

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

Vincenzo Marciano^{1,2,3} Xiaoming Zhang^{1,2} Gianluca Guglielmo⁴
 Sébastien Ourselin³ Michela Antonelli³ Maria A Zuluaga^{1,3}

¹EURECOM ²Sorbonne Université ³King’s College London ⁴Technische Universität Graz

{vincenzo.marciano, xiaoming.zhang, maria.zuluaga}@eurecom.fr
 {sebastien.ourselin, michela.antonelli}@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 [17, 30, 25] to pan-organ foundation models [18, 26, 27, 38, 41]. Evaluating these models remains fundamentally tied to per-sample ground-truth (GT) annotations at test time. Dice similarity coefficient, Hausdorff distance (HD95), and normalised surface distance (NSD) [29] each measure how accurately a predicted mask reproduces the paired GT annotation at the voxel level. Although they were effective in controlled experimental conditions, several practical challenges remain to be addressed. First, every prediction must be matched to an annotated GT, an assumption that fails during prospective deployment and when evaluating 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, 22, 31, 39]; 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

(LPIPS) [43] 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) 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 therefore 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, medical imaging spans both volumetric and 2-D modalities (e.g., CT, MRI). 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 requiring any per-sample correspondence, MID correlates with mean Dice at Spearman $\rho = -0.55$. For comparison, HD95, a paired-reference geometric metric, reaches $|\rho| = 0.72$ with Dice on the same corrupted masks but requires a ground-truth mask for every prediction. While per-sample metrics remain more precise when ground truth is available, MID provides a complementary distribution-level signal for settings where candidate predictions cannot be annotated case by case.

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 cDice [36] and topologically faithful segmentation [37]. All of these metrics require per-sample GT alignment at test-time.

FID and distribution-level evaluation. FID [16] embeds images into Inception-v3 features and computes the squared Fréchet distance [11], which equals W_2^2 between the fitted Gaussians [9]. 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 organize by anatomy and generalize across datasets and modalities [33, 32]. 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, 33]. Scaling vision-only models, such as Rad-DINO [32], and contrastive image-text pretraining, including BiomedCLIP [44] and UniMedCLIP [21], further improved generalisation across anatomies, scanner types, and modalities. More recent large-scale medical vision encoders [35, 42] 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 [24]. 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 3.1–3.4 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 image inputs, f_θ produces vectors in the same shared feature space, allowing 2-D and 3-D samples to be represented within a single joint distribution.

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*

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

where $[\cdot \parallel \cdot]$ denotes the concatenation of the image and mask embeddings and $\pi : \mathbb{R}^{2d} \rightarrow \mathbb{R}^k$ 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).

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

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) [16]. 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. Under a Gaussian assumption, the resulting approximation preserves MID up to the discarded variance, with $|\text{MID}_k - \text{MID}| \leq 2 \sum_{i>k} \lambda_i$, where λ_i are the eigenvalues of the covariance.

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 MID_{img} computed from image embeddings $f_\theta(X)$, and a mask-only term MID_{mask} computed from mask embeddings $f_\theta(m_c)$. Concretely, $\text{MID}(P, Q)$ is computed on joint embeddings, $\text{MID}_{\text{img}}(P, Q)$ on image-only embeddings, and $\text{MID}_{\text{mask}}(P, Q)$ 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:

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

Unlike MID_{img} and MID_{mask} , MID_{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. Perturbations that preserve both image-only and mask-only distributions but alter their dependence structure, such as random image-mask shuffling, leave MID_{img} and MID_{mask} approximately unchanged while increasing MID_{corr} . The derivation of Equation 3 is given in Appendix F.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 distributional metrics against per-sample geometric metrics. Experiment 4 evaluates MID on outputs from real segmentation models. Experiment 5 isolates the role of image-mask correspondence via controlled perturbations. We also provide additional analyses on within-group spread (Appendix C.1), sample complexity (Appendix H.2), dataset-specific projections (Appendix H.1), and latent space geometry (Appendix I).

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 [20], 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. 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 image–ground-truth mask pairs 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 reserved as held-out candidate sets for cross-dataset evaluation.

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, 19], 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 B). Second, we use predictions from real segmentation models. SwinUNETR [25] and MedSAM [27] 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 [30] 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 [17], A2V [13], and SPOCKMIP [34].

