Failure-Mode-Driven Multimodal Mitral Valve Segmentation for the MVAA 2026 Challenge

Yixin Chen

Institution of Medical Technology, Peking University, Beijing, China
2311110791@stu.pku.edu.cn

Abstract. Mitral valve assessment spans cardiac CT, 3D TEE, and surgical video frames. These modalities expose different failure modes: sparse CT supervision, noisy ultrasound boundaries, and video-frame false positives that harm distance metrics. Our method uses a failure-mode-driven ensemble rather than a single shared model. For cardiac CT and 3D TEE, we build on 3D nnU-Net style models with fold ensembles, probability fusion, and component filtering. For surgical video frames, we use a semi-supervised 2D ensemble with mean-teacher training, test-time augmentation, and reference-guided far-component filtering. Candidate selection used out-of-fold validation, visual auditing, and challenge validation measurements to reject changes that introduced HD- or ASD-sensitive artifacts. Validation DSC was 0.8609, 0.8599, and 0.8153 for CT, TEE, and video, respectively. For video, reference-guided cleanup reduced HD from 84.64 to 71.77 while preserving DSC. These matched comparisons show that candidate selection should account for boundary-sensitive failure modes in addition to overlap.

Keywords: Mitral valve segmentation · Multimodal medical imaging · Semi-supervised learning · Ensemble learning · Failure-mode analysis.

1 Introduction

Mitral valve repair and replacement planning uses anatomical information from preoperative imaging, intraoperative ultrasound, and surgical video. The MVAA 2026 challenge reflects these imaging settings through three segmentation tasks: cardiac CT, three-dimensional transesophageal echocardiography, and surgical video frames [2]. We abbreviate the ultrasound task as 3D TEE. Although the targets are clinically related, the imaging problems are not interchangeable. CT has limited labeled cases but relatively stable volumetric contrast. TEE has ambiguous leaflet boundaries and ultrasound artifacts. Surgical video has appearance variation, instrument occlusion, and small false-positive components that can disproportionately harm distance metrics.

We therefore used task-specific pipelines rather than forcing all modalities into a single model family, and selected components according to each task's observed failure mode. We treated candidate selection as a failure-control problem:

a candidate should improve or preserve overlap metrics while avoiding empty-case false positives, unsupported remote components, excessive shrinkage, and area inflation. This matters in MVAA because the official evaluation includes DSC, HD, and ASD for all three tasks, and a small remote component may be negligible for DSC but harmful for HD.

Our contribution is a task-specific ensemble and candidate-selection protocol that connects observed errors to metric-aware acceptance rules, rather than a new network architecture. We combine a semi-supervised CT ensemble with branch-level support filtering, a geometry-diverse 3D TEE ensemble, and a semi-supervised video ensemble with reference-guided component filtering. Across tasks, candidates are selected by combining out-of-fold evidence, visual auditing, and challenge validation measurements.

2 Data and Evaluation

The MVAA challenge contains three independent tasks. Task 1 evaluates binary mitral-valve segmentation in cardiac CT volumes. Task 2 evaluates two-class leaflet segmentation in 3D TEE volumes. Task 3 evaluates binary mitral-valve segmentation in surgical video frames. All tasks are assessed using the Dice similarity coefficient, Hausdorff distance, and average surface distance. We abbreviate these metrics as DSC, HD, and ASD. Higher DSC and lower HD or ASD are preferred.

We use the released labeled and unlabeled training data for model development. Model selection and ablation use the challenge validation protocol, in which predictions on the released validation inputs are evaluated against organizer-held annotations. Final performance and the matched comparisons in Tables 2 and 3 are validation-server measurements. Out-of-fold statistics provide complementary internal evidence and are identified explicitly. Figure 1 illustrates the three imaging settings and the type of segmentation evidence available for qualitative inspection.

3 Method

3.1 Task 1: Cardiac CT

The CT pipeline is based on 3D nnU-Net style residual-encoder U-Net models [3, 4]. We use multiple model families trained with complementary supervision. The primary supervised branch uses the standard Dice–cross-entropy objective. A Tversky-oriented branch increases the contribution of false-negative errors under sparse supervision. An offline-soft student learns from cached teacher probabilities on unlabeled CT volumes, while the teacher-refresh branch repeats this training with an updated teacher after component- and confidence-based screening of its pseudo masks. The screening retained 749 of 1,030 unlabeled volumes. This avoided treating visibly fragmented teacher outputs as equally reliable supervision.

