zone and transition-zone labels. Model: SegResNet trained for 300 epochs on Task05_Prostate [31]. Artefacts: Ghosting, Rician noise, and Random Motion at three severity levels: mild, moderate, and severe. Each artefact is applied to the T2 channel only; the ADC channel is kept clean.

MID **at inference time.** The reference distribution is computed offline from clean T2 images paired with ground-truth masks. For each pair of artefact and severity, the candidate distribution is formed using corrupted T2 images paired with SegResNet predictions. Thus, MID does not use a ground-truth mask for each candidate prediction. Dice and HD95, by contrast, require paired ground-truth masks and are used only as reference-quality indicators.

Interpretation. As reported in Section 4.3, MID increases with artefact severity and follows the degradation trends measured by Dice and HD95. This supports its use for monitoring distribution-level segmentation quality when per-case annotations are unavailable.

F Metric Comparison: Full Correlation Results

This appendix reports the full Spearman correlations with $95\,\%$ bootstrap confidence intervals for the metric comparison in Section 4.4. Each distributional metric is correlated against *both* per-sample paired metrics, Dice and HD95, computed cell-by-cell on the same 840 cells. The comparison includes joint MID and five feature-space distributional baselines: FID on Inception-v3 mid-slice features, FID on MedSigLIP image features, FID on MedSigLIP mask features, KID on MedSigLIP mask features, and Mahalanobis distance from the joint reference Gaussian. Dice and HD95 themselves are not listed as rows: they are the reference targets that the columns are computed against. The decomposition of MID into $\mathrm{MID}_{\mathrm{img}}$, $\mathrm{MID}_{\mathrm{mask}}$, and $\mathrm{MID}_{\mathrm{corr}}$ is examined separately in Experiment 5 and Appendix G and is not included here.

Table 6: Metric comparison on the controlled corruption sweep. Spearman ρ is computed cell-by-cell against mean Dice and against HD95 across 840 cells (56 organ groups $\times$ 5 corruption types $\times$ 3 severity levels); bootstrap CIs use 1000 replicates. The two reference targets are inversely related (Dice goes up, HD95 goes down for better segmentations), so a metric that tracks segmentation quality should produce a negative ρ against Dice and a positive ρ against HD95 of comparable magnitude.

Category	Metric	Encoding	vs. Dice		vs. HD95	
			ρ	$95\,\%$ CI	ρ	$95\,\%$ CI
Ours	MID	MedSigLIP	-0.55	$[-0.61, -0.49]$	$+0.72$	$[+0.66, +0.77]$
Baselines	FID-mask, 3-D	MedSigLIP	-0.47	$[-0.53, -0.40]$	$+0.61$	$[+0.54, +0.67]$
	FID-img, 3-D	MedSigLIP	-0.03	$[-0.10, +0.04]$	$+0.04$	$[-0.04, +0.12]$
	KID	MedSigLIP	-0.28	$[-0.36, -0.20]$	$+0.36$	$[+0.28, +0.44]$
	FID-Inception, 2-D	Inception-v3	-0.38	$[-0.56, -0.18]$	$+0.49$	$[+0.31, +0.65]$
	Mahalanobis (joint)	MedSigLIP	-0.42	$[-0.49, -0.34]$	$+0.55$	$[+0.47, +0.62]$

Two patterns are visible. First, MID is the strongest correlator with both Dice ($\rho = -0.55$) and HD95 ($\rho = +0.72$) among the unpaired distributional metrics; the gap to FID-mask is consistent

across the two reference targets ($+0.08$ absolute against Dice, $+0.11$ absolute against HD95), and the bootstrap CIs of MID and FID-mask overlap in both columns. Second, the HD95 column has uniformly larger $|\rho|$ than the Dice column for every metric. This is consistent with HD95 being more sensitive to the boundary-shape errors that dominate the corruption sweep (smooth erosion, dilation, elastic deformation) while Dice penalises overlap loss more uniformly. FID-img is essentially zero in both columns, as expected when images are held fixed across the sweep.

G Image-Mask Correspondence Decomposition

This appendix gives the formal block-covariance derivation that underlies the diagnostic decomposition used in Section 3.4, and then reports the empirical setup that follows from it.

