Uncertainty-Aware Fusion of Foundation and Task-Specific Models for Cardiac MRI Segmentation

Mosarrat Rumman, Heidar Davoudi, and Mehran Ebrahimi

Faculty of Science, Ontario Tech University, Oshawa, Ontario, Canada

ABSTRACT

Vision foundation models, such as the Segment Anything Model (SAM), demonstrate strong zero-shot generalization but lack precision with anatomically challenging objects. In contrast, convolutional neural network (CNN)-based models, such as nnU-Net, achieve high accuracy on domain-specific data but struggle to generalize on unseen data. To address these complementary limitations, we propose an uncertainty-aware fusion framework that integrates the generalizability of foundation models with the anatomical precision of task-specific models for cardiac MRI segmentation. The proposed approach combines Dempster-Shafer Theory (DST) with an entropy-guided fallback mechanism to perform voxel-wise fusion of calibrated probability maps. Unlike simple ensemble methods, the framework takes into consideration the inter-model agreement, conflict, and the uncertainty of the models. DST fusion is applied where the models agree, while high-conflict regions are handled by an entropy-guided fallback mechanism that selects predictions from the more reliable model. Extensive evaluation on the M&Ms dataset (in-domain) and the ACDC dataset (cross-domain) demonstrates consistent improvements in Dice and IoU across model pairings of varying strengths. In-domain gains are modest, whereas cross-domain evaluations show substantially larger improvements. Notably, the nnU-Net+SAM2 pairing achieves relative gains of approximately 8% in Dice and 11% in IoU on the cross-domain dataset. Comparisons with simple averaging, ablation studies, and statistical analysis confirm the effectiveness of the proposed framework. To our knowledge, this is the first application of voxel-wise DST-based fusion to combine vision foundation models with task-specific CNNs for cardiac MRI segmentation.

Keywords: cardiac segmentation, uncertainty quantification, SAM, MedSAM2, nnU-Net, U-Net, Fusion, MRI, Foundation model, Dempster-Shafer Theory

1. INTRODUCTION

Accurate and reliable segmentation of cardiac tissue is vital for providing insight into patient health and disease progression, and assisting in clinical decisions of diseases such as heart failure, hypertrophy, ischemia, etc.[1,2] Although manual segmentation by radiologists remains the clinical standard, it is time-consuming, subjective, and prone to inter-observer variability. The emergence of deep learning has been a breakthrough and is striving to achieve near human-level performance.[3] Many task-specific deep learning models achieve high accuracy in segmenting cardiac anatomies. However, these models are resource-intensive, require large annotated datasets, and often fail to generalize to data from different scanners, centers, or image protocols.[4] On the other hand, vision foundation models such as the Segment Anything Model (SAM) demonstrate strong zero-shot generalizability but fail to detect small anatomical structures precisely, therefore compromising the quality of segmentation masks.[5] These complementary characteristics motivate the fusion of foundation and task-specific models that can achieve both generalization and anatomical precision. However, most existing ensemble and fusion methods such as probability averaging or majority voting, assume that all models are equally reliable across the image and ignore the uncertainty of the models. Such assumptions result in unreliable segmentation, especially in cross-domain settings, where models may produce overconfident yet incorrect predictions.

Further author information: (Send correspondence to M.R.)
M.R.: E-mail: mosarrat.rumman@ontariotechu.net
H.D.: E-mail: heidar.davoudi@ontariotechu.ca
M.E.: E-mail: Mehran.Ebrahimi@ontariotechu.ca

Medical Imaging 2026: Computer-Aided Diagnosis, edited by Axel Wismüller, Thomas Martin Deserno, Proc. of SPIE Vol. 13926, 1392618 · 2026
Published by SPIE · 1605-7422 · doi: 10.1117/12.3087524

To address these challenges, we propose an uncertainty-aware fusion framework based on Dempster-Shafer Theory (DST). The framework performs voxel-wise fusion of calibrated probability maps while calculating inter-model agreement and conflict. To mitigate the known instability of DST in high-conflict regions, an entropy-guided fallback mechanism is introduced that avoids forced fusion and instead selects safer predictions when disagreement is extreme. The proposed approach is model-agnostic and does not require any retraining. This work investigates the following research questions: 1) How can Dempster-Shafer Theory be used to fuse voxel-level predictions from foundation and task-specific segmentation models? 2) How uncertainty-aware DST fusion behaves when combining models of varying segmentation capabilities? 3) Does an entropy-guided fallback mechanism improve segmentation stability in regions of high inter-model conflict? 4) How well does the proposed framework generalize across domains without additional retraining, and 5) how does the proposed approach compare to individual models and simple ensemble methods in terms of accuracy and stability?

The main contributions of this paper are threefold. First, we introduce a novel uncertainty-aware DST fusion framework for combining segmentation outputs from different models in a principled manner for each voxel. Second, an entropy-guided fallback mechanism is proposed that ensures stable and reliable segmentations in high-conflict regions. Third, to assess under what circumstances the proposed framework is most beneficial; models with varying strengths are fused and evaluated on the M&Ms dataset for in-domain testing and the ACDC dataset for cross-domain generalization. Experimental results demonstrate that the proposed framework consistently improved segmentation performance over the best base model, across all experimental settings.

2. RELATED WORK

Image segmentation is one of the primary focuses of image processing tasks using machine learning. Edge detection and thresholding[6] are some of the noteworthy methods that have been used to perform image segmentation over the years. With the advancement of deep learning, convolutional network based techniques became popular; for example, Bai W. *et al.*[7] developed and proposed the use of a fully convolutional network (FCN) for cardiovascular magnetic resonance image analysis. In the current landscape, cutting-edge methods for image segmentation increasingly involve the use of foundation models.

Ronneberger O. *et al.*[8] developed U-Net in 2015, in their seminal work, they introduced the novel method of achieving state-of-the-art segmentation accuracy with sparse training data. The authors modified the existing Fully Convolutional Network(FCN) architecture and added symmetry to it using two distinct paths of convolutional layers. They proposed supplementing the usual contracting network with successive layers that replace pooling operators with upsampling operators. This design increases feature resolution at the output and propagates contextual information to higher-resolution layers. Also to tackle the scarcity of data they used extensive use of data augmentation. The high accuracy achieved by U-Net using relatively small datasets established it as a benchmark method for medical image segmentation tasks.

Isensee F. *et al.*[9] introduced the concept of a self-adapting network with nnU-Net, The notable feature of nnU-Net is that it automatically configures its architecture and hyperparameters based on dataset fingerprints. The authors designed nnU-Net to generalize across ten different medical datasets covering multiple anatomical structures. nnU-Net is now considered one of the state-of-the-art task-specific segmentation models.[10] However, despite their high in-domain accuracy, such task-specific models often fail to generalize and remain fragile under domain shift. This limitation motivates the need for models with broader semantic understanding.

In 2023, the vision foundation model "Segment Anything Model" (SAM) was introduced.[11] SAM was trained on a massive dataset and demonstrated unprecedented zero-shot segmentation capabilities. However, Fan K. *et al.*[12] showed that vision foundation models (VFMs) struggle with domain-specific challenges in medical image segmentation, such as low-contrast soft tissue boundaries, irregular anatomical shapes, and inaccurate delineation of small structures. This highlights a fundamental trade-off between the high in-domain accuracy of CNN-based models and the strong cross-domain generalization of VFMs.

In medical imaging, uncertainty quantification (UQ) plays a critical role in understanding the confidence of deep learning models in their predictions.[13] Several uncertainty estimation methods have been explored for medical segmentation, among which Monte Carlo Dropout (MCD) and deep ensembles are widely used.[13] MCD approximates Bayesian inference by performing multiple stochastic forward passes during inference, while deep

ensembles aggregate predictions from multiple independently trained models. Mehrtash A. *et al.*[14] demonstrated that deep learning models can be highly confident yet incorrect, and that UQ can help calibrate inter-observer disagreement. Their work showed that both cross-entropy loss and Dice loss can be used for calibration, with Dice loss yielding superior calibration performance. By quantifying uncertainty at the voxel level, it becomes possible to identify regions where task-specific models fail under domain shift or where foundation models hallucinate boundaries.

Combining multiple models to improve performance and reliability is a well-established strategy in medical image analysis. Classical label fusion methods such as STAPLE,[15] SIMPLE,[16] and Joint Label Fusion[17] were originally developed for multi-atlas segmentation and rely on estimating global or regional performance weights for multiple annotators or segmenters. Ensemble learning techniques commonly used in deep learning include probability averaging, majority voting, and weighted ensembles. However, these approaches implicitly assume comparable model reliability across the image and typically fuse predictions uniformly, without explicitly accounting for voxel-wise disagreement or uncertainty. As a result, they may produce unreliable predictions when models strongly disagree.