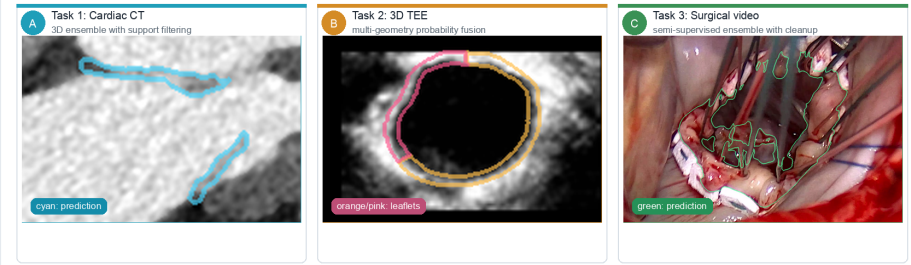

Fig. 1. Representative qualitative examples from the three MVAA modalities. The CT and surgical-video panels show prediction contours produced by the evaluated method on released validation images. The 3D TEE panel shows the two annotated leaflet classes in a labeled training example.

In the final CT configuration, branch-level connected-component support filtering is applied before probability fusion. A 27-case out-of-fold grid search indicated that the three principal branches and the refreshed student provided complementary predictions and supported the relative weight pattern

$$0.3p_{\text{sup}} + 0.1p_{\text{tver}} + 0.3p_{\text{oldsoft}} + 0.3p_{\text{refresh}},$$

The evaluated branch-filtered ensemble uses a foreground threshold of 0.15. Because branch filtering changes the probability distribution, this threshold was selected separately as an empirical operating point and is not interpreted as a calibrated probability. We apply support filtering within each branch but no post-fusion largest-component operation: the latter made masks visually more compact but reduced validation DSC and worsened both distance metrics.

3.2 Task 2: 3D TEE

The TEE pipeline uses 3D nnU-Net residual-encoder models trained under several spacing and optimization variants. The selected five-way probability ensemble combines isotropic 0.25-mm, batch-Dice, isotropic 0.30-mm, fine-z, and isotropic 0.275-mm branches. The fusion weights are:

$$(0.15, 0.15, 0.25, 0.25, 0.20).$$

The candidate was selected from out-of-fold analysis because it improved paired OOF DSC over the four-branch reference fusion and had balanced positive and negative case-level changes. Visual auditing showed local boundary and split refinements without empty-case false positives, remote large components, or mask-area expansion. In the prediction-level change audit, the candidate-to-baseline foreground ratio stayed within 0.989–1.006, the median changed-voxel fraction was 1.06% of baseline foreground, and the maximum area shift was 1.07%. Figure 2 illustrates the visual and case-level evidence used for this decision.

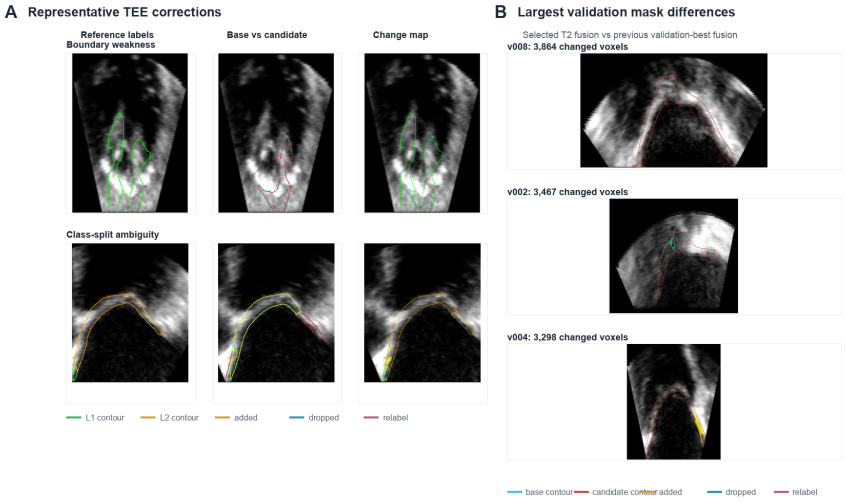

Fig. 2. Task 2 candidate-selection evidence for the isotropic 0.275-mm TEE fusion. Panel A shows representative visual-audit cases for leaflet-boundary weakness and class-split ambiguity. Each row compares reference labels, baseline-versus-candidate contours, and the corresponding change map. Panel B shows the three validation cases with the largest mask differences between the selected fusion and the four-branch reference fusion.