G.1 Covariance Block Structure

For an image-mask pair (X, m_c), write the projected feature vector as $z = (z_x; z_m)$, with image block $z_x = \pi_x(f_\theta(X))$ and mask block $z_m = \pi_m(f_\theta(m_c))$. Here π_x and π_m are PCA projections fitted on the reference image features and on the reference mask features. These block-wise projections keep the image and mask coordinates apart. For a distribution P over such embeddings, the mean and covariance admit the block form

$$\mu_P = \begin{pmatrix} \mu_P^x \\ \mu_P^m \end{pmatrix}, \qquad \Sigma_P = \begin{pmatrix} \Sigma_P^{xx} & \Sigma_P^{xm} \\ \Sigma_P^{mx} & \Sigma_P^{mm} \end{pmatrix}, \tag{4}$$

with $\Sigma_P^{xm} = (\Sigma_P^{mx})^\top$, and analogously for Q. The squared 2-Wasserstein distance between Gaussian fits $\mathcal{N}(\mu_P, \Sigma_P)$ and $\mathcal{N}(\mu_Q, \Sigma_Q)$ is

$$d_F^2(P, Q) = \|\mu_P - \mu_Q\|^2 + \mathrm{Tr}\Big(\Sigma_P + \Sigma_Q - 2\big(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2}\big)^{1/2}\Big). \tag{5}$$

Two of the three terms in (5) are linear in the block partition. The mean term decomposes as

$$\|\mu_P - \mu_Q\|^2 = \|\mu_P^x - \mu_Q^x\|^2 + \|\mu_P^m - \mu_Q^m\|^2, \tag{6}$$

and the additive trace term decomposes as

$$\mathrm{Tr}(\Sigma_P + \Sigma_Q) = \mathrm{Tr}(\Sigma_P^{xx} + \Sigma_Q^{xx}) + \mathrm{Tr}(\Sigma_P^{mm} + \Sigma_Q^{mm}). \tag{7}$$

The geometric-mean term $\mathrm{Tr}\big(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2}\big)^{1/2}$ does *not* decompose linearly in general, because the principal matrix square root mixes diagonal and off-diagonal blocks whenever either Σ_P^{xm} or Σ_Q^{xm} is non-zero.

G.2 Image, Mask, and Residual Correspondence Terms

Following the FJD-style construction [10], we define the marginal Fréchet distances on the image and mask blocks alone:

$$\mathrm{MID}_{\mathrm{img}}(P, Q) = \|\mu_P^x - \mu_Q^x\|^2 + \mathrm{Tr}\Big(\Sigma_P^{xx} + \Sigma_Q^{xx} - 2\big(\Sigma_P^{xx,1/2}\Sigma_Q^{xx}\Sigma_P^{xx,1/2}\big)^{1/2}\Big), \tag{8}$$

$$\mathrm{MID}_{\mathrm{mask}}(P, Q) = \|\mu_P^m - \mu_Q^m\|^2 + \mathrm{Tr}\Big(\Sigma_P^{mm} + \Sigma_Q^{mm} - 2\big(\Sigma_P^{mm,1/2}\Sigma_Q^{mm}\Sigma_P^{mm,1/2}\big)^{1/2}\Big). \tag{9}$$

The residual correspondence term is defined by completion to the full joint score:

$$\mathrm{MID}_{\mathrm{corr}}(P, Q) = \mathrm{MID}(P, Q) - \mathrm{MID}_{\mathrm{img}}(P, Q) - \mathrm{MID}_{\mathrm{mask}}(P, Q). \tag{10}$$

Under block-wise projections, this residual comes only from the geometric-mean term in (5). It is zero when the off-diagonal blocks Σ_P^{xm} and Σ_Q^{xm} are zero. These blocks quantify image-mask coupling under each distribution. Section 3.4 computes MID with the joint projection of Definition 1. The residual then also absorbs the difference between the joint and the block-wise projections.

Block-diagonal case. If $\Sigma_P^{xm} = 0$ and $\Sigma_Q^{xm} = 0$, then both Σ_P and Σ_Q are block-diagonal, $\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2}$ is block-diagonal with blocks $\Sigma_P^{xx,1/2}\Sigma_Q^{xx}\Sigma_P^{xx,1/2}$ and $\Sigma_P^{mm,1/2}\Sigma_Q^{mm}\Sigma_P^{mm,1/2}$, the principal square root acts block-wise on positive semi-definite block-diagonal matrices, and the trace decomposes per block. Substituting into (5) and using (6) and (7) gives

$$\Sigma_P^{xm} = \Sigma_Q^{xm} = 0 \implies \mathrm{MID}(P, Q) = \mathrm{MID}_{\mathrm{img}}(P, Q) + \mathrm{MID}_{\mathrm{mask}}(P, Q), \tag{11}$$

so $\mathrm{MID}_{\mathrm{corr}}(P, Q) = 0$. Conversely, when at least one of the off-diagonal blocks is non-zero and they differ between P and Q, the geometric-mean term picks up cross-block contributions and $\mathrm{MID}_{\mathrm{corr}}$ is generally non-zero. We therefore attribute $\mathrm{MID}_{\mathrm{corr}}$ to image-mask correspondence.

