Delving Aleatoric Uncertainty in Medical Image Segmentation via Vision Foundation Models −Supplementary Material

Overview

In this supplementary material, we provide a more detailed coverage of related work and a complete derivation of the methodological formulations. Furthermore, we present additional ablation studies and extended comparisons with existing state-of-the-art methods, accompanied by detailed analyses, visualizations, and discussions.

Related Work

Aleatoric Uncertainty Estimation.

Aleatoric uncertainty (data uncertainty) refers to the inherent and irreducible randomness in the data generation process. Typical sources include measurement noise, sensor errors, microscopic material variations, biological stochastic expression, and market fluctuations. Unlike epistemic uncertainty, which can be reduced by acquiring more data, aleatoric uncertainty persists even in the large-sample limit—it represents the residual randomness that cannot be eliminated given the model and data.

To improve the estimation of aleatoric uncertainty, researchers have proposed various schemes. In the field of deep learning, the mainstream approach is to train neural networks to estimate the parameters of heteroscedastic Gaussian distribution and maximize the log-likelihood of observed data. However, although this method is simple to train, it often suffers from variance overconfidence and gradient imbalance problems.Seitzer et al.[15] proposes the β-NLL loss to reweight each sample by the β-power of variance to solve this issue. Some studies focus on combining or optimizing uncertainty types.Kendall et al .[8] propose a Bayesian deep learning framework that integrates input-dependent aleatoric uncertainty and epistemic uncertainty and conducts research on semantic segmentation and depth regression tasks. Its explicit uncertainty formula also introduces a new loss function interpretable as learning decay for tasks, which can reduce the impact of noisy data through implicit decay. Studies such as [10] and [17] continue the main idea by adding parametric variance terms to stochastic prediction models and capturing aleatoric uncertainty with variance decay loss. Other studies start from optimizing the estimation process. Zhang et al.[21] introduces a variance approximation module, assumes that data noise follows a zero-mean distribution, and actively eliminates noise from observations to better restore the underlying signal; Stirn et al.[16] advocates Bayesian treatment of the variance (more precisely, precision) of the predictive distribution; Cui et al.[1] proposes a calibrated regression method based on Maximum Mean Discrepancy (MMD), which enables the prediction interval to have good calibration by minimizing the embedding metric of the kernel. In addition, Yi et al. [20] proposes to train the mean network and variance network separately to obtain aleatoric uncertainty. As uncertainty estimation methods continue to advance, researchers in the medical imaging field are increasingly focusing on leveraging aleatoric and epistemic uncertainty to model complex clinical tasks.

Uncertainty Estimation in Medical Image Segmentation

In medical image segmentation, uncertainty estimation refers to assigning a confidence score to each predicted pixel (or voxel) label that quantifies how likely this label deviates from the true anatomical structure. This score may arise from inherent data noise or ambiguity (aleatoric uncertainty) or from the model's knowledge deficiency caused by inadequate training, distribution shift, or unseen anatomical variations (epistemic uncertainty). By producing an uncertainty map that has the same spatial resolution as the segmentation mask, clinicians can rapidly locate low-confidence regions and give them additional scrutiny during diagnosis and treatment planning, thereby improving both model transparency and clinical safety.

Roy et al.[13] proposed a Bayesian CNN that applies Monte-Carlo Dropout at test time to sample multiple predictions and generate voxel-wise uncertainty maps. Used for whole-brain MRI segmentation, this approach can identify poorly segmented regions and serves as an effective quality-control tool. Jungo et al.[7] compared several uncertainty-estimation methods for brain-tumor segmentation and found that applying Monte-Carlo Dropout after ev-

ery convolutional layer yields more informative uncertainty estimates.By equipping a dropout-based Bayesian CNN for diabetic-retinopathy detection,Lei Biget al.[11] show that MC-dropout uncertainty enables clinically informed decision referral, boosting sensitivity/specificity to NHS-recommended levels while remaining robust across architectures and datasets. Kwon et al.[9] introduces a Bayesian neural network method that decomposes predictive uncertainty into aleatoric and epistemic components without extra variance parameters, demonstrating improved reliability of point predictions on ISLES and DRIVE segmentation datasets. Wang et al.[18] combined test-time augmentation with Monte-Carlo Dropout to estimate uncertainty in MRI segmentation; by performing multiple forward passes and computing prediction variance, their method improves boundary delineation accuracy. Seeböck et al.[14] applied a Bayesian U-Net to retinal OCT segmentation, used Monte-Carlo Dropout to estimate epistemic uncertainty, and observed that high-uncertainty regions strongly coincide with segmentation errors, offering a new route for anomaly detection in medical images.Daming et al.[5] systematically reviewed probabilistic and non-probabilistic uncertainty quantification methods applied to medical image analysis tasks.SU-ASM[19] combines Convolutional Neural Networks (CNN) with Active Shape Models (ASM) to achieve more accurate segmentation results and more reliable uncertainty estimation. Unlike previous studies based on pixel-wise uncertainty, SU-ASM integrates global shape information, effectively reducing local misjudgments and enhancing model interpretability.

