Compact Metrics, Catastrophic Failures: Trustworthy Benchmarking of Vision-Language Foundation Models in 3D Medical Segmentation

Miroslaw Narbutt
School of Electrical and Electronic Engineering
Technological University Dublin
Dublin, Ireland.
miroslaw.narbutt@tudublin.ie

Tanmoy Debnath
CSU Engineering
Charles Sturt University
Bathurst, NSW, Australia.
tdebnath@csu.edu.au

Abstract

We present a controlled benchmark of three 3D MRI segmentation paradigms, nnU-Net v2, VSmTrans, and the Universal Model, across Heart, Hippocampus, and Prostate tasks from the Medical Segmentation Decathlon using overlap, boundary, and volumetric metrics. Case-level correlation analysis reveals substantial redundancy among commonly co-reported overlap metrics, motivating an empirically validated compact eight-metric reporting subset that preserves the principal ranking and failure-detection conclusions of the full twelve-metric suite while reducing reporting overhead by 33%. Rankings are anatomy-dependent: nnU-Net v2 leads on Heart and Prostate and the Universal Model on Hippocampus. The VSmTrans emerges as the most balanced all-rounder, remaining consistently competitive across tasks and providing the strongest volumetric fidelity without catastrophic task-level collapse. However, the Universal Model exhibits severe Prostate boundary failure (HD95 42.9 ± 40.2 mm) that would be obscured by Dice-only reporting, with direct implications for trustworthy deployment in patient-specific digital-twin pipelines.

1 Introduction

Automated magnetic resonance imaging (MRI) segmentation constitutes a foundational primitive for biomedical data-mining pipelines, including radiotherapy planning, longitudinal disease monitoring, quantitative biomarker extraction, and prostate cancer risk stratification (Antonelli et al., 2022; Gao et al., 2025). Segmentation accuracy thus directly constrains Digital Twin fidelity and the reliability of downstream simulation and treatment planning. Three converging evaluation failures compound this problem. First, the Dice Similarity Coefficient conflates voxel-level overlap with boundary and volumetric accuracy, yet the latter two are often independently relevant (Fazekas et al., 2026). Second, standard evaluation batteries include functionally redundant metrics (Dice, MCC, Cohen's Kappa, and Pearson are algebraically near-equivalent on binary tasks), inflating apparent comprehensiveness without adding diagnostic information (Kocak et al., 2025). Third, aggregate means over small cohorts mask case-level outlier behavior, particularly for architectures exhibiting bimodal failure on heterogeneous structures (Debnath and Narbutt, 2025). To address these gaps, we present a controlled, multi-metric comparison of nnU-Net v2, the Universal Model, and VSmTrans across three Medical Segmentation Decathlon (MSD) tasks in cardiac, neuroanatomical, and pelvic domains (Antonelli et al., 2022), evaluated under unified preprocessing, validation, and hardware conditions.

The principal contributions are:

1. *Controlled cross-architecture evaluation* of nnU-Net v2, VSmTrans, and the Universal Model across Heart, Hippocampus, and Prostate under matched conditions.
2. *Compact metric operationalization:* Complementing Metrics Reloaded (Maier-Hein et al., 2024), case-level Spearman correlation and ranking-stability analysis over twelve commonly reported metrics identify an empirically validated compact eight-metric reporting subset. This preserves the principal conclusions of the full suite in this benchmark while reducing reporting overhead by 33%.
3. *Anatomy-stratified statistical analysis:* Holm–Bonferroni-corrected Wilcoxon signed-rank tests with effect sizes show architectural rankings are anatomy-dependent and non-transferable.
4. *Robustness analysis:* Ranking stability under metric-subset perturbation and five-fold cross-validation confirms reproducibility for Heart and Prostate, with inflated statistical power warranting caution for Hippocampus.

2 Related Work

U-Net (Ronneberger et al., 2015) and its volumetric extension (Çiçek et al., 2016) established encoder-decoder skip-connection designs as the dominant segmentation paradigm. nnU-Net (Isensee et al., 2021) automated this via dataset-fingerprint-driven configuration; nnUZoo (Mohammadi Kazaj et al., 2025) and U-Bench (Tang et al., 2025) both confirm that many purportedly superior architectures fail to achieve reproducible improvements over a well-configured CNN baseline, establishing nnU-Net v2 as the necessary reference point. Self-attention and hybrid CNN–Transformer architectures (Hatamizadeh et al., 2022; Tang et al., 2022; Ren and Li, 2025; Li et al., 2023) extend the convolutional receptive field for long-range volumetric reasoning. VSmTrans (Liu et al., 2024) couples convolutional encoding with Swin-based self-attention, enabling independent assessment of the global-attention contribution without inheriting the nnU-Net training pipeline. CLIP-conditioned architectures (Li et al., 2026) enable cross-task generalization via prompt conditioning. The Universal Model (Zhang et al., 2026) employs 3D convolutional encoders with image–text alignment. Extensions such as VoxTell (Rokuss et al., 2025), MedCLIP-SAMv2 (Koleilat et al., 2025), and efficient 3D SAM-lifting frameworks (Feng et al., 2026) demonstrate that foundation-model coverage can be extended to volumetric domains, though inference cost remains a deployment constraint in clinical pipelines. However, recent reviews document consistent generalization failure when contrastive vision–language objectives encounter domain-specific structural heterogeneity not represented in pre-training distributions (Lian et al., 2026; Guan et al., 2025; Chen et al., 2023), a pattern we also observe on Prostate segmentation and attribute to coarse contrastive alignment (Qin and Yang, 2026). The Metrics Reloaded framework (Maier-Hein et al., 2024) provides principled guidance for selecting segmentation metrics according to clinical use (Gao et al., 2025), error tolerance (Pusparani et al., 2026), and task characteristics (Qin and Yang, 2026). Our contribution is narrower and complementary: rather than introducing a new general taxonomy, we examine which commonly co-reported metrics remain empirically distinct at case level in a controlled 3D MRI benchmark, and whether a reduced subset preserves ranking stability and failure-mode visibility.