G.3 Experimental Scenarios

This subsection provides additional details for the decomposition experiment in Section 4.6. The purpose is to test whether the residual correspondence term $\mathrm{MID}_{\mathrm{corr}}$ responds to changes in image-mask dependence that are not visible from image-only or mask-only marginal distributions.

Scenario A corrupts masks while preserving image-mask pairing. Scenario B randomly shuffles masks across images, preserving the image and mask marginal distributions while disrupting their dependence. Scenario C compares TotalSegmentator CT to AMOS MRI to illustrate how the decomposition behaves under a dominant modality shift. The decomposition is visualised in Figure 4 in the main text.

Pair shuffling (Scenario B) is the cleanest empirical test of the block-structure derivation above: the candidate mask multiset is identical to the un-shuffled case, so Σ_Q^{mm} and μ_Q^m are exactly preserved; image features are also unchanged, so Σ_Q^{xx} and μ_Q^x are exactly preserved; only the off-diagonal block Σ_Q^{xm} is destroyed (driven to approximately zero). Consequently, $\mathrm{MID}_{\mathrm{img}}$ and $\mathrm{MID}_{\mathrm{mask}}$ should be invariant under shuffling, while $\mathrm{MID}_{\mathrm{corr}}$ should absorb the entire effect. This is the prediction that Section 4.6 tests.

Overall, the experiment supports interpreting $\mathrm{MID}_{\mathrm{corr}}$ as a diagnostic residual for correspondence changes rather than as a standalone metric.

H Real-Model Ranking Details

This appendix provides additional details for the real segmentation model experiment in Section 4.5. The goal is to test whether MID ranks real model outputs consistently with paired-reference metrics, rather than only responding to synthetic corruptions.

For each group of model, dataset and organ, the candidate distribution is built from model predictions on disjoint evaluation images, with each prediction paired with the corresponding image latent to form an image-mask embedding. MID is then computed against the global reference pool defined in Section 4.1, without sample-wise matching between reference and candidate. The global pool puts candidates from different models, datasets, and anatomies on a common scale. The pool is mostly CT. The ACDC offset in Figure 3a follows from this choice. It mixes cardiac segmentation quality with the small share of cardiac MRI in the reference.

For the IXI MRA vessel segmentation task, MID ranks the released methods in the same order as Dice and HD95: nnUNet obtains the lowest MID and highest Dice, followed by A2V and SPOCKMIP. All three methods are scored on the same images against the same reference. Any modality offset is thus common to the three methods. The comparison isolates the differences between the methods.

I Dataset-Specific Projections

This appendix analyses the PCA projection, the remaining estimator choice not covered in Section 4.7. Since MID relies on Gaussian statistics in a projected feature space, the choice of projection subspace can affect the stability of the estimate. A practical question concerns whether the projection should be fitted on a large universal corpus or on the target dataset alone. The default MID projection is fitted on the reference pool of Section 4.1. It gives one universal coordinate system for all datasets. A practitioner with a single target dataset can fit the projection on that dataset instead. This choice gives up cross-dataset comparability.

We test this per-organ, using TotalSegmentator CT as the healthy reference for two organs that also appear with pathological annotations in MSD: liver (class 1 in Task03, where class 2 is liver tumour) and pancreas (class 1 in Task07, where class 2 is pancreas tumour). For each organ, we fit a $k = 128$ PCA projection on the TotalSegmentator CT training split and compare against the universal projection.

Per-organ results. Table 7 reports MID under each projection for three candidate sets: a held-out split from the same TotalSegmentator distribution, the MSD healthy organ class from a different scanner and protocol, and the corresponding tumour class.

Table 7: MID between TotalSegmentator CT reference and three candidate sets, under universal and organ-specific projections ($k = 128$). Reference: TotalSegmentator CT training split; organ-specific PCA is fitted on the same reference.

