

Supplementary Table S1. Size-matched patient-exclusive multisequence U-Net sensitivity analysis.

Target	Size-matched patient-exclusive Dice (95% CI)	Sequence-specific U-Net	Cross-sequence U-Net	Full pooled U-Net	Difference vs cross-sequence (95% CI)
BTFE	0.905 (0.890–0.919)	0.908	0.870	0.914	+0.0352 (0.0218–0.0527)
ssh-TSE	0.880 (0.865–0.895)	0.888	0.822	0.894	+0.0581 (0.0413–0.0762)

Supplementary Table S2. Absolute relative volume difference across sequence-specific, cross-sequence, and pooled multisequence conditions.

Target se- quence	Pipeline	Sequence-specific abso- lute RVD, % (95% CI)	Cross-sequence absolute RVD, % (95% CI)	Pooled absolute RVD, % (95% CI)
BTFE	U-Net	5.1 (3.0–7.5)	9.8 (6.1–14.4)	4.7 (2.8–7.0)
BTFE	nnU-Net	5.0 (2.8–7.7)	4.2 (2.7–5.7)	4.3 (2.5–6.9)
BTFE	U-Mamba Bot	5.0 (2.6–8.1)	6.5 (4.0–9.1)	5.4 (3.0–8.1)
ssh-TSE	U-Net	4.8 (3.0–6.3)	9.1 (5.9–12.5)	4.6 (3.0–6.3)
ssh-TSE	nnU-Net	5.4 (3.8–6.9)	17.5 (10.7–25.5)	4.8 (3.1–6.7)
ssh-TSE	U-Mamba Bot	4.1 (2.5–6.2)	19.2 (12.4–26.9)	4.8 (3.1–6.5)

Table note: Values are mean patient-level absolute relative volume difference (RVD) with 95% bootstrap confidence intervals. Absolute RVD quantifies the magnitude of the relative difference in foreground pixel counts between predicted and reference segmentations, irrespective of over- or under-segmentation. Because physical voxel spacing and slice geometry were unavailable, RVD is interpreted as a pixel-count-based measure of relative segmentation size error rather than actual placental volume.

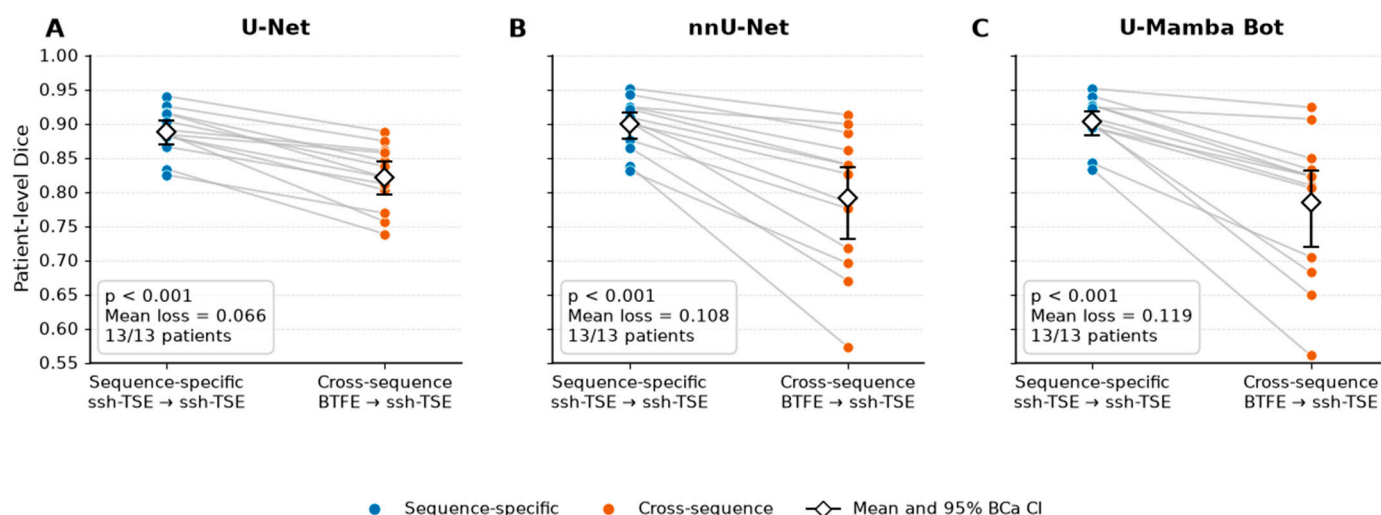
Supplementary Table S3. Additional slice-level segmentation performance and boundary-error analysis across sequence-specific, cross-sequence, and pooled multisequence conditions.