Table 1: Technical specifications of the three MSD datasets.

Attribute	Task02 Heart	Task04 Hippocampus	Task05 Prostate
Task Description	Left atrium (LA) cavity	Anterior, Posterior	Peripheral, Transitional
Clinical Application	Atrial anatomy; arrhythmia	RT planning	Prostate cancer assessment
Segmentation Challenges	Few mono-MRI cases; thin walls	Small structures; many cases	Multi-modal; limited cases
Cases (Total/Train/Test)	30/20/10	390/260/130	48/32/16
Imaging Modality	MRI (LGE)	MRI (T1 MPRAGE)	MRI (T2 + ADC)
Annotated Structures	Background, LA	Background, Ant., Post.	Background, PZ, TZ
Spatial Resolution	Variable; anisotropic	~1 mm MPRAGE	0.6×0.6×4.5 mm
Tensor Size	3D (1 ch.)	3D (1 ch.)	3D (2 ch.)
Preprocessing	Cropping; normalization; resampling	VOI extraction; resampling; normalization	Normalization; cropping

3 Methodology

3.1 Architectures

The three architectures represent distinct 3D segmentation paradigms. nnU-Net v2 is a configuration-driven CNN baseline that automates preprocessing, architecture selection, and training schedules. The Universal Model employs 3D convolutional encoders with CLIP-style image–text alignment for prompt-conditioned, cross-task inference (Zhang et al., 2026). VSmTrans couples convolutional encoding with Swin Transformer self-attention for long-range volumetric reasoning without relying on the nnU-Net training pipeline (Liu et al., 2024).

3.2 Data and Preprocessing

Experiments used Task02 (Heart), Task04 (Hippocampus), and Task05 (Prostate) from the MSD (Antonelli et al., 2022) (Table 1). For nnU-Net v2 and VSmTrans, the automated nnU-Net v2 pipeline inferred dataset properties from MSD metadata and applied RAS reorientation with task-specific median-spacing resampling. It further performed percentile-based intensity clipping (0.5–99.5%), z-score normalization, and foreground-aware cropping. The Universal Model followed its official workflow with channel-wise z-score normalization and fixed $128 \times 128 \times 128$-voxel ROI extraction. Online augmentation (random flips, rotations, scaling, intensity perturbations) was applied consistently across all models.

3.3 Implementation Details

All models were implemented in PyTorch and trained on a single NVIDIA RTX 5090 (32 GB VRAM) with Automatic Mixed Precision. Training ran for up to 1000 epochs with early stopping (patience = 100; metric: Validation Dice). nnU-Net v2 used SGD with Nesterov momentum (0.99) and polynomial decay (lr = 10^{-2}); VSmTrans used Adam (10^{-4}) with polynomial decay; the Universal Model used AdamW (2×10^{-4}) with batch size one. All checkpoints were archived with SHA-256 hashes for reproducibility. Five-fold cross-validation was applied to Hippocampus; stratified hold-out validation to Heart and Prostate given their limited cohort sizes.

3.4 Evaluation Metrics and Aggregation

We evaluate using three metric families (Debnath et al., 2026; Kocak et al., 2025): (1) *Overlap:* Dice, Precision, Recall, IoU, MCC, Cohen's Kappa, Pearson. (2) *Boundary:* HD95, ASSD. (3) *Volumetric:* Volume Similarity (VS), Relative Volume Error (RVE = $(V_{\text{pred}} - V_{\text{gt}})/V_{\text{gt}}$, signed), and Volume Similarity Error (VSE = $|V_{\text{pred}} - V_{\text{gt}}|/(V_{\text{pred}} + V_{\text{gt}})$, unsigned) (Bédard et al., 2025; Kocak et al., 2025). Dispersion is visualized with mean ±1 SD (thick whiskers) and min–max range (thin whiskers).

3.5 Statistical Analysis

Paired case-level Wilcoxon signed-rank tests (two-sided) were computed per organ and metric, with family-wise error controlled via Holm–Bonferroni correction across all metric × pairwise-comparison tests per organ ($\alpha = 0.05$). For Hippocampus, metrics were computed per test case within each fold and pooled across held-out cases. Effect sizes are reported as mean differences and standard deviations throughout.

4 Results

4.1 Metric Intercorrelation and Selection

Inflated metric batteries can mask model behavior by relying on correlated proxies. To quantify redundancy, we computed Spearman rank correlations among all twelve metrics pooled at case level across models and organs ($n = [N_{\text{Heart}} + N_{\text{Prostate}} + N_{\text{Hippo_cases}}] \times 3$ model–case pairs). Three functionally distinct clusters emerged: *(i)* overlap and agreement metrics (Dice, MCC, Kappa, Pearson, IoU; $\rho > 0.95$ within cluster); *(ii)* boundary metrics (HD95, ASSD; $\rho \approx 0.85$ within cluster and $\rho \approx 0.60$ relative to cluster i); and *(iii)* volumetric metrics (VS, VSE, RVE; moderate intra-cluster correlation and weaker coupling to clusters i and ii, $\rho \approx 0.50$–0.70). Precision and Recall exhibited intermediate correlations with cluster i ($\rho \approx 0.80$–0.90) while remaining diagnostically useful because they distinguish over-segmentation from under-segmentation errors not captured by Dice alone. The strong redundancy within cluster (i) is consistent with established relationships among Dice, MCC, Kappa, and Pearson correlation in binary classification, indicating that these theoretical dependencies persist across heterogeneous segmentation outputs. Consistent with *Metrics Reloaded* (Maier-Hein et al., 2024), these results support reducing redundant metrics while retaining complementary measures that capture distinct error modes.