Organ	Candidate	Universal	Organ-PCA
Liver	TotalSeg CT held-out (in-distribution)	0.050	0.049
	Task03 liver, class 1 (healthy, cross-dataset)	0.154	0.152
	Task03 liver tumour, class 2 (pathological)	0.165	0.162
Pancreas	TotalSeg CT held-out (in-distribution)	0.038	0.037
	Task07 pancreas, class 1 (healthy, cross-dataset)	0.238	0.236
	Task07 pancreas tumour, class 2 (pathological)	0.216	0.214

The two projections produce near-identical values: organ-specific PCA reduces MID by less than $3\,\%$ relative to the universal projection. In-distribution held-out splits yield the smallest scores in both cases. For liver, the tumour class scores above the healthy cross-dataset class, reflecting the additional distributional shift introduced by pathological appearance; for pancreas, the ordering is reversed. Pancreatic tumour masks in MSD are smaller and more variable than the full-organ masks of the reference. This scale difference lowers the measured divergence. The practical implication is that the projection choice has negligible effect on relative rankings for single-organ evaluations.

Practical recommendation. For single-dataset evaluation, a PCA fitted on a held-out reference split of the target data is appropriate. Table 7 shows nearly the same values for both projections. For cross-dataset comparisons, the universal projection is preferable because it keeps all datasets in a common coordinate system.

J Reference-Set Size and Label Quality

The experiments in the main text keep the reference distribution fixed and vary the candidate set. This appendix reports the complementary analysis: how MID behaves as the reference set itself shrinks or becomes contaminated with incorrect labels. Both are practical concerns, since the reference pool is what defines the normative population against which candidates are scored.

Reference-set size. We fix the candidate set and sweep the reference size n_P on liver features, with 30 bootstrap resamples per level. Table 8 reports the mean MID, the bootstrap confidence-interval half-width relative to the mean, and the bias relative to the largest reference size tested. The finite-sample bias falls below $25\,\%$ at $n_P = 250$ and below $10\,\%$ at $n_P = 500$, with relative CI half-widths of $11\,\%$ and $7\,\%$.

Table 8: Sensitivity of MID to the reference-set size n_P on liver features, with the candidate set held fixed. Mean, relative CI half-width, and bias are computed over 30 bootstrap resamples per level; bias is measured against $n_P = 1000$.

Statistic	Reference set size n_P				
	50	100	250	500	1000
Mean MID	0.025	0.023	0.020	0.017	0.016
CI half-width / mean	$32\,\%$	$18\,\%$	$11\,\%$	$7\,\%$	$3\,\%$
Bias vs. $n_P = 1000$	$+59\,\%$	$+45\,\%$	$+23\,\%$	$+9\,\%$	reference

Reference label quality. We next simulate wrong-organ labelling by replacing a fraction ϵ of the liver reference embeddings with spleen embeddings, keeping the reference size fixed at $n_P = 500$ and using 20 bootstrap resamples per level. Table 9 reports the result. At the levels tested, MID is tolerant of isolated wrong-organ labels: $5\,\%$ contamination inflates the score by $2.9\,\%$, which lies within the bootstrap CI at that reference size. The response is monotone in ϵ and reaches $42.8\,\%$ inflation at $\epsilon = 0.50$. We therefore read MID as robust to occasional gross labelling errors in the

reference pool, while noting that systematic degradation does shift the score, so users who suspect systematic labelling bias should audit the provenance of the reference pool.

Table 9: Effect of reference-pool label contamination. A fraction ϵ of liver reference embeddings is replaced by spleen embeddings, with $n_P = 500$ fixed and 20 bootstrap resamples per level. Inflation is measured against the uncontaminated reference.

ϵ	# contaminated	Mean MID	95 % CI	Inflation
0.00	0	0.0173	[0.0160, 0.0193]	reference
0.05	25	0.0178	[0.0165, 0.0194]	+2.9 %
0.10	50	0.0183	[0.0166, 0.0202]	+5.8 %
0.20	100	0.0196	[0.0176, 0.0219]	+13.3 %
0.50	250	0.0247	[0.0228, 0.0272]	+42.8 %

K Gaussian-Mixture Reference

MID fits a single Gaussian to the reference distribution, while Appendix M shows that a pooled CT and MRI reference is not unimodal: MRI samples are displaced from CT samples along the dominant principal component. This appendix tests whether that bimodality changes the ordering that MID induces over candidate distributions, which is the property the main experiments rely on.