In medical imaging uncertainty analysis, most existing studies primarily focus on assessing model epistemic uncertainty and evaluating model reliability. However, the inherent noise and variability in medical data directly influence the learning process of deep models, often becoming a critical bottleneck for achieving stable and accurate predictions. Motivated by this gap, our work shifts attention to quantifying the intrinsic aleatoric uncertainty of the data itself and further proposes a data optimization framework driven by aleatoric uncertainty awareness. Experimental results validate the effectiveness of this idea, demonstrating that explicitly modeling data-level stochastic uncertainty can lead to more robust and reliable medical image segmentation.

Method details

Singular Value Spectrum Analysis of vectors

From the perspective of linear algebra and representation learning, singular values characterize the variance energy distribution of a feature matrix along different orthogonal directions, serving as a key indicator of structural complexity and redundancy within the data. The magnitude of each singular value reflects the information energy or the variation strength carried along its corresponding direction, and

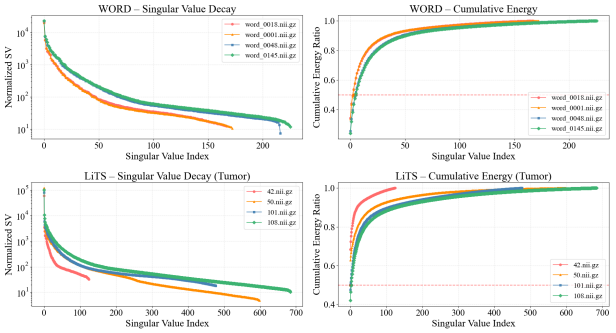

Figure 1. Singular value decay (left) and cumulative energy (right) for four WORD and LiTS samples.

the overall shape of the singular value spectrum directly reveals the richness and anisotropy of the learned vectors.

Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ denote the singular values of a feature matrix $X \in \mathbb{R}^{m \times n}$, where $r = \text{rank}(X)$. The **energy concentration** can be quantified by the ratio of the cumulative sum of the top-k singular values to the total energy:

$$E(k) = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2}$$

For a sample with **fast decay** (e.g., red/orange curves), there exists a small k_0 such that:

$$E(k_0) \geq 0.9$$

This indicates that over 90% of the total energy is captured by only k_0 components, implying that the effective rank k_0 is small. In such cases, the feature representation is **low-rank**, and the **effective dimensionality** is limited:

$$\text{Effective Rank} \approx k_0 \ll r$$

Conversely, for a sample with **slow decay** (e.g., green/blue curves), the energy is more uniformly distributed. The number of singular values required to capture a significant portion of the energy is much larger:

$$E(k) < 0.9 \quad \text{for} \quad k \gg k_0$$

This implies a higher **effective rank** and more diverse feature directions:

$$\text{Effective Rank} \approx k_{\text{slow}} \gg k_0$$

Figure 1 illustrates the singular value characteristics of four representative samples from LiTs and WORD datasets. The left panel shows the Singular Value Decay Curves, while the right panel presents the Cumulative Energy Distributions. As shown, the red and orange curves correspond to the two samples with the fastest decay of the singular value. Their singular values rapidly drop to near zero after the first

few components, and the cumulative energy ratio exceeds 90% within only a few dominant directions. This indicates that the variance information is highly concentrated and that the feature matrices can be well approximated by a low-rank structure. Such low-rank representations exhibit strong internal correlations and redundancy, which implies that the effective dimensionality of the feature space is reduced. Consequently, the information channels available to the model become narrower, and perturbations such as noise, blurred boundaries, or annotation inconsistencies are more likely to be amplified, thereby increasing aleatoric uncertainty (data uncertainty).

In contrast, the blue and green curves correspond to the two samples with the slowest decay of the singular value. Their energy is more evenly distributed, the decay process is gradual, and the cumulative energy curves rise more slowly, suggesting that the variance information is dispersed across a larger number of orthogonal directions. In this case, the feature matrices are closer to full-rank, with higher information entropy and redundancy. The model can thus extract more robust and complementary representations from multiple directions, making it less sensitive to noise or parameter perturbations.