We recommend a compact reporting subset comprising Dice, Precision, Recall, IoU, HD95, ASSD, RVE, and VS. Dice and IoU quantify overlap, Precision and Recall characterize over- and under-segmentation tendencies, HD95 and ASSD assess boundary quality, VS measures unsigned volumetric agreement, and RVE captures signed volumetric bias. This reduces the reporting suite from twelve to eight metrics while preserving the principal ranking and failure-detection conclusions of the full benchmark. We emphasize that this subset is an empirical operationalization for the present benchmarking exercise rather than a universal prescription.

4.2 Heart Segmentation (Left Atrium, Task02)

4.2.1 Overlap and Classification Metrics

Overlap and classification metric distributions are visualized across all tasks in Figure 1, and boundary metrics in Figure 2. On the Heart dataset, nnU-Net v2 achieved the highest performance across all overlap and classification metrics: Dice 0.939 ± 0.009, Precision 0.937 ± 0.024, Recall 0.942 ± 0.029, and IoU 0.885 ± 0.016, all substantially exceeding both competitors. VSmTrans performed competitively (Dice: 0.934 ± 0.006; IoU: 0.876 ± 0.011), ranking second across most measures and trailing nnU-Net v2 by narrow margins ($p < 0.05$). The Universal Model consistently underperformed, achieving Dice of 0.894 ± 0.010, with statistically significant degradation relative to both alternatives ($p < 0.01$).

Table 2: Mean ± STDEV across metrics for the three models on MSD tasks (Heart, Hippocampus, Prostate). Arrows indicate the preferred direction; the best value per row and dataset is in **bold**.

	Heart			Hippocampus			Prostate		
Metric	**nnU-Net**	**Universal**	**VsmTrans**	**nnU-Net**	**Universal**	**VsmTrans**	**nnU-Net**	**Universal**	**VsmTrans**
Dice(↑)	**0.939 ± 0.009**	0.894 ± 0.010	0.934 ± 0.006	0.890 ± 0.003	**0.902 ± 0.007**	0.890 ± 0.003	**0.800 ± 0.108**	0.512 ± 0.228	0.724 ± 0.136
Prec.(↑)	**0.937 ± 0.024**	0.872 ± 0.021	0.936 ± 0.023	0.892 ± 0.003	0.891 ± 0.006	**0.893 ± 0.004**	**0.792 ± 0.139**	0.570 ± 0.281	0.727 ± 0.189
Rec.(↑)	**0.942 ± 0.029**	0.919 ± 0.019	0.933 ± 0.026	0.890 ± 0.004	**0.915 ± 0.010**	0.890 ± 0.006	**0.821 ± 0.112**	0.537 ± 0.176	0.737 ± 0.105
IoU(↑)	**0.885 ± 0.016**	0.809 ± 0.017	0.876 ± 0.011	0.803 ± 0.005	**0.824 ± 0.011**	0.804 ± 0.005	**0.679 ± 0.144**	0.374 ± 0.198	0.585 ± 0.159
MCC(↑)	**0.939 ± 0.009**	0.895 ± 0.010	0.934 ± 0.006	0.888 ± 0.003	**0.900 ± 0.007**	0.888 ± 0.003	**0.801 ± 0.106**	0.565 ± 0.175	0.725 ± 0.135
Kappa(↑)	**0.939 ± 0.009**	0.894 ± 0.010	0.934 ± 0.006	0.887 ± 0.003	**0.900 ± 0.007**	0.887 ± 0.003	**0.798 ± 0.108**	0.509 ± 0.227	0.722 ± 0.136
Pearson(↑)	**0.939 ± 0.009**	0.895 ± 0.010	0.934 ± 0.006	0.888 ± 0.003	**0.900 ± 0.007**	0.888 ± 0.003	**0.801 ± 0.106**	0.565 ± 0.175	0.725 ± 0.135
HD95(↓)	**2.768 ± 0.048**	6.418 ± 2.337	3.083 ± 0.430	1.230 ± 0.035	**1.156 ± 0.028**	1.261 ± 0.036	**4.893 ± 1.488**	42.924 ± 40.166	5.981 ± 1.750
ASSD(↓)	**0.960 ± 0.136**	1.811 ± 0.203	0.993 ± 0.068	0.422 ± 0.012	**0.380 ± 0.025**	0.422 ± 0.012	**1.412 ± 0.388**	8.378 ± 7.306	1.886 ± 0.341
VS(↑)	0.976 ± 0.012	0.971 ± 0.014	**0.980 ± 0.015**	0.969 ± 0.002	**0.971 ± 0.002**	0.969 ± 0.002	**0.920 ± 0.056**	0.778 ± 0.261	0.914 ± 0.059
RVE(↓)	0.006 ± 0.053	0.054 ± 0.040	**−0.003 ± 0.049**	0.001 ± 0.006	0.030 ± 0.009	**0.000 ± 0.009**	0.062 ± 0.201	**−0.073 ± 0.310**	0.065 ± 0.235
VSE(↓)	0.048 ± 0.023	0.061 ± 0.029	**0.040 ± 0.029**	0.063 ± 0.005	**0.060 ± 0.004**	0.063 ± 0.003	**0.168 ± 0.126**	0.268 ± 0.172	0.188 ± 0.155

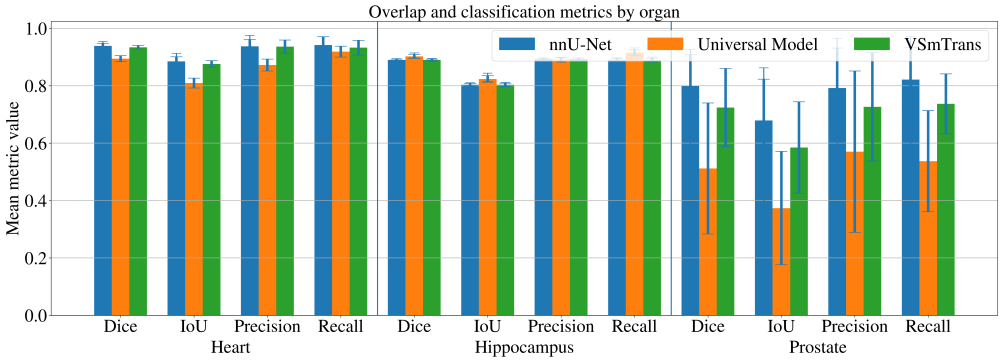

Figure 1: Overlap metrics (Dice, IoU, Precision, Recall) across three tasks and models. Bars: mean; thick whiskers: ±std; thin whiskers: min–max. Variability: Hippocampus = five folds, Heart/Prostate = cases.