Protocol. We compare the default single-Gaussian estimator with a two-component Gaussian-mixture estimator on joint embeddings. The mixture estimator scores candidates with the mixture-Wasserstein upper bound of Delon and Desolneux [9]. We use the synthetic setup of Appendix O: a two-mode reference that mirrors the CT and MRI split ($n_P = 500$), and candidates with additive corruption ($n_Q = 300$). The study uses 7 severity levels and 5 seeds.

Result. The rankings of the two estimators agree closely across seeds. Three seeds give identical orderings ($\rho = 1.00$). Two seeds show minor permutations. The two estimators differ mainly in absolute scale, because the single Gaussian absorbs the separation between the CT and MRI modes into one global covariance whereas the mixture represents the modes separately. This difference is largely monotone in severity, which explains the near-identical rankings. We also tested a nearest-mode assignment variant, which proved less stable across seeds: candidates lying near the boundary between modes can switch assignment under small perturbations, producing discontinuous score changes because the mode-specific covariances differ. We therefore recommend the single-Gaussian estimator for comparative ranking under a fixed reference protocol, and a mixture estimator when absolute calibration across heterogeneous modes is required.

L Within-Organ Comparison against Mask-Only Baselines

The metric comparison in Section 4.4 aggregates cells across organs of very different size, and the bootstrap confidence intervals of MID and FID-mask overlap in that pooled analysis (Table 6). Because the absolute scale of MID varies by roughly an order of magnitude across anatomies (Appendix D.2), between-organ scale differences can mask a within-organ difference between the two metrics. This appendix isolates the per-cell quality signal by removing those scale effects.

Protocol. We restrict the sweep to a 38-organ TotalSegmentator CT subset. Each organ has 20 cells: five corruption types at four severity levels. The fourth level is an extreme level above the three levels of Section 4.2. The subset thus has 760 cells. We rank cells within each organ before we aggregate across organs. Each organ thus contributes only the ordering of its own cells. We then compare MID against each mask-only baseline with a paired-cell bootstrap over 5000 replicates, pairing on the cell so that both metrics see exactly the same corruption instances.

Result. Under this within-organ protocol, MID improves on FID-mask against both reference targets, Dice and HD95, at $p < 0.001$. The same paired test favours MID at $p < 0.001$ against the other mask-only baselines, KID and $\mathrm{MID}_{\mathrm{mask}}$. It also favours MID against Mahalanobis distance on

joint features. With organ-scale effects removed, the joint encoding adds a quality signal to mask geometry.

M Latent Space Geometry

This appendix visualises the geometry of the MedSigLIP feature space used by MID. These analyses are not used to compute the metric; rather, they provide qualitative evidence that the encoder organises samples by modality and anatomy in a way that is meaningful for distributional comparison.

M.1 PCA Projection Structure

Figure 8 plots the first two principal components of joint MedSigLIP embeddings. CT datasets cluster in one region of the space, while MRI datasets are displaced along the dominant principal component. This supports the interpretation that modality is a major source of variation in the joint feature space.

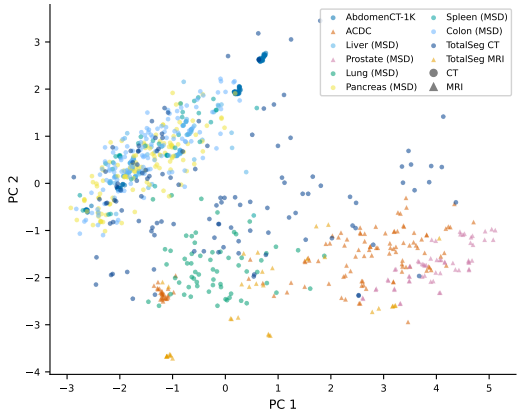

Figure 8: First two PCA components of joint MedSigLIP embeddings. CT datasets occupy a common region, whereas MRI datasets are displaced along the dominant modality-related axis.

MRI samples lie away from CT samples in the joint feature space. A modality difference between reference and candidate adds this displacement to MID.

M.2 MedSigLIP Feature Clustering

Figure 9 shows t-SNE projections of raw MedSigLIP features before PCA. We visualise both mask-only features and joint image-mask features for representative CT and MRI organs.

The clustering supports the use of MedSigLIP as the MID backbone: similar anatomical structures occupy nearby regions of the feature space, while modality shifts are reflected in the joint embedding.