Motivated by these limitations, this work adopts Dempster–Shafer Theory (DST),[18] which provides an explicit mechanism for modeling agreement and conflict between heterogeneous predictors at the voxel level. DST has been successfully applied in medical image segmentation alongside deep learning. For example, recent frameworks[19] map feature representations into Dempster–Shafer mass functions that encode evidence for class labels at each voxel, incorporating contextual discounting based on source reliability. Dempster's rule is then used to combine evidence from multiple sources, enabling uncertainty-aware fusion with improved robustness.

A review of the above literature highlights the potential of DST-based fusion for integrating vision foundation models and U-Net–based architectures to combine complementary strengths while explicitly modeling disagreement. However, existing work offers a limited systematic investigation of fusing foundation models such as SAM with task-specific architectures like nnU-Net. Moreover, there is a lack of frameworks that leverage voxel-level uncertainty to enable dynamic collaboration between heterogeneous models, particularly in rejecting confident but incorrect predictions. This work aims to address these gaps by explicitly modeling prediction uncertainty and conflict, and by fusing model outputs only when they are reliably confident.

3. METHODOLOGY

This section describes the proposed uncertainty-aware fusion framework based on DST. The dataset details and the base models that have been used for fusion are outlined. We also discuss the DST-based fusion process, including probability calibration, uncertainty modeling, and conflict-aware decision strategies.

3.1 Dataset

The proposed framework is evaluated on two publicly available cardiac cine MRI datasets. This allows assessment of both in-domain performance and performance under domain shift.

M&M Dataset (In-Domain) - The Multi-center, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge dataset[20] includes short-axis cine MRI from multiple centers and scanners. Each volume is annotated with labels for the left ventricle (LV), right ventricle (RV), and left ventricular myocardium (MYO) at both end-diastolic (ED) and end-systolic (ES) phases. A patient-wise split yielded 576 cases for training and 124 test cases for in-domain testing.

ACDC Dataset (Cross-Domain) - The Automated Cardiac Diagnosis Challenge (ACDC) dataset[21] consists of short-axis cine MRI volumes acquired under different imaging protocols and scanner settings compared to the M&Ms dataset. Each subject includes annotations for the left ventricle (LV), right ventricle (RV), and left ventricular myocardium (MYO) at both end-diastolic (ED) and end-systolic (ES) phases. All models trained on M&Ms are directly evaluated on ACDC without retraining to assess cross-domain generalization of the proposed framework.

3.2 Base Segmentation Models

The proposed fusion framework is built on the observation that the foundation and task-specific segmentation models have complementary strengths and failures when applied to cardiac MRI. Foundation models are generalizable across cross-domain datasets, while task-specific models are better at detecting edge boundaries. For this reason, models from each of these categories are fused. Furthermore, the proposed fusion framework can handle different model pairing scenarios. By combining models of different strengths, we systematically analyze fusion behavior under different model pairings. Here, the strength of the models corresponds to the segmentation performance of the models on both datasets.

3.2.1 Foundation model

Vision foundation models are pre-trained on large-scale datasets, which enables them to learn generic visual representations and exhibit strong zero-shot generalization. In medical image segmentation, where datasets are limited and annotation is highly expensive, this property of the foundation model is highly valued. In this study, we explored two variants of the Segment Anything Model[11] developed by Meta - MedSAM2, serving as the stronger model and fine-tuned SAM2 serving as the suboptimal model.

MedSAM2[22] is a medical image segmentation foundation model derived from fine-tuning the Segment Anything Model 2 (SAM2)[23] on extensive medical datasets, including over 455,000 3D image–mask pairs and 76,000 annotated video frames across multiple modalities. It retains SAM2's hierarchical transformer backbone, prompt encoder, and mask decoder, augmented with memory-based mechanisms for volumetric and temporal consistency. MedSAM2 demonstrates strong generalization to diverse anatomical structures, making it an effective and robust basis model for fusion. MedSAM2 was already trained on M&Ms and ACDC datasets, therefore showed strong segmentation performance on both datasets. For our fusion model pairings, MedSAM2 served as a **strong** model.

Fine-tuned SAM2 is intentionally used as a suboptimal model in this study. The original SAM2 is primarily trained on natural image datasets and does not transfer well to medical image segmentation tasks. When applied in a zero-shot manner on M&Ms and ACDC datasets, the original SAM2 produced unreliable and anatomically inconsistent segmentation. Therefore, we applied light fine-tuning on M&M dataset while preserving most of its pretrained weights. This limited training improved basic anatomical learning without fully optimizing the model. This resulted in a suboptimal yet functional model, which is crucial to studying how fusion works under such circumstances.

3.2.2 Task-specific CNN models

Task-specific convolutional neural networks (CNNs) are supervised models, designed, trained and optimized for a given segmentation task. Their strong hierarchical feature extraction and the ability to learn local patterns enable them to detect anatomical structures precisely and accurately. However, as these models are trained to minimize task-specific loss functions on a fixed dataset distribution, these models lack generalizability under domain shift. In this study, we explored two task-specific CNN segmentation models - nnU-Net[9,10] serving as a strong model, and UNet++[24] serving as a suboptimal model in in-domain settings and a weak model in cross-domain settings.

nnU-Net is a complete segmentation pipeline. Rather than using a single fixed network, nnU-Net systematically adapts its entire pipeline, such as preprocessing of the data, network topology, patch size, batch size, normalization, and training schedule, depending on the properties of the dataset such as image resolution, class imbalance, etc. This data-driven configuration has enabled nnU-Net to achieve state-of-the-art performance on a wide range of medical datasets, including M&Ms and ACDC datasets. However, training nnU-Net on M&Ms dataset only, its aggressive optimization and strong dependence on the dataset properties, result in the generation of overconfident probabilities under domain shift. This behavior makes nnU-Net an ideal candidate for uncertainty-aware fusion where its high confidence can be moderated by more generalizable foundation models.

UNet++ extends the original U-Net[8] architecture by introducing nested and dense skip connections between encoder and decoder, transferring feature maps from earlier layers to later layers. Thus, the network can preserve spatial information better than vanilla U-net. In this work, UNet++ is trained from scratch on 2D slices extracted from 3D volumes of the M&Ms dataset. To intentionally construct a suboptimal baseline, no additional data preprocessing such as data balancing or architectural modifications were made during training. This resulted

in a suboptimal in-domain segmentation performance and very weak performance under domain shift. This baseline allowed us to evaluate whether the proposed uncertainty-aware fusion framework can compensate for a weak performing task-specific model, by using the complementary strength of the foundation models, to produce better segmentation masks.

3.3 Uncertainty-Aware Model Fusion via Dempster-Shafer Theory

Dempster–Shafer Theory (DST),[18] also known as the theory of belief functions, is a framework for reasoning under uncertainty. This theory extends classical probability theory by explicitly modeling both uncertainty and conflict between multiple sources of evidence. Let $\Theta = \{\theta_1, \ldots, \theta_C\}$ denote the list of all possible hypotheses. A basic belief assignment (BBA) is defined as a function $m : 2^\Theta \to [0, 1]$ such that $m(\emptyset) = 0$, i.e., no belief value is assigned to the empty set, and $\sum_{A\subseteq\Theta} m(A) = 1$; all belief values sum to one. Given two independent sources with belief assignments m_1 and m_2, Dempster's rule of combination is defined as

$$m_{1\oplus2}(A) = \frac{1}{1-K} \sum_{B\cap C=A} m_1(B)m_2(C), \tag{1}$$

Here, B and C denote the subsets of Θ supported by the two evidence sources, while $A = B \cap C$ represents their intersection and agreement. The conflict term

$$K = \sum_{B\cap C=\emptyset} m_1(B)m_2(C) \tag{2}$$

quantifies the degree of disagreement between the two sources. In the context of voxel-wise medical image segmentation, probability distributions from different models can be interpreted as belief masses over anatomical classes. This allows DST to combine model outputs while explicitly measuring agreement and conflict. This property is particularly valuable when fusing heterogeneous predictors, such as vision foundation models and task-specific CNNs, which exhibit complementary strengths and failures. Although DST can produce unstable results when two models strongly disagree, the conflict term allows us to detect such cases. This conflict information is therefore used to guide fusion strategies rather than forcing fusion in unreliable regions.