Baselines. We compare MID against five baselines. **FID-Inception** computes standard FID using Inception-v3 [16] features on axial mid-slice renderings; this 2-D proxy discards volumetric information and mask 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 (Kernel Inception Distance)** [5] is an unbiased distributional baseline computed on mask features. Finally, **per-sample geometric metrics**, mean Dice and HD95 over matched –ground-truth pairs, are used as reference.

Implementation. For the main experiments, we instantiate f_θ with MedSigLIP [35]. For 3-D volumes, we follow the original MedSigLIP protocol and average patch-token embeddings across axial slices; for MAISI-based encoders (Appendix J), each 2-D axial slice is encoded as a 3-D input with depth 1. Alternative encoder choices are evaluated in Appendix J. 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 [23]. All code was implemented in Python 3.10 using PyTorch 2.0 and CUDA 12.4. All experiments use a fixed random seed. Code is available at [anonymous].

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

Table 1: MID–Dice Spearman correlation disaggregated 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

indicators. The quantitative analysis is performed over all such combinations in the full corruption sweep.

Figure 1 shows four representative cases from the sweep, selected to span different datasets, anatomies, and degradation patterns. In each panel, MID rises with degradation severity in step with the decline in Dice and the increase in HD95, mirroring per-sample paired metrics 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 C.2, Figure 7), clean-reference variability (Appendix C.1), and sample complexity (Appendix H.2), are reported in the appendix.

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 artifacts.

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 dataset and acquisition protocol as the reference. The resulting $\text{MID} = 0.011$ 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 $23\times$ increase over the floor, reflecting a modality shift under matched abdominal anatomy. Prostate MRI and ACDC cardiac MRI produce larger shifts ($\text{MID} = 0.932$ and 0.949), approximately $85\times$ above the floor, as anatomy, modality, and acquisition protocol all differ from the reference. This graded response shows that MID is sensitive to the type of shift and not only to its presence.

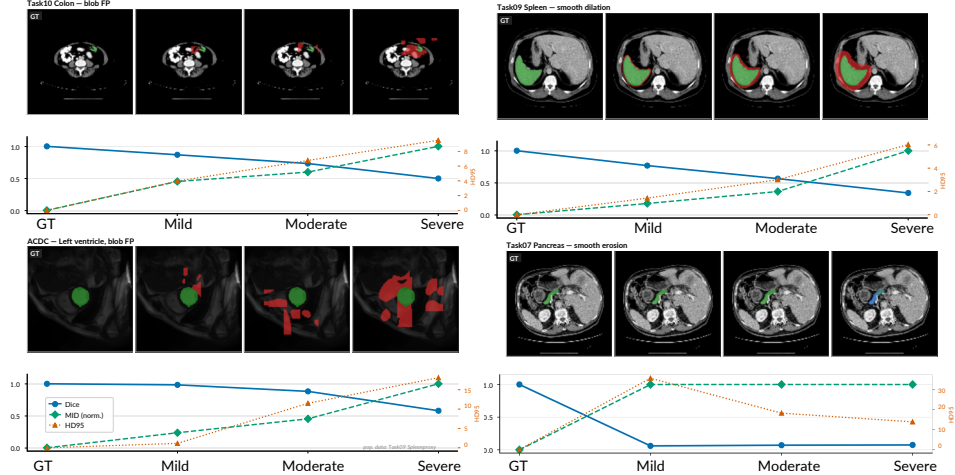


Figure 1: Representative examples from the controlled corruption sweep (Experiment 1). The four panels show Task10 Colon with blob-shaped false positives, Task09 Spleen with smooth dilation, ACDC left ventricle with blob-shaped false positives, and Task07 Pancreas with smooth erosion. Overlays show the ground-truth mask and corrupted masks at mild, moderate, and severe levels. Line plots show MID, Dice, and HD95 as a function of degradation severity. MID is normalised by the panel-wise maximum for visualisation only; all reported correlations use raw values.

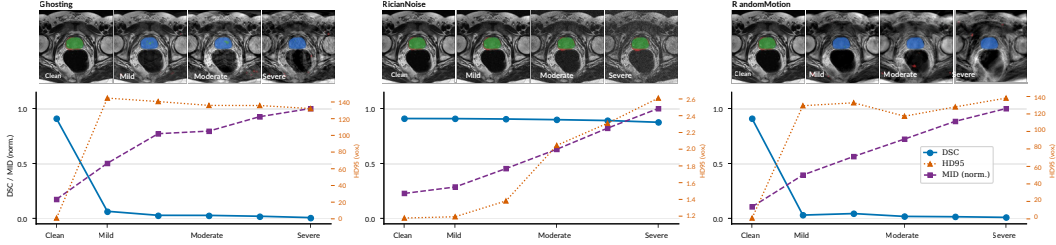


Figure 2: ROOD-MRI prostate on Task05_Prostate using SegResNet-300. For each artefact type, representative T2 slices across severity levels are shown above, with population metrics below.