N Encoder Backbone Ablation

This appendix studies the dependence of MID on the choice of visual encoder. Since MID is an encoder-space distributional metric, its behaviour depends on whether the embedding captures medically meaningful image and mask structure. We compare MedSigLIP against biomedical, natural-image, and generative encoders using the same controlled corruption sweep.

Figure 10 summarises the per-organ rank correlation between MID and Dice across five encoder backbones. MedSigLIP achieves the highest mean $|\rho| = 0.833$ $[0.742, 0.918]$; BiomedCLIP and DINOv2-L follow at 0.796 ($[0.758, 0.833]$ and $[0.750, 0.836]$, respectively), while CLIP-L/14 reaches 0.755 $[0.698, 0.807]$ and MAISI obtains 0.739 $[0.699, 0.778]$.

Two patterns emerge. First, joint image-mask encoding outperforms mask-only FID for every encoder, confirming that the image component contributes information beyond mask geometry alone. Second,

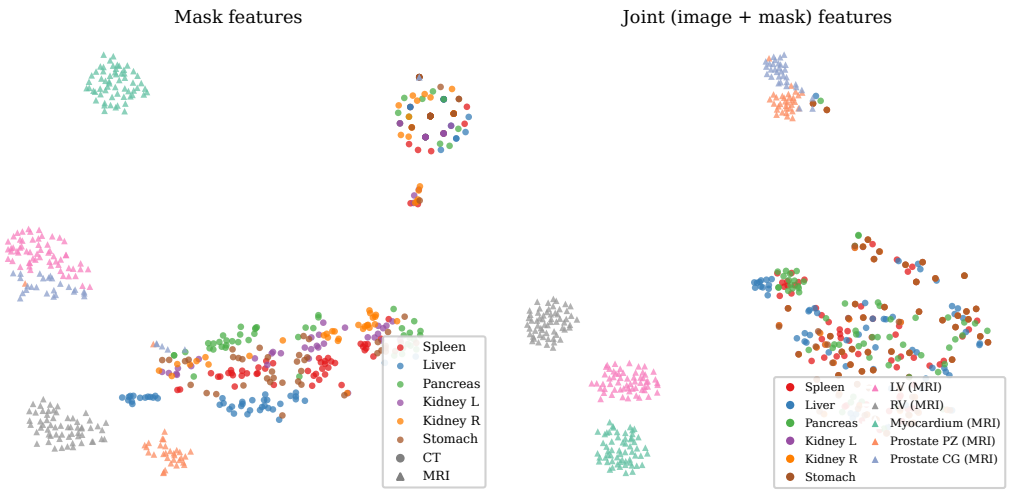

Figure 9: t-SNE visualisation of MedSigLIP embeddings for eleven organ types. Left: mask-only embeddings. Right: joint image-mask embeddings. Points are coloured by organ, with circles denoting CT samples and triangles denoting MRI samples. Anatomical structures form coherent clusters in both spaces, whereas MRI samples separate from their CT counterparts mainly in the joint embedding space, suggesting that modality-related variation is primarily captured by the image component.

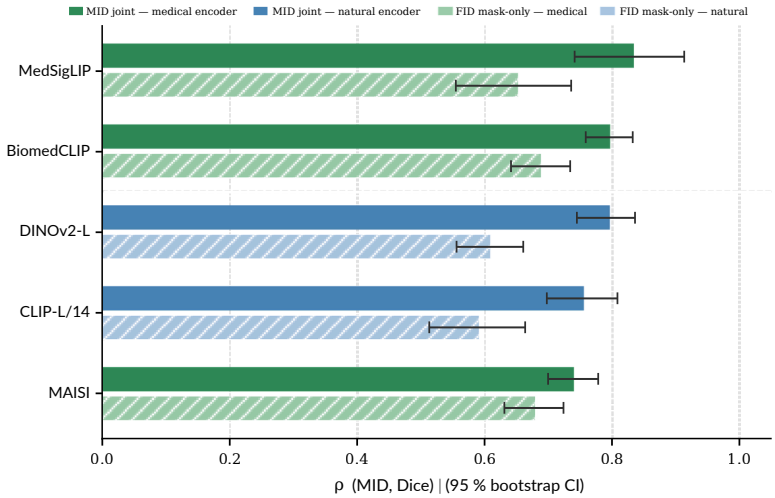

Figure 10: Mean per-organ $|\rho(\mathrm{MID}, \mathrm{Dice})|$ for five encoder backbones on the TotalSegmentator v2 CT corruption sweep. Each bar reports the mean over per-organ rank correlations.

domain-specific encoders do not uniformly outperform natural-image encoders: DINOv2-L matches BiomedCLIP despite having no medical pretraining, suggesting that representational capacity can partially compensate for domain gap in this corruption setting. The lower correlation for MAISI is consistent with its VAE-style generative objective, which regularises the latent space toward a spherical prior rather than organising it for discriminative anatomy or quality assessment.

