Label-Limited Mitral Valve Segmentation Across CT, TEE, and Surgical Video: MVAA 2026

Jay Kshirsagar[1] and Pascal Theriault Lauzier[2]

[1] University of Ottawa, Ottawa, ON, Canada
jkshi064@uottawa.ca
[2] University of Ottawa Heart Institute, Ottawa, ON, Canada

Abstract. Mitral valve repair is planned on cardiac CT, guided by 3D transesophageal echocardiography (TEE) and performed under a surgical camera, so an automated anatomy pipeline must segment one structure in three modalities that do not share appearance statistics. The MVAA 2026 challenge releases 27 labeled CT volumes, 105 labeled TEE volumes, and 180 labeled video frames, a regime in which the binding constraint is the label count rather than the model capacity. We build one pipeline per task around that constraint: semi-supervised pseudo-labeling under a shape gate for CT, a fully-supervised larger-backbone ensemble for TEE, and a leave-one-video-out ensemble with leaflet-preserving post-processing for video. On the validation leaderboard, these reach Dice scores of 0.861, 0.853, and 0.811; on the official hidden test set, they reach 0.839 on CT and 0.789 on video. The video score improved once we removed a post-processing rule that emitted an empty mask, a rule that is optimal under the validation protocol and maximally penalized under the test protocol. We separately disclose that our 0.966 hidden test dice on TEE is not a valid measurement: the external cases we added to that model overlap the pool from which the hidden test set was drawn.

Keywords: Mitral valve segmentation · Multimodal medical imaging · Semi-supervised learning · nnU-Net · Test-set contamination.

1 Introduction

Mitral regurgitation is among the most common valvular heart diseases, and a repair is planned, guided, and assessed on three different image streams: pre-procedural cardiac CT, intraoperative 3D TEE, and the intraoperative surgical camera. Automating valve delineation in all three is the target of the MVAA 2026 challenge [1] (Mitral Valve Anatomy Analysis, MICCAI 2026 Medical World Models workshop), which scores each modality as a separate task under Dice similarity coefficient (DSC), Hausdorff distance (HD), and average surface distance (ASD), and ranks only teams that submit all three.

The dominant recipe for medical segmentation is to take a self-configuring encoder-decoder such as nnU-Net [2] and scale it, or to add task-specific structure such as coarse-to-fine refinement [13]. That recipe assumes enough labeled

cases to fit the added parameters, and cardiac benchmarks that supply them exist: large annotated collections are available for coronary CT angiography [10] and for right atrial cavity segmentation from LGE-MRI [11, 12], as are curated benchmarks for surgical video understanding [14]. MVAA supplies 27, 105, and 180 labeled cases per task, so the assumption fails on at least two of the three tasks, and the question the challenge actually poses is not which architecture is strongest but how much supervision can be recovered from limited labels.

We treat the label count as the single governing constraint and design each pipeline around it. Mainly, this paper contributes:

1. A shape-gated pseudo-labeling loop for CT (Section 3.1) whose accept/reject thresholds are read off the 27 ground-truth masks rather than off model confidence, which is uninformative when the background occupies over 99% of the volume.
2. Evidence that backbone scaling pays only where labels support it: the same ResEnc-L configuration that hurts on 27 CT cases is our largest gain on 105 TEE cases (Sections 3.1, 3.2).
3. A never-empty inference cascade for video (Section 3.3) that removes an empty-mask failure mode accounting for most of a 301.2 px reduction in Hausdorff distance on the hidden test set.

2 Tasks & Label Budget

Table 1 lists what each task provides. We follow two objectives. First, the labeled sets are small and unevenly sized, ranging from 27 to 180 cases, so a configuration that overfits one task may be correctly specified on another. Second, the unlabeled pools are large but present for only two tasks, and only Task 1's pool is used in this work; Task 3's 1,379 unlabeled frames from 46 videos remain unused and are the subject of Section 7.

Each task also carries a structural differences that the pipeline must absorb. In Task 1 the valve occupies roughly 2% of the CT volume, and all 27 ground-truth masks form a single connected 3D component. In Task 2 the two leaflets are separate foreground classes of comparable size. In Task 3, 62 of the 180 labeled frames (34%) contain no valve, so an empty mask is a legitimate training target even though, as Section 3.3 shows, it is never a legitimate test-time output.