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 D.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.

4.4 Metric Comparison

We compare scalar MID against distributional baselines and per-sample metrics on the same experimental setup as Experiment 1 ($56 \text{ organ groups} \times 5 \text{ corruption types} \times 3 \text{ severities} = 840 \text{ cells}$). Standard FID applied to Inception-v3 features of axial mid-slice renderings achieves $\rho = -0.38$ with Dice, discarding both volumetric and mask information. Replacing Inception with MedSigLIP mask features in full 3-D yields $\rho = -0.47$, a gain attributable to volumetric encoding. FID-img is essentially zero ($\rho = -0.03$) because images are held fixed across the sweep and only masks are

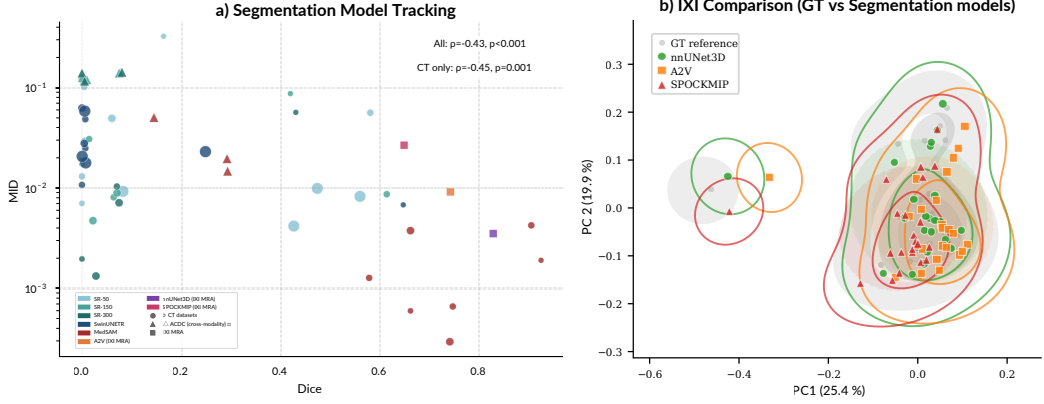


Figure 3: **a)** MID vs. Dice across real segmentation model outputs. Points are coloured by model; y-axis is log-scaled. **b)** Ground-truth and predicted-mask distributions for three IXI MRA vessel segmentation methods.

corrupted, confirming that mask information is needed to track segmentation quality. Mahalanobis distance from the joint reference Gaussian reaches $\rho = -0.42$, weaker than MID because it averages per-sample distances rather than comparing the full candidate distribution to the reference. 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 4 (Appendix E). FID-mask, the closest distributional baseline, operates on mask features alone and cannot capture image-mask agreement.

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 model–dataset–foreground-class group, 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 $\rho = -0.45$); including ACDC gives $\rho = -0.43$ ($p < 0.001$). Given that the reference is global, ACDC lies away from the CT trend: its elevated MID reflects cardiac-MRI/reference mismatch rather than segmentation quality alone. For IXI MRA, we use VesselVerse predictions from nnUNet, A2V, and SPOCKMIP [12]. MID ranks the three vessel segmentation methods in the same order as Dice and HD95: 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). The same ordering is observed for HD95 (4.6 mm vs. 20.7 mm for nnUNet and SPOCKMIP), indicating that MID recovers the model ranking without requiring per-case annotations at inference time. Additional setup details are in Appendix G.

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.