3.3 Task 3: Surgical Video Frames

The video-frame pipeline is a two-dimensional ensemble using UNet++ and related encoder-decoder segmentation networks [7, 1] with EfficientNet-style encoders [5]. Semi-supervised training uses mean-teacher style soft targets for unlabeled frames [6]. At inference time, model probabilities are averaged with flip test-time augmentation and thresholded to obtain binary foreground masks.

For Task 3, the evaluated configuration uses reference-guided cleanup of far components. For each frame, the reference is a fixed auxiliary mask generated by an earlier ensemble for the same image, rather than a manual label. The largest component is always retained. A secondary component is removed only when its area is at least 8 pixels, its centroid is more than 225 pixels from the largest component, its overlap with the reference is below 5%, and its area is below 95% of the largest component. Thus the rule targets unsupported remote predictions that have limited DSC effect but can increase HD and ASD. Reference support and unconditional retention of the largest component reduce the risk of deleting genuine anatomy, although errors shared by both predictions remain possible. The selected cleanup changed 8 of 48 validation frames and removed 2,898 pixels.

Table 1. Failure-mode taxonomy linking each task to its dominant failure source, metric signature, and mitigation.

Task	Failure source	Metric signature	Mitigation and acceptance evidence
CT	Sparse labels and fragmented pseudo masks	Missed valve tissue lowers DSC, while unsupported branches can increase HD or ASD	Complementary supervised and soft-label branches, pseudo-mask screening, and branch-level support filtering
3D TEE	Boundary ambiguity and voxel-spacing sensitivity	Small contour or class-split changes create DSC near-ties but measurable HD/ASD shifts	Multi-geometry fusion, with near-ties accepted only with improved HD/ASD and non-negative OOF and case-level evidence
Video	Remote false positives and frame-level mislocalization	Small components may barely change DSC while dominating HD and increasing ASD	Reference-guided far-component veto, largest-component preservation, and visual and frame-level change audit

3.4 Failure-Mode-Guided Candidate Selection

For each candidate, we first used out-of-fold metrics to identify potential gains. We then inspected case-level differences and visual overlays to determine whether a measured gain was associated with plausible boundary changes or with artifacts such as remote components, empty-frame predictions, or area inflation. Finally, we used challenge validation-server results as an external protocol check. A DSC-near-tie was accepted only when both HD and ASD improved, out-of-fold evidence was non-negative, and the case-level audit found no new structural failure. This rule prevents the primary overlap metric from hiding a small but distant component.

Table 1 lists the dominant failure modes and the control used for each task. Postprocessing was accepted only when it addressed a specific observed failure mode and did not remove the main valve component.

4 Implementation Complexity

The evaluated method contains four five-fold 3D CT branches, five five-fold 3D TEE branches, and 15 two-dimensional video checkpoints before test-time augmentation. The resulting 20 CT fold models and 25 TEE fold models are evaluated sequentially, and their stored probability maps are fused after the branch-level forward passes. This schedule bounds peak memory for the specified NVIDIA Tesla V100 32GB environment. Its inference cost is dominated by the model forward passes, while probability fusion and component filtering add

comparatively little computation. The method therefore prioritizes accuracy and failure diversity over latency and does not target real-time deployment.

We report the model settings, fusion weights, thresholds, and postprocessing rules needed to reproduce inference. Outputs for all 30 CT, 20 TEE, and 48 video validation cases conformed to the organizer schema.

5 Results

Table 2 summarizes the evaluated configuration. Because one final score does not explain the selection decisions, Table 3 reports matched validation comparisons for the interventions most directly linked to the identified failure modes.

Table 2. MVAA validation-server performance of the evaluated configuration. HD and ASD are distance metrics, for which lower values are better. CT and TEE distances use physical image spacing, whereas video distances are measured in pixels.