O Estimator Symmetry and Union-PCA Variant

The squared 2-Wasserstein distance between two Gaussians is symmetric in P and Q. The default MID estimator, however, is not, because the projection π is fitted on the reference set only. Swapping

Table 10: Symmetry probe on synthetic joint embeddings with the MedSigLIP feature dimensionality. $\mathrm{MID}(P, Q; \pi_P)$ and $\mathrm{MID}(Q, P; \pi_Q)$ are the two directions of the default estimator; $\mathrm{MID}^{\mathrm{sym}}$ uses union-PCA. All values are means over five seeds; relative gap is $|\mathrm{MID}(P, Q) - \mathrm{MID}(Q, P)|/\frac{1}{2}(\mathrm{MID}(P, Q) + \mathrm{MID}(Q, P))$.

Corruption severity	$\mathrm{MID}(P, Q; \pi_P)$	$\mathrm{MID}(Q, P; \pi_Q)$	$\mathrm{MID}^{\mathrm{sym}}$	rel. gap (mean)	rel. gap (max)
0.05	46.71	43.18	45.31	7.8%	8.7%
0.10	46.73	43.34	45.40	7.5%	8.4%
0.20	46.80	44.02	45.83	6.1%	7.0%
0.40	47.07	47.68	48.31	1.3%	2.8%
0.60	47.52	55.70	55.43	15.9%	18.6%
0.80	48.17	68.90	68.79	35.5%	39.3%
1.00	49.05	87.53	89.38	56.5%	60.9%

P and Q swaps the projection basis and, in general, changes the numerical value: $\mathrm{MID}(P, Q; \pi_P) \neq \mathrm{MID}(Q, P; \pi_Q)$. This appendix quantifies the direction gap and defines a symmetric variant.

Union-PCA variant. Fit the projection on the union of reference and candidate embeddings, $\pi_{P \cup Q}$, and compute MID using this basis for both distributions. This variant is symmetric by construction: swapping P and Q leaves the fitted basis unchanged, and the Fréchet distance in the projected space is symmetric in its arguments. We denote this variant $\mathrm{MID}^{\mathrm{sym}}$.

Perturbation bound on the direction gap. Let Π_P and Π_Q denote the orthogonal projectors onto the k-dimensional PCA subspaces fitted on P and Q respectively. Let $\theta_{\max}$ be the largest principal angle between the two subspaces. The operator-norm gap is $\|\Pi_P - \Pi_Q\|_2 = \sin\theta_{\max}$. Wedin's theorem [38] bounds $\sin\theta_{\max}$ by the covariance perturbation divided by the eigengap at k. The direction gap of MID is then bounded by

$$\left|\mathrm{MID}(P, Q; \pi_P) - \mathrm{MID}(Q, P; \pi_Q)\right| \;\leq\; 2\,\sin\theta_{\max}\left(\mathrm{Tr}\,\Sigma_P + \mathrm{Tr}\,\Sigma_Q + \|\mu_P - \mu_Q\|^2\right), \qquad (12)$$

where $\sin\theta_{\max}$ is computable in closed form from the singular values of $C_P C_Q^\top$ (C_P, C_Q the two component matrices).

Empirical asymmetry. We evaluate the gap on synthetic joint embeddings with the MedSigLIP feature dimensionality ($2d = 2304$, projection $k = 128$, reference size $n_P = 500$ drawn as a two-mode mixture that mirrors the CT/MRI split, candidate size $n_Q = 300$ obtained by increasingly severe additive corruption; five seeds). The synthetic features are not ℓ_2-normalised. Their MID values are therefore larger than the real values in this paper. Table 10 reports the two directions of the default estimator, the symmetric variant, and the relative gap.

