Supplementary Material Mitral-Valve Segmentation Across CT, TEE, and Surgical Video

Harshit Agrawal[1]

AIATELLA Oy., Helsinki, Finland harshit@aiatella.com

S1 System overview

S2 Training protocols and pseudo-label gating

Task 1 pseudo-label acceptance criteria. A candidate pseudo-label is accepted only if all of the following hold. Every bound is read off the 27 labeled cases rather than tuned. (i) The mask is non-empty. (ii) Its largest connected component holds $\geq$90% of the foreground – all 27 real labels are single-component. (iii) Foreground volume lies in $[1500, 5000]$ mm^3 (measured labeled range 1854–4639). (iv) Foreground fraction lies in $[0.005, 0.04]$. (v) The source volume's finest-axis spacing is $<$0.55 mm, which discards a $\sim$9% coarse out-of-distribution tail. **921 of 1,040 volumes pass (88.6%)**; accepted masks are largest-component-cleaned before use.

S3 Measured inference cost

On the evaluation V100 the deployed image processes the whole hidden set in 2,151 s of a 21,600 s allowance; the earliest single-member image took 2,010 s and the most expensive intermediate one 2,593 s. Per-case figures measured on our own GPU for the single-member container are 1.01 s (CT), 4.06 s (TEE; 7.52 s with the enlarged patch) and 0.43 s per frame (video), with peak GPU memory of 1.05 GiB per video member; the five-member CT family costs 3.14$\times$ the single model (26.2 $\rightarrow$ 82.3 s for 30 cases). Container-versus-host mask parity was mean Dice $\geq$0.9998 on all tasks. Only the end-to-end runtimes are from the target hardware, since our development GPU cannot execute the container's CUDA build.

S4 Empty-frame behaviour on Task 3

Because the hidden score averages over frames with no valve, a false positive on an empty frame costs far more than a sloppy contour. We tracked empty-frame false positives explicitly on the six-video internal split: the deployed fusion weight reduced them from 3 to 1, and any candidate that increased them was vetoed regardless of its Dice.

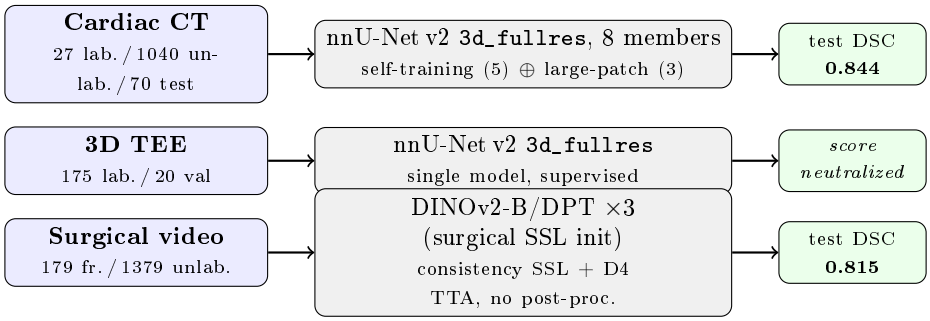

Fig. S1. The deployed system (hidden-test submission 873932); DSC values are hidden-test scores. Task 2 is excluded from the competition ranking.

Table S1. Training protocols for the deployed members. nnU-Net auto-configures normalization and augmentation for the 3D tasks; the video settings are chosen explicitly. The 3D tasks use nnU-Net defaults (SGD, Nesterov momentum 0.99, polynomial decay). Patch sizes marked † are non-default overrides.

	Task 1 (CT)	Task 2 (TEE)	Task 3 (Video)
Backbone	nnU-Net v2 `3d_fullres`	same	DINOv2 ViT-B/14 + DPT
Encoder init	from scratch	from scratch	SurgeNetXL (surgical SSL)
Members deployed	8 (5 self-train + 3 large-patch)	1	3
Input / patch	auto; 112×160×192† (family B)	128×192×224†	448 × 800
Epochs / optimizer	250, SGD 1e−2	same	70, AdamW
Loss	region + CE	region + CE	region + CE, consistency
SSL	offline self-training (921 pseudo)	none	EMA teacher, 1,379 unlabeled
Inference	mirror TTA, fuse, largest CC	mirror TTA off, largest CC	D4 TTA, thr. 0.45, no post-proc.

An explicit valve-presence gate – zeroing a mask whenever any member predicts empty – was built and rejected. On clean data it removed 71–91% of empty-frame false positives with zero false suppressions, but under a fixed panel of photometric and blur corruptions it silently zeroed **27 frames that did contain a valve** (white balance at the strongest level 13, sensor noise at the *weakest* 8), including frames from otherwise healthy videos. One false suppression scores zero Dice on a real valve, and the failure is invisible on in-distribution evaluation – which is why a clean-data check had rated the gate promising. The deployed system therefore contains no presence heuristic.

S5 Interventions that failed their gates

Beyond the reverted rows in Table 2 of the main paper, the following were built, measured against a pre-registered gate, and not deployed (Table S2).

Table S2. Shelved interventions with effect sizes. Internal surfaces are leave-one-video-out (video) or leakage-free out-of-fold / held-out (CT, TEE).

Intervention	Task	Internal effect	Reason not deployed
Valve-presence gate	T3	FP -71 to -91% clean	27 real frames zeroed under shift
Small-DINOv2 + UniMatch replacement	T3	-0.0560 DSC	decisive regression
Bidirectional copy-paste	T3	-0.0209 DSC	sign-consistent regression, both folds
Boundary distillation	T3	-0.0100 DSC, one fold -0.085	shrinks under-segmented anatomy
Strong RandAugment (unlabeled branch)	T3	-0.0100 composite at TTA	breaks teacher–student consistency
CaFormer third member	T3	-0.0194 DSC, 1/6 folds	under-fits on $\sim$149 frames
Test-time batch-norm adaptation	T3	helps noise axis only	fails empty-frame FP veto
clDice boundary loss	T2	$+0.0012$ DSC	inside ± 0.008 noise floor
Foreign-anatomy encoder init	T1	0.8603 vs 0.8614 board	below deployed model