Belief Mass Construction via Probability Calibration Deep learning models can be overconfident even when wrong. To ensure that predicted probabilities accurately reflect model confidence, we apply post-hoc temperature scaling. Given the voxel-wise logits $z^{(A)}$ and $z^{(B)}$ produced by Model A and Model B, respectively, temperature scaling is applied as

$$\hat{p}^{(A)} = \mathrm{softmax}\left(\frac{z^{(A)}}{T_A}\right), \quad \hat{p}^{(B)} = \mathrm{softmax}\left(\frac{z^{(B)}}{T_B}\right). \tag{3}$$

The temperature parameters T_A and T_B are estimated independently for each model by minimizing the multiclass negative log-likelihood (NLL) on a held-out calibration set, which is exclusive of the training and test set. T is selected by minimizing

$$T^* = \arg\min_{T>0} \sum_n \mathcal{L}_{\mathrm{NLL}}\big(\mathrm{softmax}(z_n/T), y_n\big), \tag{4}$$

where (z_n, y_n) denote the logits and ground-truth labels of the calibration samples. This calibration step mitigates overconfidence and is critical for reliable and comparable uncertainty estimation. These calibrated probabilities $\hat{p}^{(A)}$ and $\hat{p}^{(B)}$ are interpreted as singleton belief masses and used for DST fusion.

Uncertainty Estimation and Model Confidence Model confidence at each voxel corresponds to the maximum calibrated class probability of that voxel, while uncertainty is quantified using Shannon entropy:

$$H = -\sum_{c=1}^{C} \hat{p}_c \log \hat{p}_c. \tag{5}$$

Although entropy derived from softmax probabilities does not represent full epistemic uncertainty, it provides a simple, model-independent measure of how uncertain a prediction is by quantifying how evenly the predicted probabilities are distributed across classes. Since our framework utilizes uncertainty only for relative decision-making, entropy derived from softmax probabilities can be sufficient for the resolution of model conflicts. A more detailed discussion of alternative uncertainty measures is provided in the Discussion (Section 6).

Conflict-Aware Dempster-Shafer Fusion Given the calibrated singleton belief masses $\hat{p}^{(A)}$ and $\hat{p}^{(B)}$, the voxel-wise disagreement between the two models is quantified using the Dempster-Shafer conflict term

$$K = \sum_{i \neq j} \hat{p}_i^{(A)} \hat{p}_j^{(B)}, \tag{6}$$

which measures the total belief mass assigned to mutually exclusive hypotheses. A low value of K indicates agreement between the models; whereas a high value indicates strong disagreement, which reflects the models assigning high confidence to different classes. When the conflict K is below a predefined threshold τ, the predictions are considered compatible and Dempster's rule of combination is applied. In this case, the fused belief mass for each class at a given voxel is computed as

$$m_c = \frac{\hat{p}_c^{(A)} \hat{p}_c^{(B)}}{1 - K}, \quad c \in \{1, \ldots, C\}, \tag{7}$$

and the final fused class label is obtained by

$$\hat{y} = \arg\max_c m_c. \tag{8}$$

The conflict threshold τ is selected through empirical analysis on a held-out calibration set by testing a range of candidate values and choosing the threshold value that prevents unstable fusion in high conflict regions, while maintaining segmentation performance.

High-Conflict and Low-Confidence Handling. Dempster-Shafer fusion is known to become unstable when the conflict term approaches unity, i.e., when two models assign near-certain belief to different hypotheses or classes. To address this limitation, fusion is avoided whenever $K \geq \tau$, and an entropy-guided fallback mechanism is employed. We look at the model uncertainty, which is quantified using Shannon entropy, and the prediction from the model with lower entropy (i.e., higher confidence) is selected:

$$\hat{y} = \begin{cases} \arg\max_c \hat{p}_c^{(A)}, & \text{if } H_A < H_B, \\ \arg\max_c \hat{p}_c^{(B)}, & \text{otherwise.} \end{cases} \tag{9}$$

When both models are uncertain, i.e., indicated by high entropy and low prediction confidence, the voxel is considered unreliable. In such cases, the framework avoids forced fusion and instead selects the prediction from the model that is relatively more confident. This mechanism helps prevent unreliable or counter-intuitive predictions in uncertain regions. If there is a rare case of a tie, i.e., both models have the exact same confidence on a specific voxel, with different hypotheses, a tie breaker is used. More details can be found in the Discussion section 6.4.

Summary of the Fusion Procedure. The complete uncertainty-aware fusion process is summarized in Algorithm 1, with notation defined in Table 1. The key steps, such as probability calibration, conflict estimation, and entropy-guided fallback, are all unified into a voxel-wise decision process. The proposed framework overcomes the inherent limitations of classical Dempster-Shafer fusion and conventional ensemble methods. This creates a robust fusion framework that can work effectively, especially under domain shift, where model uncertainty can play a vital role.

Table 1: Notation used in Algorithm 1.

Symbol	Meaning
C	Number of classes
$z^{(A)}, z^{(B)} \in \mathbb{R}^C$	Logits of Model A and Model B
$T_A, T_B \in \mathbb{R}_+$	Temperature-scaling parameters
$\hat{p}^{(A)}, \hat{p}^{(B)} \in [0,1]^C$	Calibrated class probabilities after softmax
$H_A, H_B \in \mathbb{R}_+$	Shannon entropy (model uncertainty)
$K \in [0,1]$	Dempster–Shafer conflict between the two models
$m \in [0,1]^C$	Fused mass (belief) distribution
$\tau \in [0,1]$	Conflict threshold for fusion vs fallback
$\hat{y} \in \{1, \ldots, C\}$	Final fused class prediction

Algorithm 1 Uncertainty-aware fusion of two calibrated models via Dempster–Shafer Theory

Inputs: logits $z^{(A)}, z^{(B)}$, temperatures T_A, T_B, conflict threshold τ
Output: fused class label $\hat{y}$

1. Probability calibration & uncertainty estimation

$\hat{p}^{(A)} \leftarrow \mathrm{softmax}(z^{(A)}/T_A)$
$\hat{p}^{(B)} \leftarrow \mathrm{softmax}(z^{(B)}/T_B)$
$H_A \leftarrow -\sum_c \hat{p}_c^{(A)} \log \hat{p}_c^{(A)}$ ▷ Shannon entropy (uncertainty)
$H_B \leftarrow -\sum_c \hat{p}_c^{(B)} \log \hat{p}_c^{(B)}$

2. Dempster–Shafer fusion with conflict handling

$K \leftarrow \sum_{i \neq j} \hat{p}_i^{(A)} \hat{p}_j^{(B)}$ ▷ inter-model conflict using DST
if $K < \tau$ **then** ▷ low conflict → fuse
    $m_c \leftarrow \frac{\hat{p}_c^{(A)} \hat{p}_c^{(B)}}{1-K}$ ▷ combined belief mass using DST
    $\hat{y} \leftarrow \arg\max_c m_c$
else ▷ high conflict → trust less uncertain model
    $\hat{y} \leftarrow \begin{cases} \arg\max_c \hat{p}_c^{(A)}, & \text{if } H_A < H_B, \\ \arg\max_c \hat{p}_c^{(B)}, & \text{otherwise.} \end{cases}$
end if
return $\hat{y}$

4. EXPERIMENTAL SETUP

This section describes dataset splits used in this research, how the base models are trained and tested, how the fusion pairs are determined, details about the implementation of the approach, and the metrics used for performance evaluation.

4.1 Datasets and Evaluation Protocol

All experiments are conducted on the M&Ms and ACDC cardiac cine MRI datasets described in Methodology Section 3.1. The M&Ms dataset is used for both training and in-domain evaluation, while the ACDC dataset is used exclusively for cross-domain testing to assess generalization and robustness under domain shift. A patient-wise split is employed to avoid data leakage across training, calibration, and test sets. From the M&Ms dataset, We use 576 cases for training, 20 cases for the calibration set, and 124 test cases for in-domain testing. From the ACDC dataset, 20 cases are used for the calibration set and 80 cases are used for cross-domain testing. Notably, no model is retrained or fine-tuned on the ACDC dataset; all cross-domain results are obtained in a zero-shot manner.

4.2 Model Training and Inference

Task-specific CNN models are trained using supervised learning on the M&Ms dataset only. The nnU-Net model is trained using its self-configuring pipeline, including automated preprocessing, normalization, and a

training schedule optimized for the dataset. UNet++ is trained from scratch on 2D slices extracted from 3D M&Ms volumes using standard cross-entropy loss, without data augmentation, architectural modifications, or class balancing, in order to intentionally construct a suboptimal baseline.

For foundation models, MedSAM2 is used in a zero-shot setting, as it has already been pre-trained on diverse medical imaging datasets including M&Ms and ACDC. Fine-tuned SAM2 is obtained by applying light fine-tuning on the M&Ms data set while preserving most of its pretrained weights, resulting in a deliberately suboptimal but functional model. During inference, all models generate voxel-wise softmax probability maps, which are used as inputs to the fusion framework.

