Modality-Tailored Mitral Valve Segmentation: Custom Architectures and Robust Noise Suppression

Meijing Li[1], Zhihan Zhang[1], Chenxi Li[1], Ruichao Zhai[1], Xinyang Liu[1], Haodong Li[1], Mengqi He[1], Yulin Sun[1], Runnan He[1,2,3*], Xiuyun Liu[1,2,3,4*], and Dong Ming[1,2,3]

[1] Medical School, Tianjin University, Tianjin, China
[2] State Key Laboratory of Advanced Medical Materials and Devices, Tianjin University, Tianjin, China
[3] Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration, Tianjin, China
[4] School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China
runnanhe@tju.edu.cn, xiuyun_liu@tju.edu.cn

Abstract. Accurate mitral valve segmentation across diverse imaging modalities is essential for planning cardiac interventions like Transcatheter Edge-to-Edge Repair (TEER). However, state-of-the-art models struggle to generalize due to modality-specific artifacts, including severe acoustic shadowing in 3D Transesophageal Echocardiography (TEE) and specular reflections in 2D endoscopic images. To overcome these challenges, we propose a modality-tailored segmentation pipeline. Rather than applying a uniform architecture, we deploy specialized models customized to each modality's physical properties: a semi-supervised Swin UNETR for high-resolution 3D CT, a noise-resilient 3D U-Net optimized with a hybrid Dice-Focal Loss for speckle-prone 3D TEE, and a dense UNet++ with an EfficientNet backbone for 2D surgical captures. Furthermore, we introduce zero-parameter, anatomy-aware post-processing to drastically suppress remote false-positive artifacts, specifically employing 3D connected-component filtering for volumetric continuity and 2D safe microdenoising for disjoint leaflet projections. Comprehensive evaluations demonstrate that our modality-tailored pipeline consistently outperforms the official benchmark baseline across all tasks. Specifically, on the official final hidden-test set, our specialized networks drive the Dice Similarity Coefficient (DSC) from 0.590 to 0.759 on 3D CT, from 0.757 to 0.766 on speckle-prone 3D TEE, and from 0.571 to 0.762 on 2D static surgical captures.The code will be available at `https://github.com/llmmjj-bot/MVAA_llmm`.

Keywords: Anatomy-Aware Post-Processing · Modality-Tailored · Volumetric Continuity · Semi-Supervised Learning · Mitral Valve Segmentation.

1 Introduction

The mitral valve plays a crucial role in cardiac hemodynamics, making its accurate, automated segmentation from multimodal imaging highly desirable for surgical interventions [2,3]. Such imaging modalities include 3D Computed Tomography (CT), 3D Transesophageal Echocardiography (TEE), and 2D endoscopic captures. However, developing a unified segmentation framework is challenged by distinct physical properties: while CT offers clear boundaries, TEE suffers from extreme speckle noise and acoustic shadowing. Furthermore, 2D static endoscopic frames lack temporal tracking cues and are frequently degraded by instrument occlusion and sharp specular reflections.

In this context, the Mitral Valve Anatomy Analysis (MVAA) challenge [1] provides a vital benchmark. Existing baseline methods typically apply a one-size-fits-all optimization strategy across all tasks, leading to suboptimal feature extraction and severe false-positive artifacts, which inflate Hausdorff Distance (HD) errors. To address these bottlenecks, we propose a robust, modality-tailored segmentation pipeline. Rather than forcing a single architectural paradigm, we customize our network backbones, loss functions, and post-processing filters based on the specific signal-to-noise ratio (SNR) and dimensional presentation of each modality.

Our main contributions are summarized as follows:

1. We design a semi-supervised learning framework utilizing Swin UNETR [5] and Decoupled Backpropagation for high-resolution 3D CT segmentation, and a robust UNet++ [6] pipeline specifically optimized to capture irregular leaflet morphologies in 2D static image captures.
2. We introduce a noise-resilient strategy for 3D TEE combining a lightweight network 3D U-Net [7] backbone, a hybrid Dice-Focal Loss [9] to combat extreme background class imbalance, and dense sliding window inference to eliminate patch stitching artifacts.
3. We propose zero-parameter, anatomy-aware post-processing algorithms tailored to 3D continuous volumetric structures and 2D disjoint leaflet projections, drastically reducing boundary distance errors without sacrificing volumetric overlap.