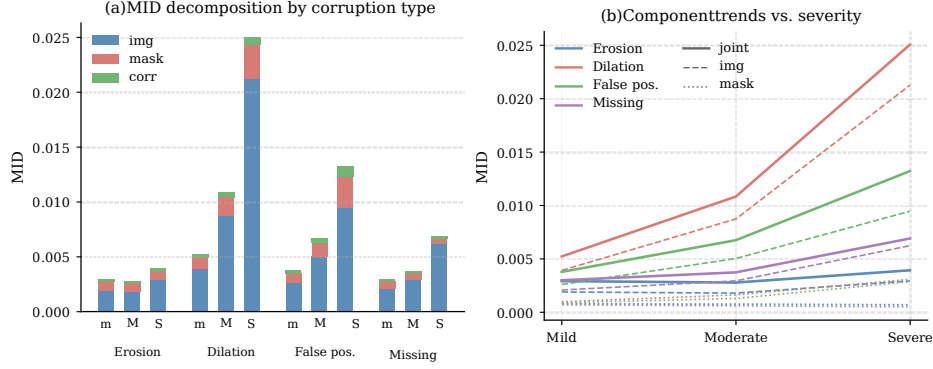


Figure 4: MID decomposition across corruption types on TotalSegmentator v2 CT. **(a)** Stacked bars show the image, mask, and residual correspondence terms for each corruption type and severity level (m = mild, M = moderate, S = severe). **(b)** Component trends as severity increases.

Under pair shuffling, MID_{img} and MID_{mask} remain near zero, while the residual correspondence term carries the signal. This isolates image-mask dependence: the marginal feature distributions are unchanged, but the joint distribution is not. Thus, MID_{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, MID_{img} dominates, consistent with modality-driven image appearance shift. Additional decomposition details are in Appendix F.

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 graded response to distributional shift, from same-distribution held-out data ($MID = 0.011$) through modality shift (0.278 – 0.949), demonstrates that MID is sensitive to the type of shift and not only its magnitude. 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 C.2). 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 metric is non-negative, symmetric, and consistent; 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 recovers correct model rankings without per-case annotations at inference time.

7 Acknowledgements

This work is funded by the European Union (ERC CARAVEL, 101171357) and by the French government, through the 3IA Côte d’Azur Investments project (ANR-23-IACL-0001). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the EU or the European Research Council. Neither the EU nor the granting authority can be held responsible for them.

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*, 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, Claudio Zotti, Frederic Cervenansky, and et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 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*, pages 1–1, 2026.
- [9] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017.
- [10] Terrance DeVries, Adriana Romero, Joelle Pineau, Graham W. Taylor, and Ross Goroshin. On the evaluation of conditional GANs. In *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 Marciano, 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] 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.
- [17] Fabian Isensee, Paul Friedrich Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: Self-adapting framework for deep learning-based bioedical image segmentation. *Nature Methods*, 2021.
- [18] 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.
- [19] Hana Jebiril, Thomas Pinetz, and Hrvoje Bogunović. Shape prior for quality assessment in octa via denoising autoencoders at the segmentation level. *IEEE Access*, 2025.
- [20] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, 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.
- [21] 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.
- [22] 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*, 2020.
- [23] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.
- [24] 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.
- [25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.
- [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, 2021.

- [29] Lena Maier-Hein et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21:195–212, 2024.
- [30] Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop*, 2018.
- [31] 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*, 2020.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyang Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Qiao Yu. SAM-Med3D. *arXiv preprint arXiv:2310.15161*, 2023.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.

- [44] 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 3 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.

Table 3: 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 for cross-dataset evaluation; the remaining 10 datasets form the reference pool described in Section 4.1. $\ast$ 56 foreground classes are used across CT and MRI; the full 104-class annotation is available. $\ddagger$ 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 $\bullet$ [20]	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 [40]	CT	104 $\ast$	7368
TotalSegmentator v2 MRI	MRI	104 $\ast$	1788
Total			17631

B 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.

C 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.

C.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.