4.3 Fusion Configurations

To systematically analyze the behavior of the proposed uncertainty-aware fusion framework, experiments are conducted under four model pairing scenarios:

- **Strong-Strong**: MedSAM2 + nnU-Net (in-domain).
- **Strong-Suboptimal**: nnU-Net+fine-tuned SAM2 (in-domain), MedSAM2+nnU-Net (cross-domain).
- **Suboptimal-Suboptimal**: Fine-tuned SAM2 + UNet++ (in-domain), Fine-tuned SAM2 + nnU-Net (cross-domain).
- **Suboptimal-Weak**: Fine-tuned SAM2 + UNet++ (cross-domain).

Model strength is defined *for descriptive and analytical purposes only* based on macro-averaged Dice performance. Models achieving Dice scores above 0.90 are referred to as *strong*, those with Dice between 0.70 and 0.90 are referred as *suboptimal*, and models below 0.70 as *weak*. These thresholds are not intended as absolute performance criteria, but are used solely to facilitate comparison and interpretation of fusion behavior across different models with different predictive capabilities.

4.4 Probability Calibration and Fusion Parameters

Temperature scaling is applied independently to each model using a held-out calibration set - a sample of 40 (20 from M&Ms and 20 from ACDC datasets). Calibration samples are excluded from both training and test sets. The temperature parameters are optimized by minimizing the multiclass negative log-likelihood (NLL). For Dempster-Shafer fusion, parameters such as the conflict threshold τ and the low-confidence margin are selected empirically on the calibration set by evaluating various candidate values and choosing the best value that keeps the fusion stable and most accurate. The same fusion parameters are kept consistent across all fusion experiments to maintain fairness in evaluation.

4.5 Evaluation Metrics

Segmentation performance is evaluated using the Dice similarity coefficient (Dice) and Intersection-over-Union (IoU), computed on a per-class basis for the left ventricle (LV), right ventricle (RV), and left ventricular myocardium (MYO).

Let P_c be the predicted class and G_c be the ground truth for class c, respectively. The Dice score for class c is defined as

$$\text{Dice}_c = \frac{2|P_c \cap G_c|}{|P_c| + |G_c|}. \tag{10}$$

Similarly, the Intersection-over-Union (IoU) for class c is defined as

$$\text{IoU}_c = \frac{|P_c \cap G_c|}{|P_c \cup G_c|}. \tag{11}$$

To summarize overall segmentation performance across all anatomical classes, macro-averaged Dice and IoU scores are computed by taking the arithmetic mean over all classes:

$$\text{Dice}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\text{Dice}_c, \tag{12}$$

$$\text{IoU}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\text{IoU}_c, \tag{13}$$

where C denotes the number of foreground classes. Macro-averaging assigns equal importance to each anatomical structure and avoids bias toward larger regions such as the ventricles.

To assess performance stability, the standard deviation (SD) of Dice and IoU scores are also reported across all test cases. This provides insight into the consistency of the segmentation results and highlights potential variability across different anatomical structures and imaging conditions.

The standard deviation of a metric $M \in \{\text{Dice}, \text{IoU}\}$ is computed as

$$\sigma_M = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(M_n - \bar{M})^2}, \tag{14}$$

where M_n denotes the metric computed for the n-th test case, $\bar{M}$ is the mean value, and N is the number of test cases.

4.6 Implementation Details

All models and fusion experiments are implemented in PyTorch. Experiments are conducted on NVIDIA A100 GPUs using Google Colab. For fair comparison, all segmentation outputs are resized to a common spatial resolution before evaluation. The same evaluation strategy is used for all models and fusion configurations.

5. RESULTS

Quantitative performance of each model pairing mentioned in section 4 is reported under both in-domain (M&Ms) and cross-domain (ACDC) evaluation settings using per-class Dice/IoU (LV, RV, MYO) and macro-averaged Dice/IoU. In addition to the main fusion results, we analyze (i) uncertainty-aware DST fusion versus simple probability averaging, (ii) the impact of entropy-guided fallback, and (iii) statistical significance of improvements across cases.

5.1 In-Domain Segmentation Performance on M&Ms

Table 2 reports in-domain results on the M&Ms test set. Among the base models, nnU-Net achieves the strongest overall performance (macro Dice 0.907 ± 0.036, macro IoU 0.835 ± 0.054), almost similar to MedSAM2 (macro Dice 0.900 ± 0.022, macro IoU 0.819 ± 0.036). In contrast, Fine-tuned SAM2 and UNet++ form weaker baselines, with notably lower performance for the RV and MYO classes, reflected in the class-wise Dice heatmap in Fig. 1. Table 2 also shows that DST fusion consistently improves the Dice score over the base models. The highest gain in in-domain settings, is achieved by the fusion of suboptimal-suboptimal pairing (fine-tuned SAM2 + nnU-Net). It can also be seen that DST fusion shows reduced standard deviation (SD) compared to the base models in the in-domain settings, indicating more stability in segmentation performance across the test cases.

MedSAM2 + nnU-Net This is a strong+strong fusion combination. DST fusion improves macro Dice from 0.907 (best base) to 0.924 ± 0.022 and macro IoU from 0.835 to 0.860 ± 0.039, corresponding to relative gains of **+1.87% Dice** and **+2.99% IoU**. Class-wise analysis (Fig. 1, left) shows consistent improvements across all structures, with noticeable gains for RV and MYO. In this case, the scope of improvement is not very large because both models are strong and produce high-confidence, similar outputs. Therefore, conflict is low; DST only refines boundary consistency and reduces prediction variance.

Table 2: **In-domain evaluation** of the proposed DST fusion framework on the M&Ms dataset. Each block reports the mean Dice and IoU (±SD) of base models and the fused result. Relative gain (%) is computed with respect to the best-performing base model within each pairing.

Experiment	Model	Mean Dice	Mean IoU	Rel. Gain (%)
MedSAM2 + nnU-Net (Strong + Strong)	MedSAM2	0.900 ± 0.022	0.819 ± 0.036	–
	nnU-Net	0.907 ± 0.036	0.835 ± 0.054	–
	DST Fusion	**0.924 ± 0.022**	**0.860 ± 0.039**	**+1.87 / +2.99**
Fine-tuned SAM2 + nnU-Net (Suboptimal + Strong)	Fine-tuned SAM2	0.757 ± 0.090	0.626 ± 0.107	–
	nnU-Net	0.907 ± 0.036	0.835 ± 0.054	–
	DST Fusion	**0.911 ± 0.028**	**0.840 ± 0.043**	**+0.44 / +0.60**
SAM2 + UNet++ (Suboptimal + Suboptimal)	UNet++	0.833 ± 0.074	0.725 ± 0.095	–
	Fine-tuned SAM2	0.757 ± 0.090	0.626 ± 0.107	–
	DST Fusion	**0.861 ± 0.053**	**0.764 ± 0.072**	**+3.36 / +5.38**

Fine-tuned SAM2 + nnU-Net This is a suboptimal+strong fusion combination. Fusion yields a smaller but consistent gain over the strongest base model, improving macro Dice from 0.907 to 0.911 ± 0.028 and macro IoU from 0.835 to 0.840 ± 0.043 (**+0.44% Dice**, **+0.60% IoU**). The benefit of DST is very limited in this case over the best base model. Since nnU-Net is a strong model here, it has much higher confidence than lightly fine-tuned SAM2 which acts as a suboptimal model in this case. Therefore, DST follows the stronger model in most cases and produces similar results to nnU-Net. Fig. 1 shows slight improvements by DST in segmenting LV and RV over nnU-Net, however DST heavily relies on nnU-Net for segmenting precise structures like MYO, showing a similar Dice score for MYO. Therefore, in this limited scope, the fusion framework primarily follows the stronger model while incorporating complementary information from the weaker model when both models agree.

Fine-tuned SAM2 + UNet++ This is a suboptimal+suboptimal fusion combination. DST fusion improves macro Dice from 0.833 to 0.861 ± 0.053 and macro IoU from 0.725 to 0.764 ± 0.072, yielding the largest in-domain relative gains among all pairings (**+3.36% Dice**, **+5.38% IoU**). Class-wise results in Fig. 1 show improvements across LV, RV, and MYO. As both models are suboptimally trained here, no model overpowers the other. However, their errors are different. UNet++ can capture precise structures more sharply but struggles with global shape consistency. On the other hand, SAM2 is more expert in capturing global shapes, but produces coarse boundaries. DST takes advantage of these complementary strengths and weaknesses and produces better segmentation masks compared to both base models. This case shows that DST fusion benefits most when there is no dominant predictor and the predictors have different weaknesses.