The direction gap is small in the mild-corruption regime relevant to segmentation-quality ranking (below 10% for severities up to 0.2), and grows only when the two distributions differ substantially, at which point either direction still identifies the same rank order. Across all five seeds, the Spearman rank correlation between the default estimator and $\mathrm{MID}^{\mathrm{sym}}$ over the seven-point severity sweep is $\rho = 1.000$, so the reference-fitted projection preserves the induced ordering of candidates even when absolute scale differs. Practitioners who require exact symmetry (e.g., pairwise comparisons without a designated reference) should use $\mathrm{MID}^{\mathrm{sym}}$; for the ranking use cases in the main paper, the default and symmetric variants give the same ordering. Their sensitivity differs: over the full severity range, the default estimator increases by $5\,\%$ and $\mathrm{MID}^{\mathrm{sym}}$ by a factor of about two (Table 10).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the three contributions (formulation of segmentation evaluation as joint distribution matching and the MID metric, the diagnostic image/mask/correspondence decomposition, and a 13-dataset evaluation covering corruption sensitivity, distribution shift, image-mask misalignment, and model ranking on IXI MRA) and the experimental scope (56 organ groups, 13 datasets); all are substantiated in the corresponding experimental sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Discussion (Section 5) covers the Gaussian approximation under heterogeneous reference pools, the anatomy-dependence of the absolute scale, dependence on the MedSigLIP encoder quality, and the computational cost of feature extraction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: Section 3 formalises the metric and uses well-established structural properties (non-negativity of the squared 2-Wasserstein distance between Gaussians, and the block decomposition of the joint covariance) without restating their proofs. The default estimator uses a reference-fitted projection and is asymmetric in P and Q; Appendix O states this explicitly, provides a union-PCA symmetric variant, and gives a Wedin-style bound $|\mathrm{MID}(P,Q) - \mathrm{MID}(Q,P)| \leq 2\sin(\theta_{\max})\,(\mathrm{Tr}\,\Sigma_P + \mathrm{Tr}\,\Sigma_Q + \|\mu_P - \mu_Q\|^2)$ on the direction gap. This bound is derived from Wedin's theorem in one step and is not given a complete proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and Appendix A specify encoder architecture, PCA dimension ($k = 128$), data splits, corruption types and severities, and the within-organ rank-correlation procedure; all datasets are publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The MID package (metric API, PCA projection, feature loading utilities, and a minimal runnable example) is public at `https://github.com/erc-caravel/pytorch-mid`. All 13 evaluation datasets are publicly available under their original licenses.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: MID has no trainable parameters beyond the reference-fitted PCA; encoder, projection dimension, corruption parameters, and evaluation splits are all reported in Section 4 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The metric comparison (Table 6) reports Spearman ρ with 95 % bootstrap CIs (1000 replicates) for all distributional baselines, and the per-corruption breakdown (Table 1) reports the same for each of the five corruption types. The encoder backbone ablation (Figure 10) reports mean per-organ $|\rho(\mathrm{MID}, \mathrm{Dice})|$ with 95 % bootstrap CIs across five encoders. Two-sided p-values for the main ρ figures are reported alongside the corresponding scores in Section 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 4.1 reports the GPU type (NVIDIA A100 80 GB); the Discussion (Section 5) reports the per-volume feature-extraction cost (about 11 s) and that PCA fitting plus MID computation runs in under one minute on a CPU node for the full 11,223-vector reference pool.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The paper uses only pre-existing anonymised public medical imaging datasets; no new patient data were collected, and no human subjects research was conducted.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Discussion (Section 5) covers the positive impact on segmentation evaluation in deployment settings without per-case annotations and notes the risk that a metric anchored on a fixed reference distribution can inherit demographic, scanner, or acquisition biases from that reference; we therefore recommend interpreting absolute MID values cautiously when the reference pool is heterogeneous.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released asset is a lightweight evaluation metric library; it does not include generative models, scraped data, or any artifact with a plausible misuse pathway.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets are cited with their original publications and used under their respective research licenses (CC-BY 4.0, Decathlon licence, TotalSegmentator licence); the MedSigLIP encoder is cited and used under its publicly released weights.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code package includes a `README`, docstrings for all public functions, a minimal runnable example (`examples/minimal_example.py`), and an MIT licence file.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or research with human subjects; all data are pre-existing anonymised medical imaging datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new patient data were collected; the paper uses only publicly released, already-anonymised datasets whose original studies obtained the relevant ethics approvals.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the proposed method. The authors used LLM assistance for editing, data processing, and result visualisation, as declared at submission.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.