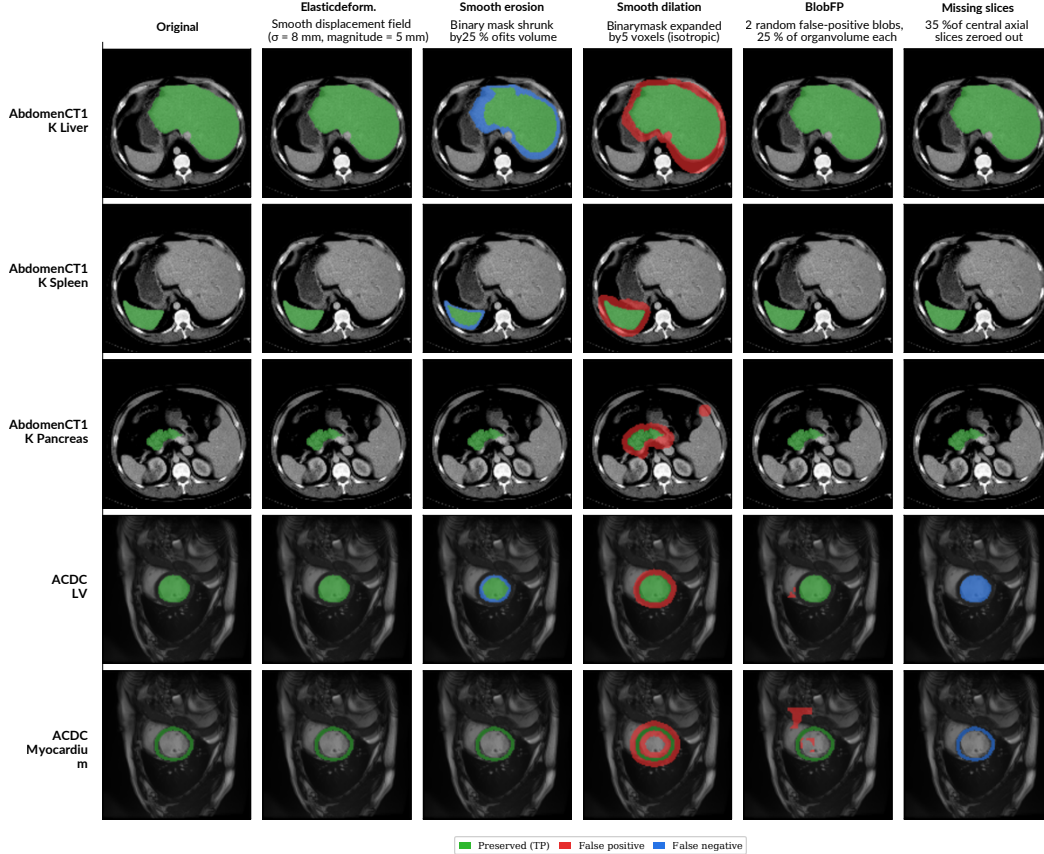


Figure 5: Anatomically realistic SDT-based mask corruptions applied to five organs. Columns show the original mask and five corruption types: elastic deformation, smooth erosion, smooth dilation, blob-shaped false positives, and partial missing slices. Overlay colours: green = preserved region, red =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

Figure 6 reports MID for clean-vs-clean comparisons within each dataset–organ group ($n = 56$ per group, with reference and candidate sets drawn from disjoint halves of the clean feature files). Dice is 1.0 for all groups by construction. The observed spread reflects natural anatomical variability and finite-sample estimation noise. This confirms that MID should be interpreted relative to anatomy-specific reference ranges rather than by a universal absolute threshold.

C.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 therefore provides a per-anatomy reference index, showing clean-baseline and corrupted-reference ranges for each foreground class.

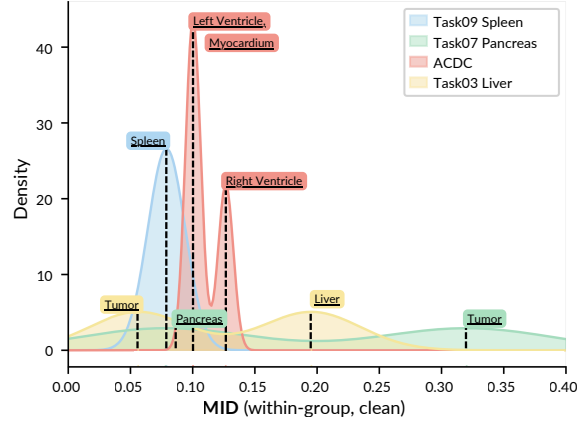


Figure 6: Within-group MID distribution across datasets. Each curve shows the kernel density estimate of clean-vs-clean MID values over organ groups within a dataset.