5.2 Cross-Domain Generalization on ACDC

Table 3 reports cross-domain performance on the ACDC dataset, where models trained on M&Ms are evaluated without any training. Under domain shift, nnU-Net's segmentation performance degrades (macro Dice 0.832 ± 0.138), especially in RV and MYO segmentation, as visible in Fig 1. In contrast, MedSAM2 remains comparatively stable across all classes (macro Dice 0.914 ± 0.024), because of large-scale medical pretraining. DST fusion outperforms the base models in all cases across cross-domain settings. The largest gains can be observed in Suboptimal + Strong pairing. The standard deviation is also lower than both the base models, except in MedSAM2+nnU-Net, where the standard deviation is similar to the best base model-MedSAM2.

MedSAM2 + nnU-Net This is a case of a strong + suboptimal fusion pair. DST fusion improves macro Dice from 0.914 to 0.922 ± 0.027 and macro IoU from 0.846 to 0.858 ± 0.044 (**+0.88% Dice**, **+1.42% IoU** relative gains). Here, MedSAM2 shows strong performance while nnU-Net shows suboptimal performance under domain shift. Class-wise results (Fig. 1, right) show consistent gains across LV, RV, and MYO, with fusion reducing nnU-Net's failures in segmenting RV and MYO. In this case, MedSAM2 overpowers nnU-Net in confidence; therefore, DST mostly follows the stronger model-MedSAM2.

Fine-tuned SAM2 + nnU-Net This is a case of a suboptimal + suboptimal fusion pair. Fusion provides a substantial improvement over the best base model, increasing macro Dice from 0.832 to 0.898 ± 0.044 and

Table 3: **Cross-domain evaluation** of the proposed DST fusion framework on the ACDC dataset. Each block shows the mean Dice and IoU (±SD) of the base models, the fusion result, and the relative gain (%) over the best base model. All base models were trained on M&Ms and directly tested on ACDC (except MedSAM2, which was trained on both M&Ms and ACDC when released).

Experiment	Model	Mean Dice	Mean IoU	Rel. Gain (%)
MedSAM2 + nnU-Net (Strong + Suboptimal)	MedSAM2	0.914 ± 0.024	0.846 ± 0.038	–
	nnU-Net	0.832 ± 0.138	0.740 ± 0.158	–
	DST Fusion	**0.922 ± 0.027**	**0.858 ± 0.044**	+0.88 / +1.42
Fine-tuned SAM2 + nnU-Net (Suboptimal + Suboptimal)	Fine-tuned SAM2	0.808 ± 0.070	0.691 ± 0.091	–
	nnU-Net	0.832 ± 0.138	0.740 ± 0.158	–
	DST Fusion	**0.898 ± 0.044**	**0.821 ± 0.068**	+7.93 / +10.95
SAM2 + UNet++ (Suboptimal + weak)	UNet++	0.362 ± 0.231	0.256 ± 0.190	–
	Fine-tuned SAM2	0.808 ± 0.070	0.691 ± 0.091	–
	DST Fusion	**0.813 ± 0.068**	**0.697 ± 0.087**	+0.62 / +0.87

macro IoU from 0.740 to 0.821 ± 0.068 (**+7.93% Dice**, **+10.95% IoU** relative gain). Here, both the models perform suboptimally, but have different strengths and weaknesses of foundation and task-specific models as discussed previously. DST fusion benefits most in cases like this and hence, the largest improvement is observed. In addition, Fig. 1 demonstrates that this specific DST fusion substantially improves class-wise Dice, especially for RV and MYO.

Fine-tuned SAM2 + UNet++ This is a suboptimal + weak fusion pair. UNet++ collapses under domain shift (macro Dice 0.362 ± 0.231), with severe degradation across all classes. Fine-tuned SAM2 performs substantially better (macro Dice 0.808 ± 0.070). DST fusion yields a modest additional improvement (macro Dice 0.813 ± 0.068, macro IoU 0.697 ± 0.087). In high regions of high uncertainty, fusion avoids UNet++ failures and follows the more reliable predictor instead, which in this case is SAM2. Class-wise Dice in Fig. 1 shows DST improves segmentation of LV and MYO. However, segmentation of RV is slightly worse than the best base model (fine-tuned SAM2). This happens because the RV is harder to segment due to frequently changing shapes across volumes, due to different cardiac phases. Without explicit training, UNet++ fails badly, while SAM2 is better but still not very confident for the RV. When one model is unreliable and the other is uncertain, DST fusion becomes cautious and avoids making strong decisions. As a result, it prefers safer predictions, which can slightly under-segment the RV compared to using SAM2 alone. This shows a limitation of uncertainty-aware fusion when one model fails, and the remaining model is not fully confident for a specific structure.

5.3 Uncertainty-Aware Fusion vs. simple Probability Averaging

To verify that the proposed fusion framework provides advantages beyond simple ensembling, we compare uncertainty-aware DST fusion against simple averaging of probability maps, as shown in Table 4. Given two models with voxel-wise class probability predictions $\mathbf{p}^{(1)}(x)$ and $\mathbf{p}^{(2)}(x)$, simple averaging computes the fused probability for class c at voxel x as

$$\bar{p}_c(x) = \frac{1}{2}\left(p_c^{(1)}(x) + p_c^{(2)}(x)\right), \tag{15}$$

with the final label obtained by $\hat{y}(x) = \arg\max_c \bar{p}_c(x)$. Across all evaluated model pairings, uncertainty-aware DST fusion consistently outperforms this simple probability average baseline. While in-domain relative gain is modest (0.6-1.9%), cross-domain gains are substantially larger, reaching up to 21.7% relative improvement in Dice and 30.5% relative gain in IoU for the most challenging setting (Fine-tuned SAM2 + UNet++ pairing in cross-domain evaluation). These results demonstrate that leveraging inter-model conflict and following uncertainty-guided decision rules yields more reliable predictions than simple probability averaging.

5.4 Effect of Entropy-Guided Fallback

Table 5 evaluates the impact of the entropy-guided fallback mechanism in high-conflict regions. In most settings, disabling fallback reduces performance, with the highest decline observed for the most challenging setting - Fine-tuned SAM2 + UNet++ (suboptimal + weak) under domain shift (macro Dice drops from 0.813 to 0.387, macro

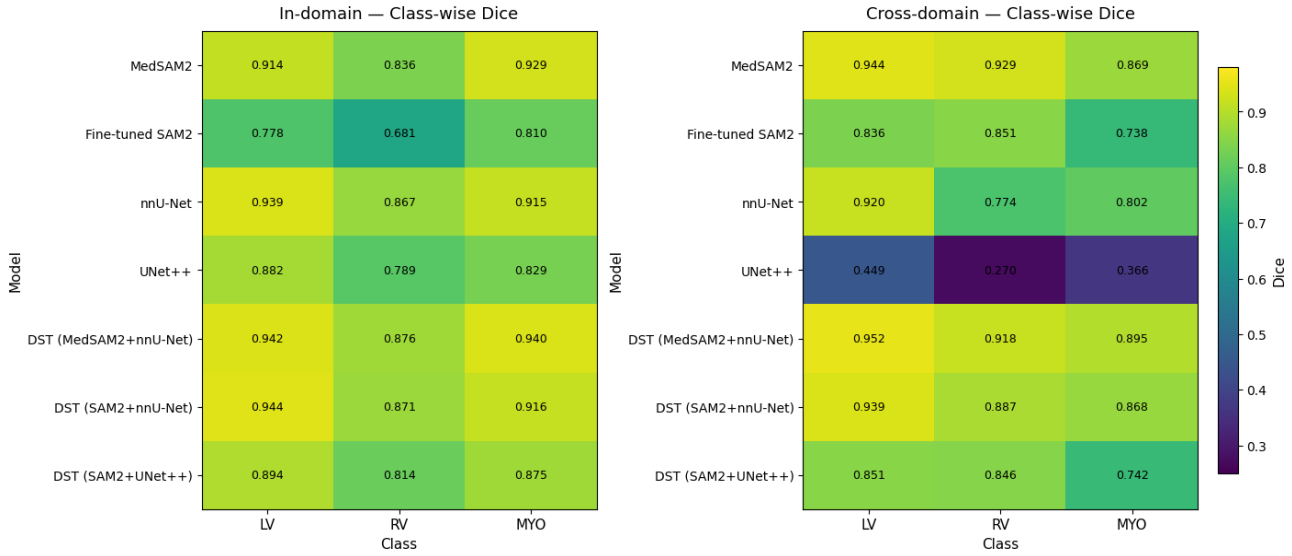


Figure 1: Class-wise Dice scores for LV, RV, and MYO on the M&Ms (left) and ACDC (right) datasets. Results are shown for individual base models and their respective fusions. Fusion consistently improves performance across classes, with the most pronounced gains observed under domain shift on ACDC dataset.

Table 4: Comparison of uncertainty-aware DST fusion against simple probability averaging. Results are macro-averaged Dice/IoU (± SD).