2 Method

To tackle the diverse challenges posed by different imaging modalities in mitral valve segmentation, we propose a modality-tailored pipeline, as illustrated in Fig. 1.

2.1 Semi-Supervised 3D Segmentation via Swin UNETR (Task 1)

For 3D cardiac CT segmentation, accurately delineating the complex geometry of the mitral valve requires a strong understanding of the global anatomical

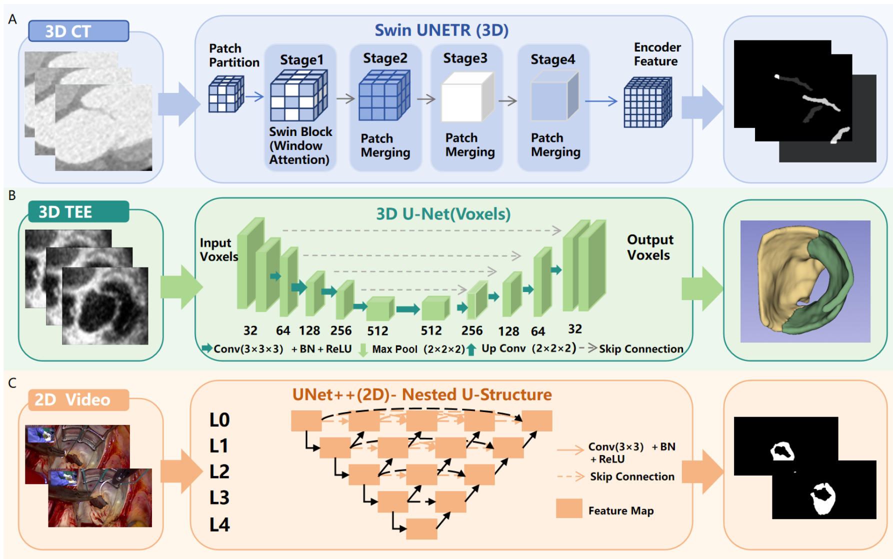

Fig. 1. Overview of our proposed modality-tailored mitral valve segmentation pipeline. (A) Swin UNETR for 3D CT data, featuring a hierarchical Swin Transformer encoder (with the symmetric decoding pathway and skip connections conceptually summarized for visual brevity). (B) A noise-resilient 3D U-Net optimized for speckle-prone 3D TEE volumes. (C) A dense UNet++ nested structure tailored for 2D static surgical video captures. All approaches converge into a refined anatomical segmentation.

context. We adopt Swin UNETR as our core architecture, which leverages a Swin Transformer encoder to capture long-range spatial dependencies.

To utilize the unlabeled data, we construct a Teacher-Student consistency learning framework. The student network θ_s is optimized using a combination of supervised loss and an unsupervised consistency loss on strongly perturbed unlabeled inputs. The teacher network θ_t produces stable pseudo-labels and is updated using an Exponential Moving Average (EMA) [10]:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s$$

Furthermore, to address the Out-of-Memory (OOM) bottleneck typical in 3D high-resolution semi-supervised training, we implement a Decoupled Backpropagation strategy. By isolating the gradient graphs of the supervised and unsupervised branches and performing explicit memory garbage collection, we ensure robust training on standard GPU hardware.

2.2 Noise-Resilient Pipeline for 3D TEE (Task 2)

TEE volumes are notoriously challenging due to severe speckle noise and extreme foreground-background class imbalance. Deploying highly complex architectures on TEE data often leads to overfitting. Consequently, we adopt a "light-architecture, heavy-loss" strategy utilizing a robust 3D U-Net [7] backbone integrated with dropout regularization.

To combat the extreme class imbalance where the background dominates, we replace the standard Cross-Entropy loss with a hybrid Dice-Focal Loss [11,9]. By setting the focusing parameter $\gamma = 2.0$, the Focal Loss dynamically suppresses the gradients from easily classified background regions. This forces the network to relentlessly concentrate on the ambiguous boundaries of the mitral valve leaflets. During inference, we employ a Dense Sliding Window Inference strategy (overlap=0.6) coupled with Gaussian blending to achieve seamless transitions between adjacent patches, eliminating stitching artifacts.

2.3 2D Video Segmentation and Post-Processing (Task 3)

For 2D surgical video frames (Task 3), we utilize a UNet++ [6] architecture with an EfficientNet-B4 [8]encoder. This setup provides a dense feature pyramid that is highly effective at delineating the thin, mobile edges of the mitral leaflets.