4.2.2 Boundary Localization

nnU-Net v2 achieved the lowest HD95 (2.768 ± 0.048 mm), significantly outperforming both the Universal Model (6.418 ± 2.337 mm; $p < 0.001$) and VSmTrans (3.083 ± 0.430 mm; $p < 0.01$). ASSD followed congruently: nnU-Net v2 (0.960 ± 0.136 mm) was significantly better than the Universal Model (1.811 ± 0.203 mm; $p < 0.001$) and marginally better than VSmTrans (0.993 ± 0.068 mm; $p < 0.05$, $\Delta = 0.033$ mm). The narrow VSmTrans gap confirms competitive surface alignment; nnU-Net v2's superior HD95 reflects greater robustness to extreme atrial boundary deviations.

4.2.3 Volumetric Fidelity and Bias

VSmTrans achieved significantly higher Volume Similarity than both nnU-Net v2 and the Universal Model (0.980 ± 0.015 vs. 0.976 ± 0.012 and 0.971 ± 0.014, respectively; $p < 0.05$), with a correspondingly lower VSE (0.040 ± 0.029; $p < 0.05$). Directional bias analysis revealed that the Universal Model exhibited systematic over-segmentation (RVE: 0.054 ± 0.040), VSmTrans showed mild under-segmentation (RVE: −0.003 ± 0.049), and nnU-Net v2 remained approximately unbiased (RVE: 0.006 ± 0.053). Despite nnU-Net v2's superior overlap metrics, VSmTrans's volumetric fidelity suggests a dissociation between voxel-wise agreement and volume preservation on cardiac data, with implications for chamber-volume quantification applications.

4.3 Hippocampus Segmentation (Anterior/Posterior, Task04)

4.3.1 Overlap and Classification Metrics:

On the Hippocampus dataset, the Universal Model reversed the performance hierarchy observed on Heart, achieving the highest Dice (0.902 ± 0.007) with statistically significant improvements over both nnU-Net v2 (0.890 ± 0.003; $p < 0.001$) and VSmTrans (0.890 ± 0.003; $p < 0.001$). Recall was similarly dominant at 0.915 ± 0.010 (vs. nnU-Net v2: 0.890 ± 0.004; VSmTrans: 0.890 ± 0.006; $p < 0.01$), indicating superior true-positive rates on small neuroanatomical structures. IoU followed this pattern (Universal Model: 0.824 ± 0.011 vs.

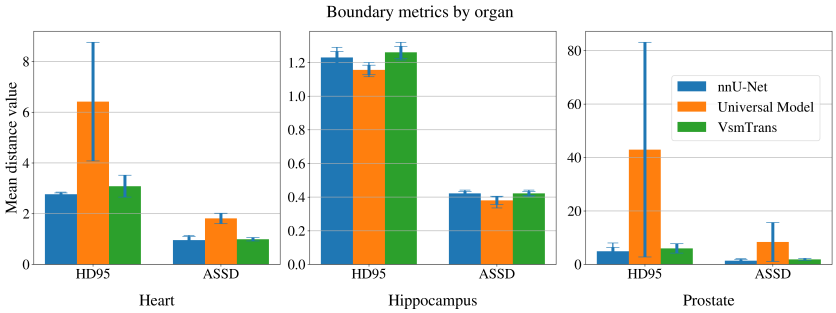

Figure 2: HD95 and ASSD for Heart, Hippocampus, and Prostate. Lower is better.

nnU-Net v2: 0.803±0.005 and VSmTrans: 0.804± 0.005; $p < 0.01$). VSmTrans achieved marginally highest Precision (0.893 ± 0.004), though the difference was not statistically significant ($p > 0.1$). Dice differences of 0.010–0.012 between models, while statistically significant, require careful interpretation. The large sample size afforded by five-fold cross-validation ($n_{\text{cases}} = 390$) endows even small effect sizes with statistical power. Whether a 1.2% Dice improvement constitutes a clinically meaningful distinction depends on downstream application (e.g., neurosurgical guidance versus research phenotyping).

4.3.2 Boundary Localization

The Universal Model achieved the lowest HD95 (1.156 ± 0.028 mm), better than both nnU-Net v2 (1.230 ± 0.035 mm) and VSmTrans (1.261 ± 0.036 mm; $p < 0.05$). ASSD followed congruently: Universal Model 0.380 ± 0.025 mm versus nnU-Net v2 0.422 ± 0.012 mm and VSmTrans 0.422 ± 0.012 mm ($p < 0.01$). These differences, while statistically significant, are numerically small in absolute terms (HD95 gap: 0.074–0.105 mm). VSmTrans and nnU-Net v2 showed comparable boundary results with no significant difference in ASSD ($p > 0.1$), suggesting that convolutional-based approaches—whether pure CNN or hybrid—exhibit similar boundary mapping behavior on hippocampal substructures.