Model Pairing	**Uncertainty-aware DST**		**Average Fusion**	
	Dice	**IoU**	**Dice**	**IoU**
MedSAM2 + nnU-Net (in-domain)	0.924 ± 0.022	0.860 ± 0.039	0.918 ± 0.027	0.851 ± 0.045
MedSAM2 + nnU-Net (cross-domain)	0.922 ± 0.027	0.858 ± 0.044	0.883 ± 0.099	0.807 ± 0.122
SAM2 + nnU-Net (in-domain)	0.911 ± 0.028	0.840 ± 0.043	0.905 ± 0.031	0.832 ± 0.049
SAM2 + nnU-Net (cross-domain)	0.898 ± 0.044	0.821 ± 0.068	0.877 ± 0.098	0.798 ± 0.120
SAM2 + UNet++ (in-domain)	0.861 ± 0.053	0.764 ± 0.072	0.852 ± 0.049	0.750 ± 0.068
SAM2 + UNet++ (cross-domain)	0.813 ± 0.068	0.697 ± 0.087	0.668 ± 0.174	0.534 ± 0.178

IoU from 0.697 to 0.275). This confirms that when the conflict between the models approaches unity, especially when one model becomes unreliable under domain shift, falling back to the more confident model yields better results.

Table 5: Effect of entropy-guided fallback in high-conflict regions. Results are macro-averaged Dice/IoU (± SD).

Model Pairing	**With fallback**		**Without fallback**	
	Dice	**IoU**	**Dice**	**IoU**
SAM2 + nnU-Net (in-domain)	0.911 ± 0.028	0.847 ± 0.042	0.903 ± 0.032	0.825 ± 0.052
SAM2 + nnU-Net (cross-domain)	0.898 ± 0.044	0.821 ± 0.068	0.841 ± 0.119	0.748 ± 0.138
SAM2 + UNet++ (in-domain)	0.851 ± 0.045	0.749 ± 0.062	0.858 ± 0.051	0.760 ± 0.070
SAM2 + UNet++ (cross-domain)	0.813 ± 0.068	0.697 ± 0.087	0.387 ± 0.238	0.275 ± 0.199

5.5 Statistical Significance of Improvements

We further test whether the observed gains are statistically significant[25] using paired comparisons, across cases where DST showed significant improvement over the simple average method (Table 6). For each case, the difference in Dice score between DST fusion and averaging is computed, and statistical tests are performed on these paired differences.

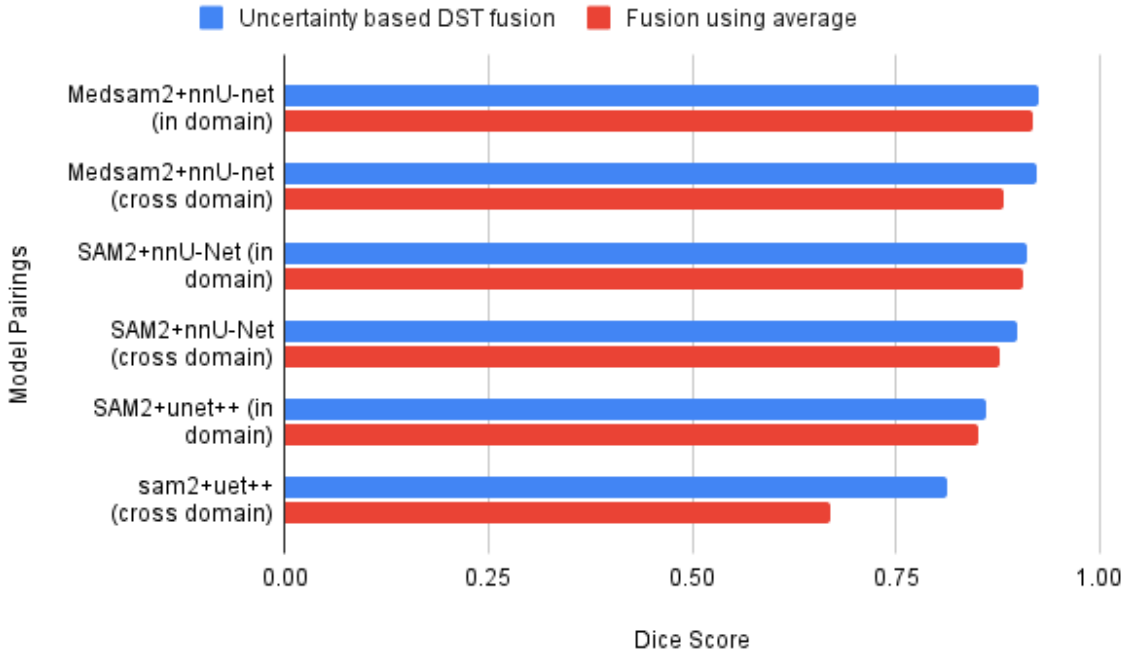


Figure 2: Performance comparison of Dice scores between uncertainty-aware DST fusion and simple probability averaging is illustrated across different model pairings under in-domain and cross-domain settings. DST fusion consistently outperforms averaging, with substantially larger gains under domain shift.

Table 6: Statistical significance of Dice improvements (paired across cases). ΔDice is reported as the median improvement. Holm correction is applied for multiple comparisons.

Model Pairing	n	ΔDice (median)	95% CI	p_{raw}	p_{holm}
MedSAM2 + nnU-Net (cross-domain)	80	0.0113	[0.0054, 0.0190]	6.88×10^{-9}	1.38×10^{-8}
SAM2 + nnU-Net (in-domain)	124	0.0091	[0.0060, 0.0110]	2.24×10^{-15}	8.95×10^{-15}
SAM2 + nnU-Net (cross-domain)	80	0.0033	[0.0006, 0.0058]	2.84×10^{-2}	2.84×10^{-2}
SAM2 + UNet++ (cross-domain)	80	0.0951	[0.0696, 0.1447]	4.11×10^{-13}	1.23×10^{-12}

The p_{raw} values represent the original significance levels obtained from paired tests, without any correction. The p_{holm} values are corrected using the Holm-Bonferroni method to account for multiple comparisons.[26] A result is considered statistically significant when $p_{\text{holm}} < 0.05$. The 95% confidence interval (CI)[27] of the median Dice improvement (ΔDice) indicates the reliability of the observed gain. If the CI does not include zero, the improvement is consistently positive across cases.

Table 6 shows that the improvements are statistically significant for MedSAM2 + nnU-Net (cross-domain) and Fine-tuned SAM2 + nnU-Net (in-domain), with Holm-corrected p-values of 1.38×10^{-8} and 8.95×10^{-15}, respectively. Under the most challenging setting (Fine-tuned SAM2 + UNet++ cross-domain), DST fusion yields a large median Dice improvement ($\Delta\text{Dice} = 0.0951$) with strong statistical significance ($p_{\text{holm}} = 1.23 \times 10^{-12}$). These results confirm that the observed improvements are not due to random variation, but strongly suggest the advantage of uncertainty-aware DST fusion over simple averaging, particularly under domain shift.

5.6 Qualitative Analysis

Figure 4 illustrates entropy and conflict maps for different model pairings. This highlights that uncertainty is prominent near anatomical boundaries which triggers the fallback mechanism during fusion. Figure 5 presents qualitative examples from the ACDC dataset across different model pairings, illustrating that uncertainty-aware DST fusion improves anatomical consistency and mitigates model-specific failures under domain shift.