3 Three Label-Efficient Pipelines

We build one model per task rather than a shared multimodal backbone, which the challenge rules permit. The modalities differ in dimensionality, voxel scale, and noise process, and with 27 to 180 labels per task there is no budget to pay for a shared representation that must then be specialized three ways. All three pipelines ensemble by averaging fold predictions; what differs between them is where the additional supervision comes from, which the three subsections below take in turn.

Table 1. Data released per task. Counts are cases (CT, TEE) or frames (video). Task 3's labeled count includes 62 frames with no valve in view and one frame excluded as mislabeled, leaving 179 frames used for training of which 118 are valve-bearing. The unlabeled column is the pool released for semi-supervised use, not the pool this work consumed: Task 3's pool is unused.

Task	Modality	Dim.	Foreground labels	Train (lab./unlab.)	Val
1	CT	3D	1 class (valve)	27 / 1,040	30
2	TEE	3D	2 classes (leaflets)	105 / –	20
3	Video	2D	1 class (valve)	180 / 1,379	48

3.1 Shape-Gated Pseudo-Labeling for Cardiac CT (Task 1)

The Task 1 model is a 3D full-resolution nnU-Net in its default configuration, trained as a 5-fold cross-validation ensemble.

We first train that ensemble on the 27 labeled scans alone and predict a mask for each of the 1,040 unlabeled scans. Admitting a predicted mask as a pseudo-label requires passing three checks, each threshold derived from the 27 ground-truth masks rather than from model output. The volume check keeps a case whose largest 3D connected component (26-connectivity) has physical volume in $[1{,}298, 6{,}030]\,\mathrm{mm}^3$, built as $0.7\times$ the smallest and $1.3\times$ the largest ground-truth valve volume (observed range 1,854 to $4{,}639\,\mathrm{mm}^3$). The dominance check requires that largest component to hold at least 85% of all predicted foreground voxels, rejecting fragmented output. The third check rejects empty predictions. An admitted case is reduced to its largest component before use. The gate admits 947 of 1,040 cases, rejecting 3 as too small, 82 as too large, 8 as too fragmented, and 0 as empty. That 82 of the 93 rejections are oversized identifies what the gate actually filters: over-segmentation leaking into surrounding tissue, not low-confidence output. We then retrain the 5-fold ensemble from scratch on 27 ground-truth plus 947 pseudo-labeled scans, adding pseudo-labels to the training split of every fold while restricting each fold's validation split to ground-truth cases only.

The volume check does almost all of the filtering: 85 of the 93 rejections are volume failures and only 8 are dominance failures. We did not train an ungated model, so the gate's contribution to the final score is not isolated here; what the rejection tally does establish is what the gate removes, which is over-segmentation rather than low-confidence output.

Reducing to the largest component is admissible for CT specifically because the annulus and both leaflets form one connected volume in all 27 ground-truth masks. Section 3.3 shows the same operation is destructive in 2D video, where the leaflets can be genuinely disjoint in-plane.

Confidence, the usual pseudo-label criterion, is unusable on this task. Because background occupies over 99% of each CT volume, the teacher ensemble's mean predicted confidence is 0.999, so a confidence gate removes boundary supervision rather than bad labels. The same statistic explains why a mean-teacher consis-

tency loss [6] did not train: the teacher is already saturated almost everywhere, leaving the unsupervised term with a near-zero gradient.

3.2 Backbone Scaling and an External-Data Error (Task 2, TEE)

The Task 2 model is a fully-supervised 3D full-resolution nnU-Net with the ResEnc-L backbone, trained for 1,000 epochs as a 5-fold ensemble on the 105 released cases, with mirror test-time augmentation disabled so that the full ensemble meets the validation phase's 10 s per-case budget.

ResEnc-L differs from the default configuration in the encoder alone: each plain convolutional stage is replaced by a stack of residual blocks, and nnU-Net's GPU-memory target is raised, which its planner spends on a deeper encoder and a larger patch size. The decoder, loss, and data pipeline are unchanged. Task 2 is the one setting in this challenge where the larger encoder pays: the identical substitution degrades Task 1's 27-case model (Section 6), and the labeled-set size is the only difference between the two settings. We selected this configuration on validation score, but the encoder and the 1,000-epoch schedule were changed together and we did not run the matched-schedule comparison, so their contributions are not separated here.