4.3.3 Volumetric Fidelity and Cross-Fold Consistency

All models achieved high volumetric consistency (VS: 0.969–0.971), with no statistically significant difference between nnU-Net v2 and VSmTrans ($p > 0.1$). The Universal Model achieved a marginal but statistically significant improvement over nnU-Net v2 (0.971 ± 0.002 vs. 0.969 ± 0.002; $p < 0.05$). VSE favored the Universal Model (0.060 ± 0.004) over both nnU-Net v2 (0.063±0.005; $p < 0.01$) and VSmTrans (0.063 ± 0.005; $p < 0.01$). Signed RVE revealed mild over-segmentation for the Universal Model (0.030±0.009), near-zero bias for VSmTrans (0.000 ± 0.009), and stable behavior for nnU-Net v2 (0.001 ± 0.006). Five-fold cross-validation provides direct evidence of reproducibility: standard deviations across all models remain narrow (Dice SD ≤ 0.007; HD95 SD ≤ 0.036), confirming highly stable cross-fold behavior and validating the pooling of case-level metrics across folds for paired statistical tests.

4.4 Prostate Segmentation (Peripheral/Transitional Zones, Task05)

4.4.1 Overlap and Classification Metrics

The Prostate dataset revealed the greatest inter-model divergence. nnU-Net v2 achieved the highest performance across overlap metrics (Dice: 0.800 ± 0.108; Precision: 0.792 ± 0.139; Recall: 0.821 ± 0.112; IoU: 0.679 ± 0.144), with statistically significant improvements over both VSmTrans (Dice: 0.724 ± 0.136; $p < 0.01$) and the Universal Model (Dice: 0.512 ± 0.228; $p < 0.001$). The Universal Model exhibited substantial degradation (IoU: 0.374 ± 0.198; MCC: 0.565 ± 0.175), constituting a failure regime rather than graceful degradation. VSmTrans achieved intermediate performance with moderate variability (Dice SD: 0.136).

4.4.2 Boundary Localization

Boundary metrics revealed extreme sensitivity to anatomical heterogeneity. nnU-Net v2 achieved the lowest HD95 (4.893 ± 1.488 mm), significantly outperforming VSmTrans (5.981 ± 1.750 mm; $p < 0.05$) and dramatically outperforming the Universal Model (42.924 ± 40.166 mm; $p < 0.001$). The Universal Model's mean HD95 represented an order-of-magnitude degradation, while its standard

deviation (40.166 mm) approached the mean, indicating substantial case-level instability. ASSD analysis further confirmed the superiority of nnU-Net v2 (1.412 ± 0.388 mm), compared with VSmTrans (1.886 ± 0.341 mm; $p < 0.05$) and the Universal Model (8.378 ± 7.306 mm; $p < 0.001$). The Universal Model's high relative variance (SD/mean ≈ 0.87) suggests that a subset of cases experienced severe boundary delineation failures, consistent with the known sensitivity of vision–language models to morphologically atypical or heterogeneously imaged anatomy.

4.4.3 Volumetric Fidelity and Bias

nnU-Net v2 achieved significantly higher VS (0.920±0.056) than both VSmTrans (0.914±0.059) and the Universal Model (0.778 ± 0.261; $p < 0.01$), and the lowest VSE (0.168 ± 0.126; $p < 0.05$ vs. VSmTrans). Signed RVE revealed asymmetric bias across models: the Universal Model showed unstable under-segmentation (−0.073 ± 0.310), with variance exceeding the mean magnitude by more than fourfold, indicating that directional bias is not a reliable summary of its systematic behavior. In contrast, nnU-Net v2 and VSmTrans maintained stable near-zero RVE (0.062 ± 0.201 and 0.065 ± 0.235, respectively), consistent with substantially more controlled volumetric bias.

4.5 Statistical Significance Summary

Table 3: Holm–Bonferroni-corrected p-values for pairwise Wilcoxon signed-rank tests. † denotes $p < 0.05$; ‡ denotes $p < 0.001$; n.s. denotes not significant.

Comparison	Heart	Hippo.	Prostate
nnU-Net vs. Universal (Dice)	‡	‡	‡
nnU-Net vs. VSmTrans (Dice)	†	n.s.	‡
Universal vs. VSmTrans (Dice)	‡	‡	‡
nnU-Net vs. Universal (HD95)	‡	†	‡
nnU-Net vs. VSmTrans (HD95)	†	n.s.	†
nnU-Net vs. Universal (VS)	n.s.	†	‡
nnU-Net vs. VSmTrans (VS)	†	n.s.	n.s.

Table 3 summarizes Holm–Bonferroni-corrected Wilcoxon signed-rank test results for each pairwise model comparison across organs and metric families. Effect sizes, reported throughout Sections 4.2–4.4 as mean differences and standard deviations, confirm that statistically significant differences correspond to practically meaningful gaps in most cases. The exception is the Hippocampus dataset, where Dice differences of 0.010–0.012, although statistically significant under the large sample enabled by five-fold cross-validation, may not correspond to clinically meaningful separation and therefore must be interpreted in the context of absolute effect size. By contrast, the Prostate difference between nnU-Net v2 and the Universal Model (Dice: 0.800 vs. 0.512; $\Delta = 0.288$) constitutes a substantial practical gap that would materially affect clinical utility regardless of sample size.

4.6 Robustness and Sensitivity Analysis

The reliability of architectural conclusions depends on whether observed rankings are stable across (1) variation in the metric subset used for evaluation, (2) case-level outlier influence, and (3) cross-validation fold perturbation.