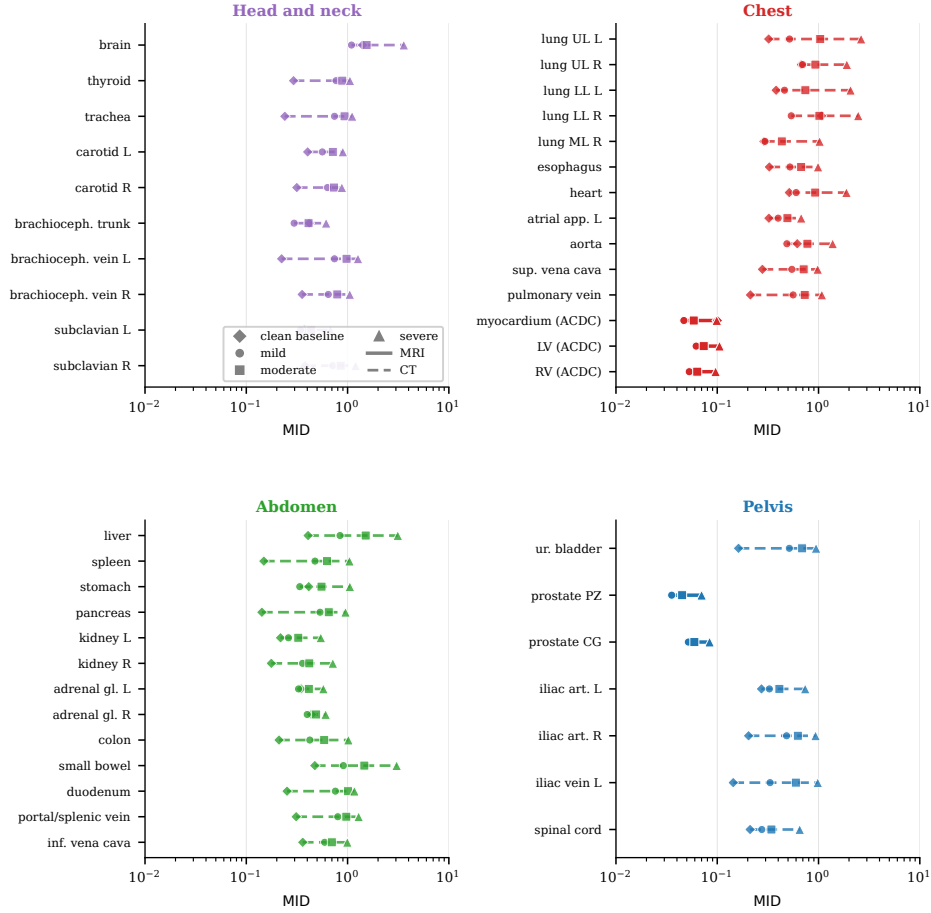


Figure 7: Per-anatomy MID reference index across the organ groups: CT/MRI structures from TotalSegmentator and overlapping MSD/AbdomenCT-1K annotations, plus cerebral arteries from IXI MRA. Each horizontal bar spans the within-group clean-baseline MID (diamond) to the severe-corruption MID (triangle), with mild (circle) and moderate (square) markers at intermediate levels. Values are averaged over five SDT-based corruption types. Solid lines indicate MRI-derived organs; dotted lines indicate CT organs.

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.

D 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.

D.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–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.

D.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 [30]. 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 artefact–severity condition, 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.

E 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 MID_{img} , MID_{mask} , and MID_{corr} is examined separately in Experiment 5 and Appendix F and is not included here.

Table 4: 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
<i>Ours</i>	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.

F 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.

F.1 Covariance Block Structure

For an image-mask pair (X, m_c) and the per-class joint embedding of Definition 1, write the projected joint vector as $z = (z_x; z_m) \in \mathbb{R}^k$. For a distribution P over such embeddings, the mean and covariance admit the block form

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

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

$$d_F^2(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{Tr}\left(\Sigma_P + \Sigma_Q - 2(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2})^{1/2}\right). \quad (5)$$

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

$$\|\mu_P - \mu_Q\|^2 = \|\mu_P^x - \mu_Q^x\|^2 + \|\mu_P^m - \mu_Q^m\|^2, \quad (6)$$

and the additive trace term decomposes as

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

The geometric-mean term $\text{Tr}(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2})^{1/2}$ 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.