For the final test phase we added 50 cases from the public MVSeg2023 dataset [9], which the challenge's data-use rules permit, reducing cross-validation from 5 folds to 3 to keep training cost manageable on the larger 155-case set. This was a mistake. As Section 5 sets out, those 50 cases are drawn from the same pool as the hidden test set, so the resulting model was trained on data overlapping the data it was scored on, and its hidden-test score measures memorization rather than generalization. The uncontaminated Task 2 model is the one trained on the 105 released cases alone.

3.3 Presence Estimation and the Never-Empty Fallback (Task 3, Video)

The Task 3 model is a UNet++ [4], a nested variant of U-Net [3], with an ImageNet-pretrained [5] encoder, trained on 179 labeled frames with a combined Dice [8] and Focal [7] loss. A presence head, a global average pool over the deepest encoder features followed by a linear layer to one logit trained with its own binary cross-entropy term, predicts whether a frame contains the valve at all. Folds are split so that frames from one source video never appear in both the training and validation split of a fold, a leave-one-video-out protocol over 6 folds. At inference the segmentation probability maps are averaged across folds and the presence probabilities are averaged separately.

Patient-level diversity bounds what cross-validation can measure. The 179 labeled frames come from only six distinct patients. A random frame-level split would put frames from one patient on both sides of the split, and because frames from one patient share illumination, camera pose, instrument set, and valve morphology, such a split reports interpolation within a patient rather than generalization to a new one. Leave-one-video-out removes that leakage,

which is why we use it, but it cannot remove the underlying limit: each fold's held-out set is essentially one patient, so the fold-to-fold variance is large and the pooled mean is an average over very few effective samples.

The presence gate is wrong at test time. Our validation-leaderboard submission thresholded the averaged presence probability at 0.5 and emitted an empty mask below it, on the reasoning that an empty prediction against an empty reference scores a perfect Dice while a false positive scores zero. That reasoning does not transfer: the challenge specifies that every frame scored in the test phase contains the valve, so the empty reference never occurs, and an empty prediction is instead charged the maximum-distance penalty on both HD and ASD. We therefore demote the presence head to a diagnostic output, retained behind a flag for ablation, and guarantee a non-empty mask through a three-level cascade in which each level fires only when the one above it returns empty: (i) threshold the averaged map and keep every connected component of at least 200 px; (ii) if that is empty, take the largest region of a lower peak-relative threshold; (iii) if that is also empty or implausibly large, indicating a near-flat probability map, emit a size-bounded blob centered on the single highest-probability pixel. We additionally moved error handling from per-video to per-frame scope, so a single failed frame writes a bounded fallback rather than discarding every remaining frame in that video.

Component filtering must preserve the second leaflet. On the averaged mask we run 2D connected-component labeling and keep every component of at least 200 px, rather than the conventional largest-component-only rule. The mitral valve presents as two leaflets that appear as two disjoint regions in many 2D views, and a largest-only rule deletes the smaller leaflet whenever the two do not touch in-plane. We compared no filtering, a threshold relative to the largest component, and absolute cutoffs of 200 and 400 px on the pooled held-out frames; all four fall within 0.001 DSC of each other, so the aggregate metric does not select among them. The choice is therefore made on the failure mode rather than on the score, and we take the most conservative rule that cannot remove a real leaflet.

The presence head plays two roles, and only one of them survives. As an output rule it is harmful at test time, for the reason given above. As an auxiliary training signal it may still help, by forcing the encoder to represent valve absence rather than assuming every frame contains a valve, and that effect is independent of whether its output is ever consulted at inference. We did not retrain without the auxiliary term, so the head's value as a regularizer is untested; what the hidden-test results in Section 4 do establish is the cost of its second role, as an output gate.

Augmentation targeted at the worst videos. We added three augmentations chosen by inspecting which validation videos the model handled worst: wider color and contrast jitter, random spatial scale jitter, and a mask-aware synthetic occluder covering part of the labeled valve region during training. Table 2 reports the effect of retraining the identical 6-fold protocol with them.