4.6.1 Ranking Stability Under Metric Subset Perturbation

To assess whether model ordering reflects genuine architectural differences or metric-selection artifacts, we computed aggregate model ranks under four metric subsets: the full twelve-metric suite, the non-redundant eight-metric battery proposed in Section 4.1, a boundary-only subset (HD95, ASSD), and an overlap-only subset (Dice, Precision, Recall, IoU). For each subset and organ, models were ranked by mean position across included metrics (rank 1 = best; ties broken by mean metric value). Model orderings are largely preserved across metric subsets. For Heart and Prostate, rankings are invariant regardless of which metric family is prioritized, confirming that nnU-Net v2's superiority is not a consequence of metric-selection bias. For Hippocampus, the only rank perturbation occurs in the boundary-only subset, where VSmTrans (HD95: 1.261 ± 0.036 mm) and nnU-Net v2 (HD95: 1.230 ± 0.035 mm) swap positions relative to the overlap-only ordering. This swap occurs within a numerically narrow band (ΔHD95 = 0.031 mm) and is not statistically significant ($p > 0.1$ after Holm–Bonferroni correction). It does not constitute a meaningful reversal of the architectural conclusion. The Universal Model's dominance on hippocampal segmentation and its severe degradation on Prostate are both stable across all four subsets, reinforcing the anatomical specificity of these findings.

4.6.2 Inter-Case Variability and Outlier Sensitivity

Aggregate metric means can be dominated by a small number of failure cases, particularly in small test cohorts. Table 4 presents the coefficient of variation (CV = SD/mean) for Dice and HD95 by model and task, where CV exceeding 0.20 indicates substantial sensitivity to individual cases. On Heart,

Table 4: Coefficient of variation (CV = SD/mean) > 0.20 (marked †) indicates substantial case-level sensitivity.

	Heart		Hippocampus		Prostate	
Model	Dice	HD95	Dice	HD95	Dice	HD95
nnU-Net v2	0.010	0.017	0.003	0.028	0.135†	0.304†
Universal	0.011	0.364†	0.008	0.024	0.445†	0.936†
VSmTrans	0.006	0.139†	0.003	0.029	0.188†	0.293†

all models exhibit low inter-case variability (Dice CV < 0.02), indicating robust generalization across the ten test cases. On Hippocampus, cross-fold variability is negligible (Dice CV ≤ 0.008; HD95 CV ≤ 0.029), validating the five-fold cross-validation approach. On Prostate, inter-case heterogeneity becomes pronounced. nnU-Net v2 exhibits moderate sensitivity (Dice CV: 0.135; HD95 CV: 0.304); the Universal Model reaches extreme outlier sensitivity (Dice CV: 0.445; HD95 CV: 0.936), with RVE variance exceeding its mean by more than fourfold. This pattern indicates that the Universal Model performs adequately on anatomically typical cases but fails catastrophically on morphologically atypical anatomy. As a result, it produces a bimodal case-level distribution that aggregate means cannot reliably represent (Debnath and Narbutt, 2025). This analysis underscores the importance of reporting full distributional summaries, including means, standard deviations, and min–max ranges. It also supports the use of robust location statistics, such as the median and interquartile range, as complementary summaries in small-cohort benchmarking.

4.7 Cross-Task Performance Heatmap

Figure 3 provides a rank-based synthesis of relative model performance, consolidating the twelve metrics reported in Table 2 into a compact visual summary across organs (green: rank 1, best; yellow: rank 2; red: rank 3, worst). Three distinct cross-task performance profiles emerge.

nnU-Net v2 exhibits the strongest overall ranking stability. It places first across overlap, boundary, and volumetric metric families on Heart and Prostate, and second on Hippocampus. This pattern indicates that its configuration-driven design generalizes well across anatomies with differing levels of morphological complexity, with the clearest advantage observed on the more heterogeneous organs.

VSmTrans exhibits the most balanced cross-task profile. It remains consistently competitive across overlap and boundary metrics while frequently at-

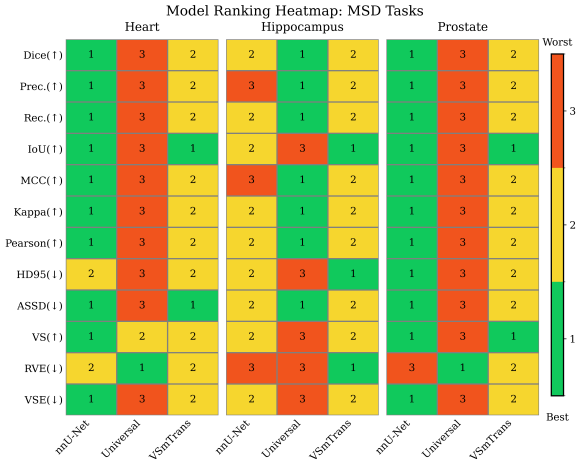

Figure 3: Color-coded models ranking.

taining the strongest volumetric fidelity, particularly on Heart. Rather than dominating a single task, it maintains a stable intermediate-to-high ranking pattern across anatomies, distinguishing it from the more polarized behavior of the other two models.

The Universal Model demonstrates the highest degree of task dependence. It ranks first on Hippocampus across overlap, boundary, and volumetric metrics, but ranks last across most metric families on Prostate and underperforms on Heart overlap measures. This pattern suggests that its vision–language conditioning is effective for small, morphologically constrained structures, but less robust for anatomies characterized by substantial heterogeneity, weak tissue contrast, or ambiguous zonal boundaries.

These results indicate that architectural ranking is anatomy-dependent rather than globally transferable. The observed performance differences are consistent with three plausible task-level factors: variation in morphological heterogeneity across organs, differences in effective foreground-to-background class imbalance introduced by preprocessing, and the degree of boundary ambiguity present in each segmentation target. Hippocampus tends to reward models that benefit from constrained structural context. By contrast, Heart and Prostate place greater demands on robustness to anatomical variability and uncertain boundary definition.

5 Discussion

5.1 Multi-Metric Evaluation as a Necessity

Single-metric evaluation fundamentally mischaracterizes model behavior. Two case studies illustrate this.