Task	DSC ↑	HD ↓	ASD ↓
Task 1: CT	0.8609	4.5940	0.2719
Task 2: 3D TEE	0.8599	9.3361	0.5041
Task 3: Video	0.8153	71.7682	11.9414

Table 3. Decision-critical validation comparisons. Within each task, only the named component changes, while the other two task outputs are held fixed. Bold marks selected variant labels and the best value for each metric within a task.

Task	Variant	DSC ↑	HD ↓	ASD ↓
CT	**Branch support without post-fusion LCC**	**0.860900**	**4.5940**	**0.2719**
CT	Additional post-fusion LCC	0.860768	4.6444	0.2754
TEE	DSC-forward fusion	**0.860017**	9.4867	0.5071
TEE	**Distance-aware selected fusion**	0.859921	**9.3361**	**0.5041**
Video	No reference-guided cleanup	0.815008	84.6440	12.2208
Video	200-pixel veto	0.813662	74.9368	12.3584
Video	**225-pixel veto**	**0.815253**	**71.7682**	**11.9414**
Video	240-pixel veto	0.815203	**71.7682**	11.9473

For TEE, the selected fusion sacrificed approximately 0.000097 DSC but improved HD by 0.1506 and ASD by 0.0030. Its paired OOF proxy improved by 0.000554 over the DSC-forward fusion across 105 paired cases, with 53 positive

and 52 negative case-level changes. This satisfied the stated near-tie rule and was therefore preferred to strict validation-DSC ordering. For video, the unfiltered prediction illustrates the metric asymmetry. Removing 2,898 unsupported pixels from only 8 frames increased DSC by 0.000245 but reduced HD by 12.88 and ASD by 0.28. The neighboring 200- and 240-pixel variants further show that the distance veto is not monotonic. Overly aggressive cleanup reduced overlap, whereas overly conservative cleanup gave no additional HD benefit.

Among the eight decision-critical variants in Table 3, three were selected and five were rejected. Four of the five rejected variants improved none of the official metrics relative to the corresponding selected configuration. The remaining TEE alternative had approximately 0.000097 higher DSC but worse HD and ASD, and was rejected by the stated near-tie rule. We restrict this comparison to matched decisions because correlated OOF grid points and output variants derived from the same predictions do not constitute independent candidates.

6 Discussion

The validation results support a task-specific failure-control strategy rather than a single shared architecture. In CT, complementary supervision sources improve robustness, but an additional post-fusion largest-component rule loses both overlap and boundary accuracy. TEE is sensitive to small geometry and training-procedure differences, so probability fusion is evaluated with all three official metrics. In video, the marked HD reduction after changing only 8 frames shows that overlap alone can conceal a remote-component failure. These observations motivate the taxonomy and the staged acceptance rule, but they do not establish that the same controls will transfer unchanged to other datasets.

The main limitation is that the evidence comes from a single challenge dataset and its validation protocol. Validation-server differences among top candidates are small, and the selected candidate prioritizes distance-metric behavior over a strict validation-DSC ordering. This should not be interpreted as evidence of generalization to independent clinical cohorts. The candidate search contains correlated model sweeps and output variants derived from shared predictions. Counting them as independent candidates would overstate the breadth of evidence, so Table 3 reports only matched decision-critical comparisons. Computational cost is another limitation. Sequential evaluation bounds peak memory but increases latency, and deployment-oriented work should examine model pruning, distillation, or branch selection. Finally, visual audit helps screen for remote components and localization failures, but it is not a replacement for labeled case-level evaluation.

7 Conclusion

We presented a failure-mode-driven framework for selecting task-specific ensembles and geometry-aware postprocessing across CT, 3D TEE, and surgical video. Matched validation comparisons show why candidate selection should consider

overlap and distance metrics jointly. Evaluation on independent clinical cohorts and reduction of ensemble inference cost are important directions for further study.

Acknowledgments. The author thanks the MVAA 2026 organizers for releasing the multimodal challenge data, baseline code, and evaluation platform.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European conference on computer vision. pp. 833–851. Springer (2018)
2. Hammad, M., Bai, J., Zhao, J., Jia, T., Zhang, X., Xu, X., Ma, J., Li, S.: Mitral valve anatomy analysis using multimodal imaging data (Apr 2026). https://doi.org/10.5281/zenodo.19726755, `https://doi.org/10.5281/zenodo.19726755`
3. Isensee, F., Jaeger, P.F., Kohl, S.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)
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
5. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PmLR (2019)
6. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
7. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)