Table 2. Task 3 augmentation ablation, pooled across the 6 leave-one-video-out held-out splits. Both rows use the identical training protocol and differ only in the augmentation set. HD and ASD in px.

Configuration	DSC	HD (px)	ASD (px)
6-fold baseline	0.844	98.7	37.3
+ all three augmentations	**0.855**	**85.4**	**28.4**

The pooled figures understate the effect on the videos the augmentations targeted: the single worst validation video improved from 0.750 to 0.804 DSC with its Hausdorff distance roughly halved, and the colour and contrast jitter contributed most of the gain while the synthetic occluder changed little on its own.

Finally, we treat each frame as an independent 2D image. Labeled and unlabeled frames are sparse and non-consecutive, with temporal gaps of up to 15 frames, so no adjacent frame exists to support a temporal or optical-flow model.

4 Results

All models were trained on an HPC cluster with NVIDIA A40 (45 GB) GPUs. Tasks 1 and 2 use nnU-Net v2 with its automatic preprocessing and configuration; Task 3 uses `segmentation_models_pytorch`. All three tasks ensemble by averaging fold predictions.

Table 3 reports our validation-leaderboard scores and the organizers' hidden-test evaluation of our submitted Docker containers side by side. Task 1 transfers with a modest drop, from 0.861 to 0.839 DSC, and its surface metrics degrade in proportion (HD 4.60 to 6.59 mm, ASD 0.272 to 0.576 mm), which is the expected behaviour of a model that generalizes.

Task 3 is where the two phases diverge, and the cause is post-processing rather than modeling. The submission carrying the presence gate scored 0.759 DSC with 518.1 px HD and 384.1 px ASD; removing the gate moved those to 0.789, 216.9 px, and 55.2 px. A 301.2 px drop in HD against a DSC change of only +0.030 is the signature of removing maximum-distance penalties, not of better delineation, and it matches the empty-mask mechanism identified in Section 3.3. One attribution limit applies: this submission changed the presence gate and the augmentation set together, and Table 2 shows augmentation alone moves pooled HD by 13.3 px, or 4% of the observed change. The presence gate is the only mechanism that can produce the remainder, but we did not submit the two changes separately and cannot close the attribution from the hidden test set alone. Note also that Task 3's validation HD of 111.9 px is high relative to its 0.811 DSC, which localizes the residual error: the model omits an entire valve region on some frames, an error Dice largely forgives and the surface metrics do not.

Table 3. Our validation-leaderboard and official hidden-test results. Validation splits are $n = 30$ (CT), $n = 20$ (TEE), $n = 48$ (video), with no missing cases. HD and ASD are in mm for Tasks 1–2 and px for Task 3, so magnitudes are not comparable across tasks. Higher DSC and lower HD/ASD are better. The italicised Task 2 row is invalidated by Section 5 and is not a performance claim.

Task	Evaluation	DSC	HD	ASD
1: CT	Validation	0.861	4.60	0.272
	Hidden test	0.839	6.59	0.576
2: TEE	Validation	0.853	10.95	0.613
	Hidden test, released data only	0.853	9.88	0.526
	Hidden test, + external data	*0.966*	*2.14*	*0.081*
3: Video	Validation	0.811	111.9	15.69
	Hidden test, presence gate active	0.759	518.1	384.1
	Hidden test, gate removed + aug.	**0.789**	**216.9**	**55.2**

Task 2 scores 0.853 DSC on both the validation leaderboard and the hidden test set when trained on the released data alone. The 0.966 row is reported for completeness only; Section 5 explains why it is not a measurement of generalization.

5 Data Provenance: The Task 2 Hidden-Test Overlap

After receiving the 0.966 score we traced the provenance of MVAA's own Task 2 data and found that it is MVSeg2023. The Task 2 training and validation cases MVAA released to participants are byte-identical to MVSeg2023's own training split and to part of its validation split; the 50 cases we added for the final submission are exactly the MVSeg2023 cases MVAA never released; and those withheld cases are the only pool from which MVAA's hidden test cases could have been drawn. Hence, we treat the 105-case results in Table 3 as the only supportable estimate of Task 2 performance, and we report the contaminated number.