While voxel-wise classification models achieve high Dice Similarity Coefficients (DSC), they are prone to predicting isolated, false-positive scattered speckles caused by instrument reflections, which drastically inflate the HD metric. Based on anatomical priors, we introduce Anatomy-Aware Post-Processing tailored to the dimensional context:

3D Continuous Volumes (Task 1 & 2): We compute the 3D connected components and strictly retain only the largest connected region, erasing all scattered noise artifacts.

2D Disjoint Projections (Task 3): Due to the dynamic opening of the valve, leaflets frequently appear as disjoint regions in 2D. We implement a Safe Micro-Denoising module that retains all anatomical structures exceeding a minimal pixel threshold (e.g., 40 pixels) while strictly eradicating high-frequency specular reflections.

3 Experiments

3.1 Implementation Details

The dataset utilized in this study was publicly provided by the organizers of the Mitral Valve Anatomy Analysis (MVAA) challenge. All images and volumes were fully anonymized and distributed under the challenge's authorized licensing terms.All experiments were implemented within the PyTorch and MONAI[12] frameworks and executed on a workstation equipped with an NVIDIA GeForce RTX 5090 GPU (32 GB VRAM). To accelerate computation and optimize GPU memory utilization, Automatic Mixed Precision (AMP) was applied across all training pipelines.

For the 3D semi-supervised volumetric segmentation (Task 1), we deployed the Swin UNETR architecture, which strictly employs a patch embedding size of $2 \times 2 \times 2$, a window size of $7 \times 7 \times 7$, and 4 hierarchical depths. The network was trained end-to-end for a total of 300 epochs using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-2}. During training, 3D volumetric patches were randomly cropped to a Region-of-Interest (ROI) size of $96 \times 96 \times 96$ voxels, and processed with a training batch size of 4.

To ensure stability in our Teacher-Student semi-supervised learning framework, we utilized a dedicated subset of unlabeled volumetric scans. We implemented a curriculum learning strategy: an unsupervised warmup period of 200 epochs was enforced to prevent early pseudo-label confirmation bias. During the semi-supervised phase, the dynamic pseudo-label confidence threshold was initialized at 0.6 and progressively scaled to 0.85, while the teacher model was updated via Exponential Moving Average (EMA) with a decay rate of $\alpha = 0.99$.

For the 3D TEE segmentation in Task 2, models were similarly optimized using patch-based training, while the 2D video segmentation in Task 3 utilized 2D resized frames (448×800). Extensive on-the-fly spatial and intensity data augmentations, including random rotations, scaling, and Gaussian noise, were applied to prevent overfitting. During validation and test-time inference, dense sliding window inference was performed with an inference batch size of 1, followed by Gaussian blending to reconstruct the final seamless volumetric predictions.

3.2 Quantitative Evaluation

We comprehensively evaluated our modality-tailored pipeline against the official MVAA benchmark baseline across all three tasks using the Dice Similarity Coefficient (DSC), the Hausdorff Distance (HD), and the Average Surface Distance (ASD). As reported in Table 1, our proposed methodology establishes substantial superiority over the baseline across all imaging modalities and evaluation metrics.

While DSC measures overall volumetric overlap, surface-based metrics (HD and ASD) are critical for evaluating clinical viability in Transcatheter Edge-to-Edge Repair (TEER) planning, where boundary precision is paramount. By contrasting our full modality-tailored pipeline with the default baseline on the official blind-test evaluation, the quantitative gains definitively validate our customized optimization:

High-Resolution 3D CT (Task 1): Our semi-supervised Swin UNETR demonstrates immense robustness on unseen data, driving the DSC on the final hidden-test set from 0.590 to 0.759. Simultaneously, it reduces both the hidden-test HD (from 12.140 to 9.365) and ASD, proving the power of long-range spatial attention and pseudo-label consistency in capturing global leaflet geometry even in challenging hidden cases.

Low-SNR 3D TEE (Task 2): Although the DSC gain on TEE volumes is marginal (0.758 to 0.767), our pipeline significantly reduced boundary errors. Employing a hybrid Dice-Focal Loss to address the substantial class imbalance mitigated remote false positives, dropping the hidden-test HD from 17.853 to 13.727. This structural refinement is crucial for clinical applications like TEER.