Necessary Conditions for Dynamic Uncertainty-aware Optimization.

Consider the optimization problem of the total loss function $\mathcal{L}_{\text{total}}$ of DUO:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{\mathcal{L}_{\text{seg}}(f(x_i)_{\theta_1}, y_i - \hat{\epsilon}_i \cdot f(x_i)_{\theta_2})}{1+\alpha\cdot\mathcal{S}(f(x_i|c)_{\theta_1})}, \\ \text{s.t.} \quad & \frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i = 0, \quad \frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i^2 = 1. \end{aligned} \tag{1}$$

subject to:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i = 0, \quad \frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i^2 = 1. \tag{2}$$

To handle the constraints, we introduce the Lagrange multipliers λ_1 and λ_2 and formulate the Lagrangian function [21] as follows:

$$\begin{aligned} \mathcal{L}_{\text{total}} = &\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{\mathcal{L}_{\text{seg}}(f(x_i)_{\theta_1}, y_i - \hat{\epsilon}_i \cdot f(x_i)_{\theta_2})}{1+\alpha\cdot\mathcal{S}(f(x_i|c)_{\theta_1})} \\ &+\lambda_1\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i\right) + \lambda_2\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i^2 - 1\right). \end{aligned} \tag{3}$$

During the optimization process, when the parameters reach a local minimum, the partial derivatives of the Lagrangian with respect to all variables (including $\hat{\epsilon}$, λ_1, and λ_2) should vanish. We first focus on the partial derivative with respect to the estimated noise $\hat{\epsilon}$.

Partial Derivative Conditions for $\hat{\epsilon}$. For each $i = 1, 2, \ldots, N$, compute $\frac{\partial L}{\partial \hat{e}_i} = 0$:

$$\frac{\partial L}{\partial \hat{e}_i} = \frac{\partial L_{\text{total}}}{\partial \hat{e}_i} + \frac{\lambda_1}{N} + \frac{2\lambda_2}{N}\hat{e}_i = 0.$$

Here, $\frac{\partial L_{\text{total}}}{\partial \hat{e}_i}$ needs to be computed explicitly. According to the chain rule:

$$\frac{\partial L_{\text{total}}}{\partial \hat{e}_i} = \frac{1}{N}\sum_{c=1}^{C}\frac{\partial}{\partial \hat{e}_i}\left[\frac{L_{\text{seg}}(f(x_i)_{a_i}, y_i - \hat{e}_i \cdot f(x_i)_{a_i})}{1+\alpha\cdot S(f(x_i|c)_{b_i})}\right].$$

Let $z_i = y_i - \hat{e}_i \cdot f(x_i)_{a_i}$, then:

$$\frac{\partial L_{\text{seg}}}{\partial \hat{e}_i} = \frac{\partial L_{\text{seg}}}{\partial z_i}\cdot\frac{\partial z_i}{\partial \hat{e}_i} = \frac{\partial L_{\text{seg}}}{\partial z_i}\cdot(-f(x_i)_{a_i}).$$

Therefore,

$$\frac{\partial L_{\text{total}}}{\partial \hat{e}_i} = -\frac{1}{N}\sum_{c=1}^{C}\frac{f(x_i)_{a_i}\cdot\frac{\partial L_{\text{seg}}}{\partial z_i}}{1+\alpha\cdot S(f(x_i|c)_{b_i})}.$$

Substituting into the partial derivative condition:

$$-\frac{1}{N}\sum_{c=1}^{C}\frac{f(x_i)_{a_i}\cdot\frac{\partial L_{\text{seg}}}{\partial z_i}}{1+\alpha\cdot S(f(x_i|c)_{b_i})} + \frac{\lambda_1}{N} + \frac{2\lambda_2}{N}\hat{e}_i = 0.$$

Multiplying both sides by N and simplifying:

$$-\sum_{c=1}^{C}\frac{f(x_i)_{a_i}\cdot\frac{\partial L_{\text{seg}}}{\partial z_i}}{1+\alpha\cdot S(f(x_i|c)_{b_i})} + \lambda_1 + 2\lambda_2\hat{e}_i = 0. \tag{4}$$

Partial Derivatives for Constraint Conditions. The partial derivatives of the Lagrangian function with respect to λ_1 and λ_2 directly yield the constraint conditions:

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = \frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda_2} = \frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i^2 - 1 = 0. \tag{5}$$

These conditions ensure that the estimated noise has zero mean and unit variance.