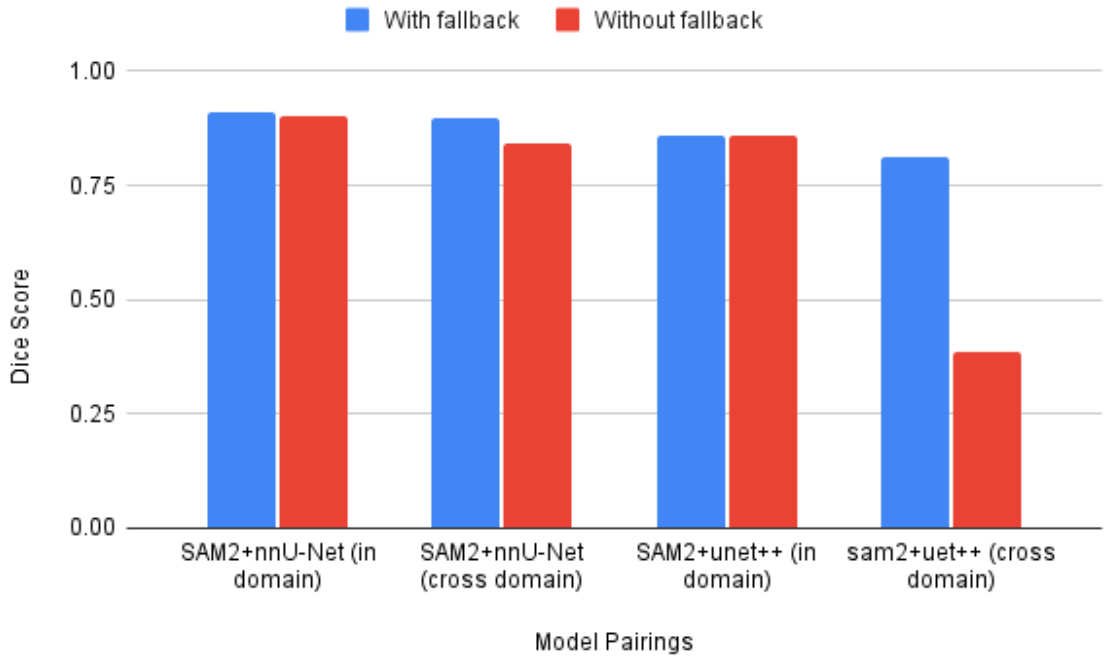


Figure 3: The impact of entropy-guided fallback on fusion performance in terms of Dice score is illustrated. Removing the fallback mechanism leads to performance degradation, particularly under cross-domain conditions, where high inter-model conflict causes unstable fusion. The fallback mechanism prevents catastrophic failures by selecting the model with the lower uncertainty in high-conflict regions.

6. DISCUSSION

This section discusses the experimental findings presented in the Results Section 5 and interprets the observed trend of the proposed fusion framework. The section also highlights when and why Dempster-Shafer Theory based fusion is particularly effective and also discusses how it compares to conventional ensemble strategies. Finally the limitations of the proposed approach and future directions are discussed.

6.1 Overall Trends in Fusion Performance

Across all evaluation settings (in-domain and cross-domain), the proposed DST fusion framework outperforms the best individual model. In-domain improvements on M&Ms datasets were generally modest, whereas cross-domain gains on the ACDC dataset were substantially larger. When models are trained and evaluated on the same data distribution, their predictions are largely consistent, limiting the scope for corrective fusion. However, under domain shift, disagreement and mismatch of confidence increase, which allows the fusion framework to suppress unreliable predictions. This is a core strength of the proposed fusion framework.

The Results (section 5) also highlighted that, compared to simple probability averaging, uncertainty-aware DST fusion achieved better performance in all settings. The benefit of uncertainty-aware DST based fusion over simple probability averaging was most significant under cross-domain settings. This demonstrates that when model reliability varies, simple ensembling collapses. This highlights the importance of leveraging inter-model agreement and conflict in fusion to produce more reliable segmentation masks.

Ablation experiments further demonstrate the importance of the entropy-guided fallback mechanism proposed in the framework in high-conflict areas. Removing the entropy-guided fallback mechanism led to a significant decrease in fusion performance, especially when one model collapsed under domain shift. In addition to ablation study, paired statistical tests also confirmed that the observed gains are statistically significant across pivotal experimental settings. Together, these findings show that the proposed DST fusion is particularly useful in challenging cross-domain conditions.

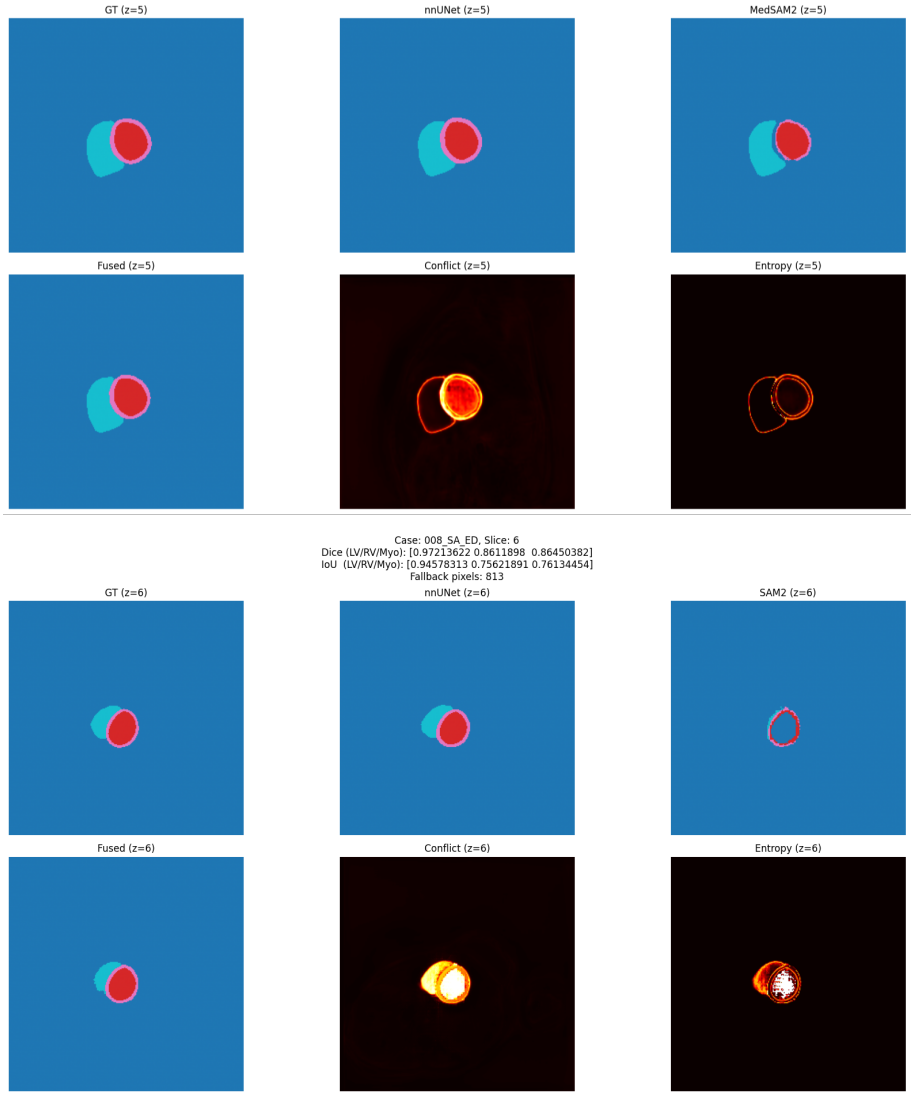


Figure 4: Voxel-wise uncertainty visualization for strong-strong (MedSAM2 + nnU-Net, top) and strong-suboptimal (SAM2 + nnU-Net, bottom) fusion cases. Shown are ground truth, individual model predictions, fused output, conflict maps, and entropy maps computed from the fused probability distribution. The lighter region corresponds to high conflict and entropy, which is prominent around boundaries.

6.2 Effects of Model Strengths and Complementarity

Though the proposed DST fusion framework has consistently shown better performance compared to the base models, it has a limited scope of contribution when both models produce equally high confidence results and have low inter-model conflict. In this case, the proposed fusion framework mainly acts as a refining tool for boundary consistency and the performance improvement of fusion is modest. In cases where one model is highly uncertain, the proposed fusion framework performs limited but safe by following the stronger model. In such cases, the proposed fusion framework produces similar results to the best-performing model. This maintains the performance consistency of the proposed fusion framework, i.e., if it cannot produce a better result, it will at least produce similar results to the best base model. The proposed fusion framework is particularly beneficial when both models have complementary strengths and weaknesses. Results have shown proposed DST based fusion achieved the largest gains in cases where the model pair has balanced disagreements, i.e., neither model consistently dominated, and their errors are different in nature.