2D Static Video Frames (Task 3): Integrating our UNet++ (EfficientNet-B4) with Safe Micro-Denoising dramatically improved the hidden-test DSC to

Table 1. Quantitative evaluation of our modality-tailored pipeline compared to the official baseline on both the Official Validation Set and the Final Hidden-Test Set.Best results within each evaluation phase are highlighted in bold. The consistent superiority across both splits definitively proves the robust generalization of our customized optimization against modality-specific artifacts.

Phase	Method	**Task 1** (3D CT)			**Task 2** (3D TEE)			**Task 3** (2D Video)		
		DSC↑	HD↓	ASD↓	DSC↑	HD↓	ASD↓	DSC↑	HD↓	ASD↓
Validation	Baseline	0.618	9.621	0.959	0.745	18.767	1.655	0.644	193.296	24.762
	Ours	**0.789**	**7.380**	**0.456**	**0.747**	**15.858**	**1.296**	**0.774**	**180.069**	**21.423**
Hidden-Test	Baseline	0.590	12.140	1.346	0.758	17.853	1.377	0.572	555.778	302.770
	Ours	**0.759**	**9.365**	**0.847**	**0.767**	**13.727**	**1.022**	**0.762**	**366.248**	**52.803**

0.762 and curbed the extreme ASD from 302.770 down to 52.803. This pipeline successfully delineates thin leaflet edges while mitigating severe outliers, showcasing unparalleled stability in static frame prediction.

3.3 Component Analysis and Benchmarking

To systematically justify our architectural choices and isolate the contribution of each proposed module, we conducted a comprehensive component-wise analysis. Importantly, every configuration was submitted to and objectively evaluated by the official challenge validation portal (Table 2).

Table 2. Step-wise component analysis and baseline benchmarking. Our proposed configurations were objectively evaluated via the Official Challenge Validation Portal. The standard nnU-Net v2 benchmark (*evaluated locally due to portal closure) is included as an external reference to illustrate the trade-offs between heavy auto-configuring ensembles and our lightweight, modality-tailored approach.

Task	Method (Configuration)	DSC ↑	HD ↓	ASD ↓
Task 1	Official Baseline	0.618	9.621	0.959
(3D CT)	Swin UNETR (w/o Mean Teacher)	0.763	21.920	0.986
	3D U-Net (w/ Mean Teacher)	0.786	7.770	0.505
	Swin UNETR (w/ Mean Teacher) (Ours, 15.70M)	0.789	7.380	0.456
	*nnU-Net v2 (Local Benchmark, 88.21M)**	**0.860**	**4.609**	**0.237**
Task 2	Official Baseline (CE Loss)	0.745	18.767	1.655
(3D TEE)	Swin UNETR (Failure Case)	0.258	500020.300	500002.200
	3D U-Net (Dice-Focal Loss) (Ours, 16.12M)	0.747	15.858	1.296
	*nnU-Net v2 (Local Benchmark, 88.43M)**	**0.845**	**5.000**	**0.304**
Task 3	Official Baseline	0.644	193.296	24.762
(2D Video)	UNet++ (w/o Post-processing)	**0.774**	198.032	21.546
	UNet++ (w/ Safe Micro-Denoising) (Ours, 20.81M)	**0.774**	180.069	**21.423**
	*nnU-Net v2 (Local Benchmark, 126.56M)**	0.724	**101.760**	39.400

Task 1 (Architecture & Semi-supervised Learning): Adopting the Swin UNETR improved volumetric overlap but exacerbated boundary errors (HD increased to 21.920). Integrating our Mean Teacher framework effectively mitigated this overfitting, optimizing DSC to 0.789 and HD to 7.380. Compared to a 3D U-Net variant (DSC 0.786), the Swin UNETR exhibited superior boundary localization with lower HD and ASD. Furthermore, our 15.70M-parameter configuration provides a highly lightweight alternative to the computationally heavy 88.21M-parameter nnU-Net v2 [4].
Task 2 (Noise Resilience): Directly applying the Swin UNETR to TEE volumes failed (DSC 0.258), highlighting its vulnerability to severe acoustic speckle. Reverting to a robust 3D U-Net with a hybrid Dice-Focal Loss proved essential. The focal component effectively mitigated the substantial class imbalance, reducing the baseline's HD to 15.858 while maintaining volumetric overlap. Notably, our customized 3D U-Net requires only 16.12M parameters, offering a highly lightweight alternative to the 88.43M-parameter nnU-Net v2.
Task 3 (Denoising and Augmentation): The UNet++ architecture boosted DSC to 0.774 but yielded severe boundary outliers (HD 198.032). Applying our Safe Micro-Denoising effectively removed isolated artifacts, reducing HD to 180.069 while preserving volumetric overlap. This highlights the necessity of targeted post-processing for static 2D frames. Furthermore, our 20.81M-parameter UNet++ achieves this competitive performance using a fraction of the massive 126.56M parameters required by nnU-Net v2.