Case 1 (Prostate overlap): VSmTrans achieved Dice 0.724 ± 0.136 versus the Universal Model's 0.512 ± 0.228, yet both are superficially presented as "inadequate" under Dice-only reporting. The divergence reveals different decision strategies: VSmTrans maintains comparative voxel-level confidence; the Universal Model produces poorly calibrated outputs that do not track spatial accuracy. Practitioners requiring a single model deployable across structurally diverse targets, without task-specific retraining, should consider VSmTrans as the default all-rounder.

Case 2 (Heart volumetric bias): All three models achieved high Dice (0.894–0.939), implying comparable performance. However, the Universal Model exhibited systematic over-segmentation (RVE 0.054 ± 0.040) versus nnU-Net v2's near-zero bias (RVE 0.006 ± 0.053). A 5.4% systematic overestimate of chamber volume is clinically meaningful for ablation planning and cardiac function assessment despite apparently similar Dice scores.

5.2 Deployment Recommendations

- *Highest overlap and boundary accuracy on heterogeneous anatomy* (Heart, Prostate): nnU-Net v2, whose adaptive configuration yields the strongest overall performance and the most reliable surface-distance behavior.
- *Balanced versatile performance with strong volumetric fidelity:* VSmTrans, which remains consistently competitive across organs, frequently leads on volumetric measures, and avoids catastrophic anatomy-specific collapse. Practitioners requiring a single model deployable across structurally diverse targets, without task-specific retraining, may consider VSmTrans as the default all-rounder.
- *Small, standardized structures with constrained morphology* (e.g., Hippocampus): the Universal Model, with the caveat that it is not recommended for heterogeneous multi-zone anatomy without task-specific adaptation and explicit boundary-quality monitoring.

6 Conclusions

We benchmarked nnU-Net v2, VSmTrans, and the Universal Model across twelve metrics and three MSD tasks under unified conditions. This design addresses two compounding evaluation failures, metric inflation and metric incompleteness, that undermine model selection in biomedical data pipelines (Figure 4).

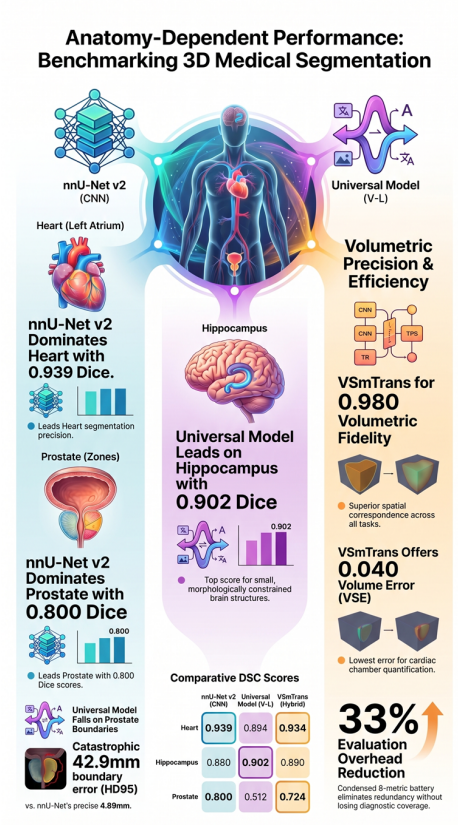

Figure 4: Visual summary of anatomy-dependent performance across three segmentation architectures.

Four principal findings emerge. *First*, standard evaluation batteries are substantially redundant: several commonly co-reported overlap and agreement metrics track near-identical case-level behavior on these binary tasks. An empirically validated compact eight-metric reporting subset preserves the principal conclusions of the full suite at 33% lower reporting overhead. *Second*, rankings are anatomy-dependent and non-transferable: nnU-Net v2 dominates on structurally variable organs, the Universal Model leads on morphologically constrained neuroanatomy, and VSmTrans emerges as the most balanced all-rounder with particularly strong volumetric fidelity. *Third*, failure modes of foundation-style models can be invisible to Dice-only evaluation: the Universal Model's severe Prostate boundary failure is masked by a superficially moderate overlap score, with direct consequences for digital-twin reliability. *Fourth*, architectural conclusions are robust for Heart and Prostate but warrant caution for Hippocampus because large pooled sample size can make numerically small differences appear disproportionately decisive.

Limitations

This work has at least four limitations. First, the compact eight-metric subset was derived empirically for this controlled 3D MRI benchmark. It is intended as a practical reporting subset for the tasks, metric definitions, and failure modes studied here rather than as a replacement for broader task-aware metric-selection frameworks.

Second, the evaluation covers only three MSD organs, and the resulting architectural patterns may not transfer directly to other anatomical targets or imaging modalities, such as liver, pancreas, lung nodules, CT, or ultrasound (Gao et al., 2025).

Third, the Heart and Prostate analyses rely on small hold-out cohorts (Heart $n = 10$, Prostate $n = 16$), which limits statistical power and increases sensitivity to individual failure cases despite Holm–Bonferroni correction. Conversely, the pooled five-fold Hippocampus evaluation provides substantially greater power, so statistically significant differences on that task should be interpreted together with their absolute magnitude and practical relevance.

Fourth, the observed performance gaps reflect the combined effects of architecture, optimization strategy, preprocessing, and training objective. This design improves ecological validity with respect to real-world use of each method, but it limits strictly causal attribution of performance differences to architecture alone.

Future work should therefore extend the benchmark along four directions. First, multi-center and multi-scanner validation is needed to test whether the observed anatomy-dependent rankings remain stable under realistic acquisition variability. Second, broader task coverage should include additional anatomies and modalities, particularly settings with severe class imbalance, topological complexity, or lesion-centric targets. Third, follow-up studies should investigate unified or partially unified training protocols, as well as multi-objective training strategies that jointly optimize overlap, boundary accuracy, and volumetric fidelity (Gong et al., 2021). Fourth, robust deployment-oriented evaluation should include uncertainty estimation through conformal prediction or deep ensembles for boundary outlier detection, class-imbalance-aware metrics for anatomies with pronounced size asymmetry such as PZ versus TZ, and improved prompt conditioning for the Universal Model on heterogeneous glandular anatomies (Gu et al., 2022).