6 Interventions That Did Not Help

Several changes we expected to help did not, and we record them because the pattern is informative. On Task 1, neither capacity nor training length helped: the ResEnc-L backbone and a 1,000-epoch schedule both scored below the 250-epoch default configuration on 27 labeled cases, and confidence-gated pseudo-labeling with loss re-weighting failed for the reason given in Section 3.1, that a saturated teacher makes confidence uninformative. On Task 3, a transformer-based encoder with a boundary-distance loss, a recall-biased Tversky loss, and an auxiliary head predicting valve depth and anatomy all improved our internal

cross-validation while failing to improve, and in one case reducing, the leaderboard score. All three failures share one signature, that offline cross-validation moved and the board did not, which is the patient-count limit analysed in Section 3.3 rather than three independent modeling accidents. Taken together with the Task 2 result, where the larger encoder did help at 105 labels (Section 3.2), these outcomes support treating the label count rather than the architecture as the design variable.

7 Limitations

Task 3's 1,379 unlabeled frames were never used. The shape gate of Section 3.1 does not port to them, because it assumes a singly connected target and the valve is a two-blob target in 2D; a cross-fold agreement gate is the replacement this requires.

The three per-task pipelines share no parameters, so nothing here tests whether a shared representation would help; the design assumes it would not pay at these label counts, and that assumption is untested.

8 Conclusion

Across three modalities and 312 total labeled cases, the interventions that moved MVAA 2026 scores were those that changed how much supervision each label delivers, or that removed a metric failure mode at inference, and not those that added parameters: capacity helped on the one task with 105 labels and hurt on the task with 27. The sharpest instance is not a modeling result at all. A single post-processing rule, emitting an empty mask when a presence classifier fires, was defensible under the validation protocol and catastrophic under the test protocol, accounting for most of a 301.2 px change in Hausdorff distance, because the two protocols differ in whether an empty reference can occur. Challenge pipelines are tuned against the metric as much as against the anatomy, and a scoring assumption that changes between phases is a larger risk than a suboptimal backbone. This argues for the direction taken by challenges that specify evaluation in terms of the clinical assessment being automated [15], where the scored quantity does not shift underneath a submission between phases.

References

1. Hammad, M., Bai, J., Zhao, J., Jia, T., Zhang, X., Xu, X., Ma, J., Li, S.: Mitral Valve Anatomy Analysis Using Multimodal Imaging Data. In: MICCAI 2026. Zenodo (2026). https://doi.org/10.5281/zenodo.19726755
2. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)

4. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis (DLMIA). pp. 3–11 (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
6. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS. pp. 1195–1204 (2017)
7. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
8. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571 (2016)
9. Carnahan, P., Moore, J., Bainbridge, D., Eskandari, M., Chen, E.C.S., Peters, T.M.: DeepMitral: Fully Automatic 3D Echocardiography Segmentation for Patient Specific Mitral Valve Modelling. In: MICCAI. pp. 459–468 (2021)
10. Zeng, A., Wu, C., Lin, G., et al.: ImageCAS: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. Computerized Medical Imaging and Graphics **109**, 102287 (2023)
11. Zhu, J., Bai, J., Zhou, Z., et al.: RAS dataset: A 3D cardiac LGE-MRI dataset for segmentation of the right atrial cavity. Scientific Data **11**(1), 401 (2024)
12. Bai, J., et al.: A benchmark framework for right atrial cavity segmentation from LGE-MRIs. IEEE Transactions on Medical Imaging (2025)
13. Chen, Z., Bai, J., Li, S., et al.: CFVP-Net: Coarse-to-fine visual perception network for aortic dissection segmentation. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 5809–5816 (2025)
14. Zia, A., Berniker, M., Perreault, C., Nespolo, R., Jarc, A.: Surgical Visual Understanding (SurgVU) Challenge. In: MICCAI 2025. Zenodo (2024). `https://doi.org/10.5281/zenodo.14054184`
15. Bai, J., Khobo, I., Lu, Y., et al.: Landmark detection challenge for intrapartum ultrasound measurement meeting the actual clinical assessment of labor progress. In: MICCAI 2025. Zenodo (2025). `https://doi.org/10.5281/zenodo.15172238`