Model	Evaluation setting	Pa-tients	Slices	Slice Dice, mean $\pm$ SD	Slice Dice, median [IQR]	HD95, pixels, median [IQR]	Slices with Dice below 0.75, %	Empty predic-tions, n (%)
nnU-Net	BTFE BTFE	→ 13	345	0.906 $\pm$ 0.070	0.927 [0.882, 0.950]	5.39 [3.00, 13.96]	3.2	0 (0.0)
nnU-Net	BTFE ssh-TSE	→ 13	371	0.774 $\pm$ 0.166	0.826 [0.702, 0.893]	19.30 [8.54, 38.40]	34.2	1 (0.3)
nnU-Net	ssh-TSE ssh-TSE	→ 13	371	0.889 $\pm$ 0.076	0.911 [0.861, 0.942]	6.02 [3.62, 15.48]	5.1	0 (0.0)
nnU-Net	ssh-TSE BTFE	→ 13	345	0.846 $\pm$ 0.142	0.890 [0.815, 0.927]	13.60 [5.87, 31.30]	11.9	0 (0.0)
nnU-Net	Pooled BTFE	→ 13	345	0.905 $\pm$ 0.071	0.925 [0.884, 0.949]	5.45 [3.00, 13.94]	3.5	0 (0.0)
nnU-Net	Pooled ssh-TSE	→ 13	371	0.892 $\pm$ 0.070	0.917 [0.862, 0.944]	6.36 [3.61, 13.35]	4.3	0 (0.0)
U-Mamba Bot	BTFE BTFE	→ 13	345	0.907 $\pm$ 0.071	0.927 [0.886, 0.949]	5.08 [3.00, 11.78]	3.2	0 (0.0)
U-Mamba Bot	BTFE ssh-TSE	→ 13	371	0.769 $\pm$ 0.180	0.820 [0.703, 0.894]	17.03 [7.89, 34.40]	31.8	2 (0.5)
U-Mamba Bot	ssh-TSE ssh-TSE	→ 13	371	0.894 $\pm$ 0.070	0.919 [0.864, 0.944]	6.08 [3.61, 12.79]	4.0	0 (0.0)
U-Mamba Bot	ssh-TSE BTFE	→ 13	345	0.845 $\pm$ 0.164	0.893 [0.833, 0.928]	10.21 [5.71, 24.11]	11.9	3 (0.9)
U-Mamba Bot	Pooled BTFE	→ 13	345	0.908 $\pm$ 0.067	0.926 [0.887, 0.951]	5.02 [3.00, 12.20]	2.9	0 (0.0)
U-Mamba Bot	Pooled ssh-TSE	→ 13	371	0.895 $\pm$ 0.075	0.917 [0.863, 0.943]	5.83 [3.57, 11.27]	3.2	0 (0.0)
U-Net	BTFE BTFE	→ 13	345	0.894 $\pm$ 0.099	0.922 [0.879, 0.945]	6.40 [3.49, 16.86]	4.6	1 (0.3)
U-Net	BTFE ssh-TSE	→ 13	371	0.802 $\pm$ 0.140	0.842 [0.753, 0.886]	19.35 [9.00, 39.33]	24.0	2 (0.5)
U-Net	ssh-TSE ssh-TSE	→ 13	371	0.876 $\pm$ 0.082	0.901 [0.839, 0.928]	7.98 [4.47, 15.82]	6.5	0 (0.0)
U-Net	ssh-TSE BTFE	→ 13	345	0.844 $\pm$ 0.143	0.881 [0.825, 0.920]	11.36 [6.32, 23.26]	11.3	5 (1.4)

U-Net	Pooled BTFE	→ 13	345	0.902 ± 0.073	0.922 [0.882, 0.946]	5.76 [3.61, 11.68] 3.8	0 (0.0)
U-Net	Pooled ssh-TSE	→ 13	371	0.884 ± 0.077	0.904 [0.856, 0.936]	7.03 [4.00, 17.94] 5.4	0 (0.0)

Table note: HD95, 95th-percentile Hausdorff distance; IQR, interquartile range; SD, standard deviation. Slice-level Dice and HD95 were calculated on a common 256×256 evaluation grid for all three segmentation pipelines. U-Net predictions were generated directly on this grid, whereas nnU-Net and U-Mamba Bot foreground probability maps were resized to 256×256 using bilinear interpolation before thresholding; reference masks were resized to the same grid using nearest-neighbor interpolation. HD95 is reported in pixels because physical pixel spacing was unavailable. HD95 values are summarized as median [IQR] among finite measurements only. No evaluated test slice had an empty reference placental mask; therefore, all non-finite HD95 values resulted from empty predictions for non-empty reference masks and represent complete segmentation misses. These failures are reported separately in the “Empty predictions” column and are not included in the finite HD95 summary. Slice-level Dice describes performance for individual test images and is complementary to the patient-level aggregated Dice used for the primary analysis. For pooled multisequence evaluation, results are reported separately for BTFE (345 slices) and ssh-TSE (371 slices), comprising the same 13 held-out patients.