F.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:

$$\text{MID}_{\text{img}}(P, Q) = \|\mu_P^x - \mu_Q^x\|^2 + \text{Tr}\left(\Sigma_P^{xx} + \Sigma_Q^{xx} - 2(\Sigma_P^{xx,1/2}\Sigma_Q^{xx}\Sigma_P^{xx,1/2})^{1/2}\right), \quad (8)$$

$$\text{MID}_{\text{mask}}(P, Q) = \|\mu_P^m - \mu_Q^m\|^2 + \text{Tr}\left(\Sigma_P^{mm} + \Sigma_Q^{mm} - 2(\Sigma_P^{mm,1/2}\Sigma_Q^{mm}\Sigma_P^{mm,1/2})^{1/2}\right). \quad (9)$$

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

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

This residual is the contribution of the geometric-mean term in (5) that is not captured by either marginal Fréchet distance, and therefore depends only on the off-diagonal blocks $\Sigma_P^{xm}, \Sigma_Q^{xm}$ that quantify image–mask coupling under each distribution.

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)–(7) gives

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

so $\text{MID}_{\text{corr}}(P, Q) = 0$. 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 MID_{corr} is generally non-zero. We therefore interpret MID_{corr} as the part of segmentation quality that is attributable to image–mask correspondence rather than to image or mask marginals alone.

F.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 MID_{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, MID_{img} and MID_{mask} should be invariant under shuffling, while MID_{corr} should absorb the entire effect. This is the prediction that Section 4.6 tests.

Overall, the experiment supports interpreting MID_{corr} as a diagnostic residual for correspondence changes rather than as a standalone metric.

G 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 model–dataset–organ group, 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 (the same 10-dataset reference used throughout the paper), without sample-wise matching between reference and candidate. Using the global reference pool, rather than a group-specific reference, lets us compare candidates from different models, datasets, and anatomies on a common scale; the ACDC offset visible in Figure 3a is a direct consequence of this choice and reflects the

CT/MRI mismatch with the predominantly CT reference rather than poor cardiac segmentation quality.

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. This setting is particularly informative because all methods are evaluated on the same anatomy and modality, avoiding cross-modality reference mismatch.

H Projection and Sample-Size Analyses

This appendix analyses two estimator choices used by MID: the PCA projection and the candidate sample size. Since MID relies on Gaussian statistics in a projected feature space, both the projection subspace and the number of candidate samples can affect the stability of the estimate.

H.1 Dataset-Specific Projections

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 full MedSigLIP feature corpus, providing a universal latent coordinate system. A practitioner with access to a single target dataset may instead fit the projection on that dataset, trading cross-dataset comparability for within-domain sensitivity.

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 5 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 5: 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 at most 1.5% 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, because pancreatic tumour masks in MSD are smaller and more variable than the full-organ masks used for the reference, and this scale difference reduces rather than increases 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, fitting PCA on a held-out reference split of the target data is appropriate and can improve within-domain sensitivity. For cross-dataset comparisons, the universal projection is preferable because it keeps all datasets in a common coordinate system.

H.2 Sample-Size Sensitivity

We next study how MID varies with the candidate set size. This analysis is important because medical imaging datasets often contain substantially fewer cases than natural-image benchmarks used for FID evaluation.

Table 6 reports the mean and coefficient of variation (CV) of MID as a function of candidate set size n , using 50 bootstrap replicates per n in a representative cross-dataset setting (TotalSeg CT reference vs. TotalSeg MRI candidate, asymptotic MID ≈ 0.30 , see Table 2). The reference set is fixed at $n = 500$.

Table 6: Sample-size sensitivity of MID in a cross-dataset setting (TotalSeg CT reference vs. TotalSeg MRI candidate; asymptotic MID ≈ 0.30). The reference set is fixed at $n = 500$; candidate sets are subsampled at different sizes. Mean and coefficient of variation (CV) are computed over 50 bootstrap replicates. The decreasing mean reflects the well-known 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%

The mean estimate decreases monotonically with n , consistent with the finite-sample positive bias of the Fréchet estimator, and stabilises around the asymptotic value (≈ 0.30) for $n \gtrsim 250$. The CV decreases roughly as $n^{-1/2}$, as expected for bootstrap variability of a moment-based estimator. Comparisons are most reliable when candidate sets have similar sample sizes, or when the evaluation protocol fixes the candidate set size across methods.

I 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 organizes samples by modality and anatomy in a way that is meaningful for distributional comparison.

I.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.

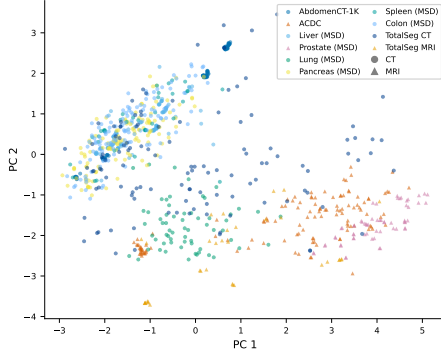


Figure 8: First two PCA components of joint MedSigLIP embeddings. CT datasets occupy a common region, whereas MRI datasets are displaced along the dominant modality-related axis.