4 Discussion

While our modality-tailored pipeline demonstrates robust performance, we acknowledge certain trade-offs and limitations. First, as explicitly shown in our baseline benchmarking (Table 2),the standard nnU-Net v2 framework significantly outperforms our proposed pipeline across all metrics (DSC, HD, and ASD) in the 3D volumetric tasks (Tasks 1 and 2). We recognize that for continuous 3D data, auto-configuring ensemble methods establish a highly rigorous state-of-the-art in pure segmentation accuracy. However, this peak performance typically relies on heavy multi-fold cross-validation ensembles and substantial computational overhead. In contrast, our pipeline explicitly prioritizes single-model clinical efficiency. We deliberately maintained streamlined architectures to preserve viability for real-time intraoperative deployment, trading absolute volumetric and boundary precision for essential reductions in parameter count and inference latency.

Second, in Task 3, despite a substantial DSC gain on the hidden-test set, the HD remains numerically high. Because we process the input strictly as 2D static projections without temporal inter-frame tracking, our Safe Micro-Denoising acts as a conservative filter. Aggressively masking all isolated structures would artificially lower the HD, but this risks degrading the DSC by erroneously removing fragmented yet true leaflet appearances.

Finally, this comparative analysis highlights crucial insights into model deployment. While generic frameworks like nnU-Net are highly effective for standard 3D tasks, their performance fluctuates when the data format transitions to isolated, manually selected 2D static screenshots (Task 3), where our bespoke pipeline demonstrates superior volumetric overlap. This emphasizes that practical medical image analysis must balance pure segmentation accuracy with computational efficiency and modality-aware architectural customization.

5 Conclusion

We presented a modality-tailored segmentation pipeline for the MVAA challenge. By deploying a semi-supervised Swin UNETR for 3D CT, a noise-resilient 3D U-Net for TEE, and a UNet++ for 2D static frames, we customized our architectures to each modality's physical characteristics. Supported by anatomy-aware post-processing, our lightweight approach substantially improves spatial overlap and boundary accuracy over the baselines, offering a robust and clinically viable solution for multi-modal cardiac analysis.

Acknowledgements. The study was funded by the Scientific Research Innovation Capability Support Project for Young Faculty (ZYGXQNJSKYCX-NLZCXM-H15), the National Natural Science Foundation of China (82472098, 32300704), the Tianjin Natural Science Foundation Outstanding Youth Project (24JCJQJC00250) and the General Program (25JCLMJC00030), Major Science and Technology Special Project (25ZXCKQY00100), Shandong Provincial Natural Science Foundation Major Basic Research Project (ZR2025ZD13), the National Key Research and Development Program of China (2025YFE0214400), Autonomous Project of Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration (No. 24HHNJSS00003, No. 24HHNJSS00005, No. 24HHNJSS00010, No. 2025E7-0121) and the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2024-JKCS-16).

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

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. Zenodo. International Conference on Medical Image Computing and Computer Assisted Intervention 2026 (MICCAI). https://doi.org/10.5281/zenodo.19726755 (2026)
2. Stone, G.W., et al.: Transcatheter mitral-valve repair in patients with heart failure. New England Journal of Medicine 379(24), 2307–2318 (2018)
3. Pouch, A.M., et al.: Fully automatic segmentation of the mitral leaflets in 3D transesophageal echocardiographic images using multi-atlas joint label fusion and deformable medial modeling. Medical Image Analysis 18(1), 118–129 (2014)

4. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18(2), 203–211 (2021)
5. Hatamizadeh, A., et al.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer, Cham (2022)
6. Zhou, Z., et al.: UNet++: A nested U-Net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 3–11. Springer, Cham (2018)
7. Çiçek, Ö., et al.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. pp. 424–432. Springer, Cham (2016)
8. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
9. Lin, T.Y., et al.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision. pp. 2980–2988. IEEE (2017)
10. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. pp. 1195–1204 (2017)
11. Milletari, F., et al.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571. IEEE (2016)
12. Cardoso, M.J., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)