References

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

Sandrine Bédard, Enamundram Naga Karthik, Charidimos Tsagkas, Emanuele Pravatà, Cristina Granziera, Andrew Smith, Kenneth Arnold Weber II, and Julien Cohen-Adad. 2025. Towards contrast-agnostic soft segmentation of the spinal cord. *Medical Image Analysis*, page 103473.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. 2023. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *Advances in Neural Information Processing Systems*, 36:51758–51777.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9901 of *Lecture Notes in Computer Science*, pages 424–432.

T. Debnath and M. Narbutt. 2025. The battle of the titans: Comparing three leading DL methods for 3D medical imaging. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 691–700, Washington, DC, USA.

Tanmoy Debnath, Shayne Chau, and Md Geaur Rahman. 2026. Brain mri segmentation. In *Artificial Intelligence and Data Analytics in Medical Imaging*, pages 94–129. CRC Press.

Szuzina Fazekas, Bettina K Budai, Viktor Bérczi, Pál Maurovich Horvat, and Zsolt Vizi. 2026. A clinically informed automated evaluation pipeline for medical image segmentation based on medical similarity index. *Physics and Imaging in Radiation Oncology*, 38:100950.

Weilun Feng, Mingqiang Wu, Zhiliang Chen, Chuanguang Yang, Haotong Qin, Yuqi Li, Xiaokun Liu, Guoxin Fan, Zhulin An, Libo Huang, Yulun Zhang, Michele Magno, and Yongjun Xu. 2026. Fast-sam3d: 3dfy anything in images but faster. *Preprint*, arXiv:2602.05293.

Yuxiao Gao, Yang Jiang, Yanhong Peng, Fujiang Yuan, Xinyue Zhang, and Jianfeng Wang. 2025. Medical image segmentation: A comprehensive review of deep learning-based methods. *Tomography*, 11(5):52.

Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengxin Dai, and Luc Van Gool. 2021. Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8344–8354.

Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. 2022. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer.

Weifan Guan, Qinghao Hu, Aosheng Li, and Jian Cheng. 2025. Efficient vision-language-action models for embodied manipulation: A systematic survey. *arXiv preprint arXiv:2510.17111*.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. 2022. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 574–584.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. 2021. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211.

Burak Kocak, Michail E. Klontzas, Arnaldo Stanzione, Aymen Meddeb, Aydin Demircioğlu, Christian Bluethgen, Keno K. Bressem, Lorenzo Ugga, Nathaniel Mercaldo, Oliver Díaz, Renato Cuocolo, and 1 others. 2025. Evaluation metrics in medical imaging AI: Fundamentals, pitfalls, misapplications, and recommendations. *European Journal of Radiology Artificial Intelligence*, 3:100030.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. 2025. Medclip-samv2: Towards universal text-driven medical image segmentation. *Medical Image Analysis*, 106:103749.

M.-X. Li, W.-H. Deng, Z.-X. Wu, C.-X. Jin, J.-M. Wu, Y. Han, J. K. H. Tsoi, G.-S. Xia, and C. Huang. 2026. Metadent: Labeling clinical images for vision-language models in dentistry. *Journal of Dental Research*.

Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. 2023. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289.

Y. Lian, Y. Xie, Y. Jiang, and 1 others. 2026. A data-efficient 3D medical vision-language model using only a 2D encoder. *Scientific Reports*, 16:8809.

Tiange Liu, Qingze Bai, Drew A. Torigian, Yubing Tong, and Jayaram K. Udupa. 2024. VSmTrans: A hybrid paradigm integrating self-attention and convolution for 3D medical image segmentation. *Medical Image Analysis*, 98:103295.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, and 1 others. 2024. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212.

Pooya Mohammadi Kazaj, Giovanni Baj, Yazdan Salimi, Anselm W. Stark, Waldo Valenzuela, George C. M. Siontis, Habib Zaidi, Mauricio Reyes, Christoph Graeni, and Isaac Shiri. 2025. From claims to evidence: A unified framework and critical analysis of CNN vs. transformer vs. mamba in medical image segmentation. *arXiv preprint arXiv:2503.01306*.

Y. Pusparani, C.-Y. Lin, Y.-K. Jan, and 1 others. 2026. Evaluation of deep learning models for segmentation of hippocampus volumes from MRI images in Alzheimer's disease. *Scientific Reports*, 16:7878.

J. Qin and Y. Yang. 2026. ProSeg: Multi-scale context fusion for high-precision prostate segmentation in MRI. *Scientific Reports*, 16:13620.

Sucheng Ren and Xiaomeng Li. 2025. HResFormer: Hybrid residual transformer for volumetric medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10558–10566.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopoulos, Tobias Norajitra, and Klaus Maier-Hein. 2025. Voxtell: Free-text promptable universal 3d medical image segmentation. *arXiv preprint arXiv:2511.11450*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241.

Fenghe Tang, Chengqi Dong, Wenxin Ma, Zikang Xu, Heqin Zhu, Zihang Jiang, Rongsheng Wang, Yuhao Wang, Chenxu Wu, and Shaohua Kevin Zhou. 2025. U-Bench: A comprehensive understanding of U-Net through 100-variant benchmarking. *arXiv preprint arXiv:2510.07041*.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 2022. Self-supervised pre-training of Swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740.

Yichi Zhang, Zhenrong Shen, Lanlan Li, Wenbo Zhang, and Le Xue. 2026. Unleashing the potential of generalist segmentation foundation models for biomedical image and video analysis. *iRADIOLOGY*, 4(2):127–136.