Table 7: Mean MID between clean and corrupted latents, averaged over organ groups for each dataset. Larger values indicate greater displacement from the reference distribution in the universal PCA space.

Dataset	Modality	Groups	Mean MID
TotalSegmentator CT	CT	44	0.046
Task10 Colon	CT	1	0.077
Task09 Spleen	CT	1	0.081
AbdomenCT-1K	CT	4	0.093
Task03 Liver	CT	2	0.099
Task07 Pancreas	CT	2	0.105
Task06 Lung	CT	1	0.260
ACDC	MRI	3	0.072
Task05 Prostate	MRI	2	0.146
TotalSegmentator MRI	MRI	2	0.895

The PCA visualisation explains why MID naturally assigns larger scores to cross-modality shifts: MRI samples are displaced from CT samples in the joint feature space. For within-modality evaluation, this modality axis contributes much less to the total score.

I.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.

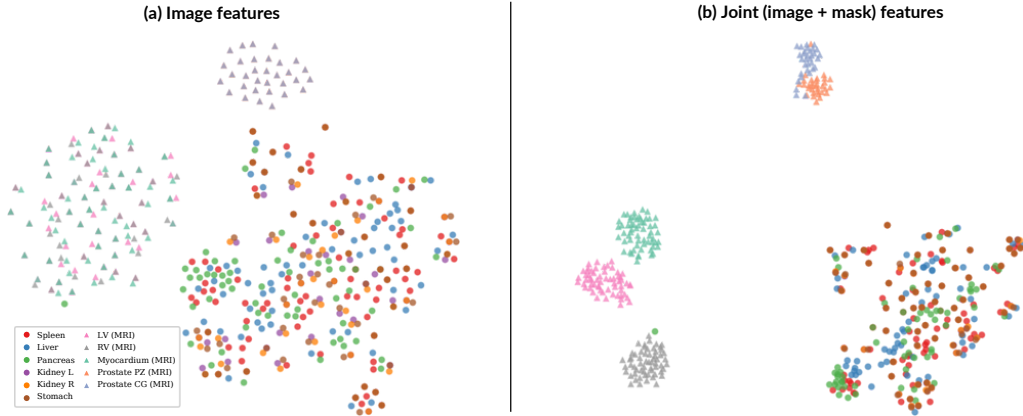


Figure 9: t-SNE visualisation of MedSigLIP embeddings for eleven organ types. Left: mask-only embeddings. Right: joint image-mask embeddings. Points are coloured by organ, with circles denoting CT samples and triangles denoting MRI samples. Anatomical structures form coherent clusters in both spaces, 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.

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.

J 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

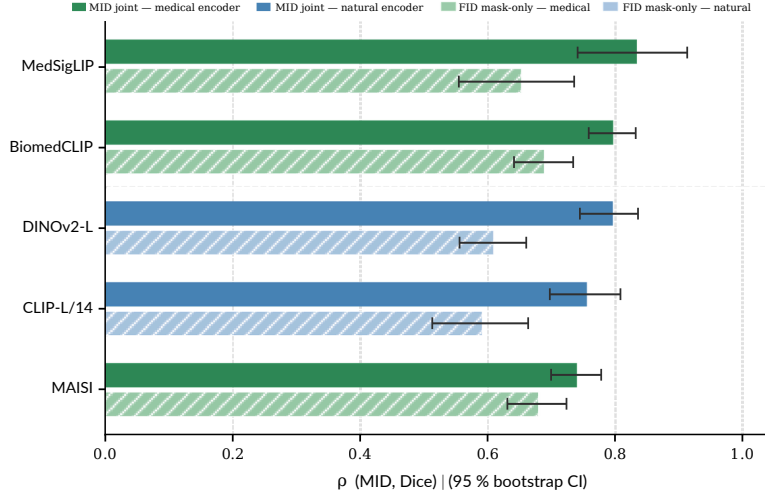


Figure 10: Mean per-organ $|\rho(\text{MID}, \text{Dice})|$ for five encoder backbones on the TotalSegmentator v2 CT corruption sweep. Each bar reports the mean over per-organ rank correlations.

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, 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.