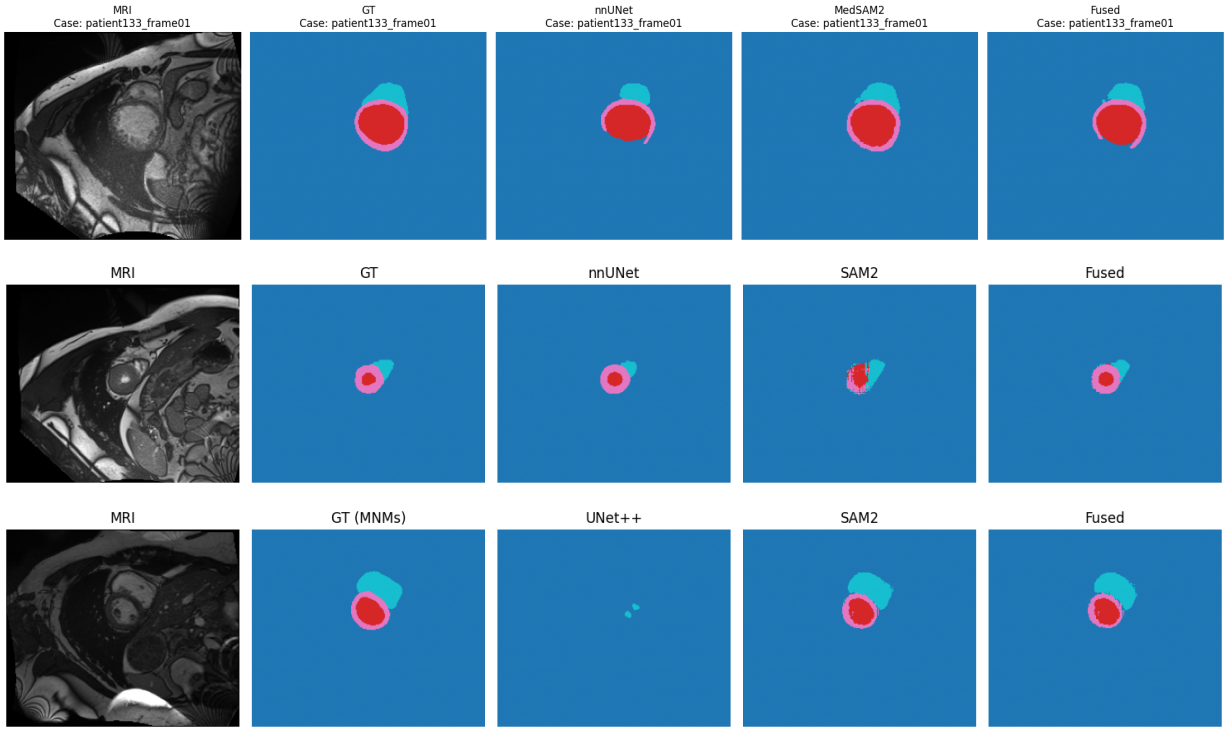


Figure 5: Qualitative comparison of individual models and uncertainty-aware DST fusion on the ACDC dataset. From left to right: input MRI, ground truth, predictions from task-specific model, predictions from foundation models, and DST fused output. The examples illustrate that the fusion improves anatomical consistency and mitigates model-specific failures, particularly under domain shift. The first row, illustrating the fusion of nnU-Net and MedSAM2, shows one of the very few cases where fusion hurts the segmentation output. This happens because of the overconfidence of nnU-Net and can be mitigated by better calibration of nnU-Net probabilities.

6.3 Role of Uncertainty Estimation and Calibration

Uncertainty estimation plays a vital role in the proposed framework for calculating model conflict and confidence. In this study, model uncertainty is quantified by calculating Shannon-Entropy of calibrated probabilities, and model confidence is defined as the maximum calibrated class probability. The optimum value for Temperature is determined using a held-out calibration set, and temperature scaling is applied to each model under each setting. As we have seen that, the performance of the fusion highly depends on the confidence of the models; this calibration step is essential for mitigating the model's overconfidence and ensuring that probability estimates are comparable across models. It is also important to note that entropy derived from softmax probabilities does not capture true epistemic uncertainty. It provides uncertainty in prediction, not uncertainty in models. However, in this study, uncertainty in prediction is sufficient as fusion decisions are made voxel-wise and relies on local uncertainty rather than global model reliability.

6.4 Handling Limitations of Dempster-Shafer Fusion

An inherent limitation of Dempster–Shafer Theory itself is its instability when conflicting belief masses approach unity ($K \approx 1$, i.e., maximum conflict), leading to counter-intuitive behavior. The proposed framework identifies regions of high conflict and avoids fusion by employing an entropy-guided fallback mechanism that chooses the model with lower uncertainty (lower entropy). This mechanism avoids counter-intuitive outcomes by DST. However, in rare cases, where the entropies of both models are equal, imposing a tie, a predefined model is selected. This is done by calculating the Negative log-likelihood (NLL) of the models on the held-out calibration set. The model with lower NLL on the calibration set is selected as the pre-defined model. Thus, the proposed framework detects and avoids unsafe fusion effectively.

6.5 Future directions

In this work, uncertainty is estimated only at the voxel level using prediction entropy. Epistemic uncertainty of the models is not explicitly measured. Incorporating epistemic uncertainty could further strengthen the proposed uncertainty-aware fusion framework by considering global model reliability. This may help resolve high-conflict regions more effectively and reduce counter-intuitive outcomes when Dempster–Shafer fusion becomes unstable.

The proposed framework focuses on the fusion of two segmentation models only. However, DST is not inherently limited to two sources only. While Dempster's rule applies to pairwise combinations, it can be applied iteratively to combine more than two belief masses. Therefore, extending the proposed framework to accommodate more than two segmentation models through iterative fusion can be a promising future direction.

In addition, the current experimentation explores only the Cardiac Cine MRI dataset. Though extensive evaluation on both in-domain and cross-domain datasets is performed, future work can include evaluating different imaging modalities, as well as different anatomical regions, to assess further generalizability. Also, given the model-agnostic nature of the proposed framework, evaluating its performance on the natural image segmentation task is also a potential research direction.

7. CONCLUSION

We present a novel uncertainty-aware DST-based fusion framework with entropy-guided fallback for cardiac MRI segmentation, uniting the complementary characteristics of foundation and task-specific models. By calibrating confidences and resolving disagreements, it enhances reliability in uncertain regions and generalizes across domains. Statistical significance testing confirms that accounting for inter-model conflicts can yield significantly better segmentation performance compared to simple ensemble methods, especially under domain shift. Results show consistent gains with reduced variability, with the largest benefits when compensating weaker models or under domain shift. Ablation studies further validate the benefit of the entropy-guided fallback mechanism, which prevents DST fusion from producing counter-intuitive results. Overall, this work shows that uncertainty-aware DST fusion offers a robust and principled method for merging heterogeneous segmentation models. The framework is model-agnostic, training-free, and low-resource, highlighting its practicality for real-world deployment.

REFERENCES

[1] Caroline Petitjean and Jean-Nicolas Dacher. A review of segmentation methods in short axis cardiac mr images. *Medical image analysis*, 15(2):169–184, 2011.

[2] Avan Suinesiaputra, David A Bluemke, Brett R Cowan, Matthias G Friedrich, Christopher M Kramer, Raymond Kwong, Sven Plein, Jeanette Schulz-Menger, Jos JM Westenberg, Alistair A Young, et al. Quantification of lv function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *Journal of cardiovascular magnetic resonance*, 17(1):63, 2015.

[3] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020.

[4] Peter M Full, Fabian Isensee, Paul F Jäger, and Klaus Maier-Hein. Studying robustness of semantic segmentation under domain shift in cardiac mri. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 238–249. Springer, 2020.

[5] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171:108238, 2024.

[6] Nida M. Zaitoun and Musbah J. Aqel. Survey on image segmentation techniques. *Procedia Computer Science*, 65:797–806, 2015. International Conference on Communications, Management, and Information Technology (ICCMIT'2015).

[7] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, 2018.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[9] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[10] Fabian Isensee, Tassilo Wald, Constantin Ulrich, et al. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[12] Kangxu Fan, Liang Liang, Hao Li, Weijun Situ, Wei Zhao, and Ge Li. Research on medical image segmentation based on sam and its future prospects. *Bioengineering*, 12:608, 2025.

[13] Nyaz Jalal, Małgorzata Śliwińska, Wadim Wojciechowski, et al. Evaluating uncertainty quantification in medical image segmentation: A multi-dataset, multi-algorithm study. *Applied Sciences*, 14(21):10020, 2024.

[14] Alireza Mehrtash, William M Wells, Clare M Tempany, et al. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020.

[15] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[16] Thomas Robin Langerak, Uulke A. van der Heide, et al. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *IEEE Transactions on Medical Imaging*, 29(12):2000–2008, 2010.

[17] Guorong Wu, Qian Wang, Daoqiang Zhang, et al. A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Medical Image Analysis*, 18(6):881–890, 2014.

[18] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[19] Ling Huang, Su Ruan, Pierre Decazes, and Thierry Denoeux. Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. *Information Fusion*, 113:102648, 2025.

[20] Olivier Bernard, Alain Lalande, Clement Zotti, Frédéric Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel A. G. Ballester, et al. Multi-centre, multi-vendor and multi-disease cardiac image segmentation: The m&ms challenge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 68–79. Springer, 2020.

[21] Olivier Bernard, Alain Lalande, Clement Zotti, Frédéric Cervenansky, Xin Yang, Pheng-Ann Heng, Oscar Camara, Miguel A. G. Ballester, Gerard Sanroma, Gemma Piella, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[22] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[23] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[24] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018.

[25] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, 1999.

[26] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[27] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.