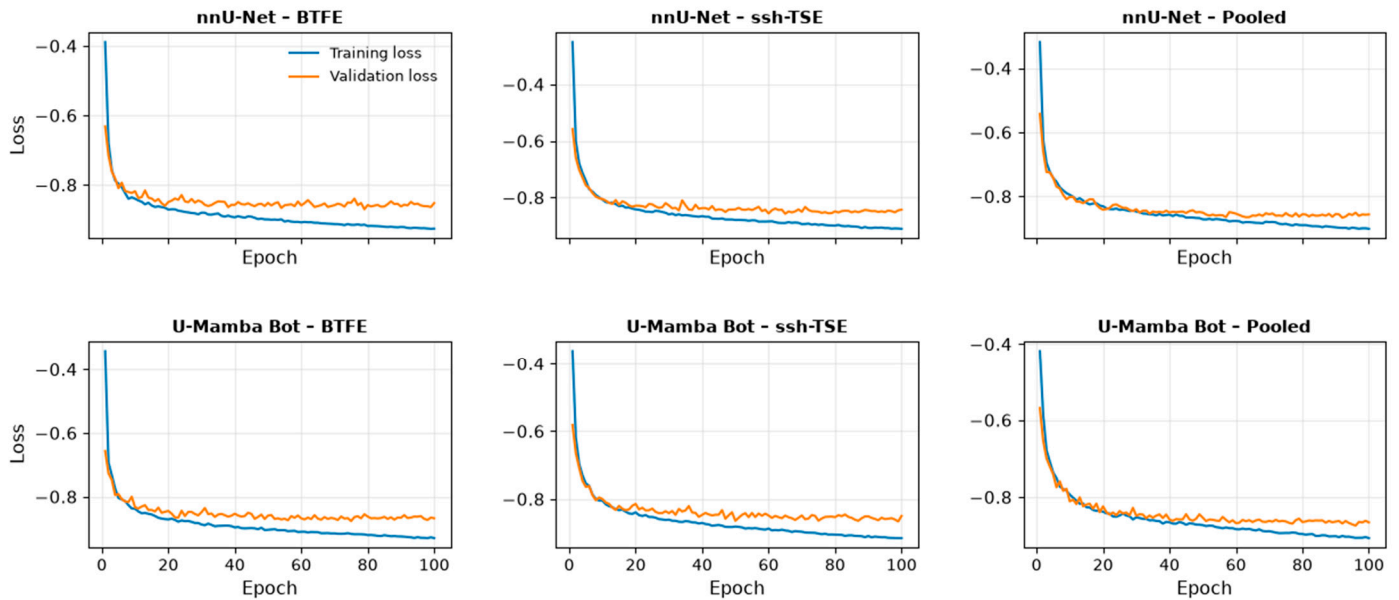
Supplementary Figure S1. Patient-level paired changes in Dice under cross-sequence transfer.



Supplementary Figure S1. Patient-level paired Dice performance under BTFE→ssh-TSE cross-sequence transfer for (A) U-Net, (B) nnU-Net, and (C) U-Mamba Bot. Each connected pair represents the same held-out patient across the two evaluation conditions. Colored circles indicate individual patient values, while white diamonds with whiskers show the mean and 95% bias-corrected and accelerated (BCa) bootstrap confidence interval. The inset boxes report the paired Wilcoxon signed-rank p-value, mean Dice loss, and number of patients showing degradation. Cross-sequence transfer reduced Dice in all 13 patients for each pipeline.

Supplementary Figure S2. Training and validation loss trajectories for nnU-Net and U-Mamba Bot.

Training and validation loss across 100 epochs



Supplementary Figure S2. Training and validation loss trajectories for nnU-Net and U-Mamba Bot. Training and validation loss are shown across the 100-epoch training schedule for BTFE-only, ssh-TSE-only, and pooled multisequence models. Validation loss decreased rapidly during early training and showed only small changes during the final portion of training across all six configurations. Inference used the final checkpoint after epoch 100 for each model. Negative loss values reflect the Dice-based loss-function conventions used by the training frameworks; validation loss generally plateaued toward the end of training, although training loss continued to decrease modestly in some configurations.