Solution and Interpretation of the Necessary Conditions. Equation 4 holds for each i, forming a system of equations. To solve for $\hat{\varepsilon}_i$, we can rewrite Equation 4 as:

$$2\lambda_2\hat{\varepsilon}_i = \sum_{c=1}^{C}\frac{f(x_i)_{a_i}\cdot\frac{\partial L_{\text{seg}}}{\partial z_i}}{1+\alpha\cdot S(f(x_i|c)_{b_i})} - \lambda_1. \tag{6}$$

This indicates that the estimated noise $\hat{\varepsilon}_i$ is proportional to the gradient of the segmentation loss $\frac{\partial L_{\text{seg}}}{\partial z_i}$, modulated by the uncertainty-aware scale $S(\cdot)$. Specifically: (1) When $S(f(x_i|c)_{b_i})$ is large (i.e., the model exhibits high representation ability for class c), the denominator increases, thereby reducing the adjustment magnitude of $\hat{\varepsilon}_i$. This helps maintain model stability on uncertain classes. (2) The constraint conditions ensure the overall statistical properties of $\hat{\varepsilon}_i$, preventing the noise estimates from deviating significantly from the actual distribution.

Statistical Properties of Noise Estimation. From the constraint conditions $\sum \hat{\epsilon}_i = 0$ and $\sum \hat{\epsilon}_i^2 = N$, we can derive the expected behavior of the noise estimates. Assuming that the true noise ϵ_i has zero mean and variance σ_ϵ^2, the estimated noise $\hat{\epsilon}_i$ is normalized through constrained optimization. This helps separate label noise from model prediction errors. In the asymptotic limit as $N \to \infty$, the sample distribution of the estimated noise $\hat{\epsilon}_i$ converges to a distribution with zero mean and unit variance. This ensures stability in the denoising process and reduces errors introduced by model epistemic bias. In addition, the effectiveness of this optimization constraint is verified in the following experiments.

Dateset details

LiTS: A liver tumor segmentation dataset comprising CT scans. It includes annotations for two semantic categories: liver and liver tumors, with a total of 131 annotated cases. The original data has an average volume size of $432 \times 512 \times 512$ voxels, an average voxel spacing of $1.0 \times 0.76 \times 0.76$ mm, and a pixel intensity range of [-983, 5420] Hounsfield Units (HU).

TotalSegmentator: A comprehensive multi-organ segmentation dataset comprising 3D CT scans. It includes annotations for 104 anatomical structures across major anatomical systems: abdominal organs (spleen, kidneys, gallbladder, liver, stomach, pancreas, adrenals, intestines), thoracic structures (lungs, heart chambers, great vessels, trachea, esophagus), spinal vertebrae (cervical C1-C7, thoracic T1-T12, lumbar L1-L5), musculoskeletal components (ribs, clavicles, scapulae, humeri, pelvis, femora, sacrum), vascular system (aorta, venae cavae, portal vein, iliac vessels, pulmonary artery), and other structures (brain, facial bones, bladder, myocardium, paraspinal muscles). The dataset contains 1,138 annotated cases with an average volume size of $116 \times 116 \times 120$ voxels, average voxel spacing of $3.0 \times 3.0 \times 3.0$ mm, and intensity range of [-1007, 1241] Hounsfield Units (HU).

FeTA: A multi-organ segmentation dataset comprising 3D MR modality data. It includes seven semantic categories: external cortex, gray matter, white matter, ventricles, cerebellum, deep gray matter, and brainstem, with a total of 80 samples. The original data has an average size of $180 \times 512 \times 512$, mean voxel spacing of $2.5 \times 0.81 \times 0.81$ mm, and intensity range of [-1024, 3071] HU.

WORD: A multi-organ segmentation dataset containing 3D CT modality data. It encompasses 16 semantic categories: liver, spleen, left kidney, right kidney, stomach, gallbladder, esophagus, pancreas, duodenum, colon, small intestine, adrenal glands, rectum, bladder, left femur, and right femur, with 120 total samples. The original data features an average size of $210 \times 512 \times 512$, mean voxel spacing of $3.0 \times 0.97 \times 0.97$ mm, and intensity range of [-3024, 3071] HU.

KiTS23: A kidney tumor segmentation dataset with 3D CT modality data. It contains three semantic categories: kidney, kidney tumor, and renal cyst, comprising 489 samples. The original data maintains an average size of $104 \times 512 \times 512$, mean voxel spacing of $3.0 \times 0.78 \times 0.78$ mm, and intensity range of [-1022, 3071] HU.

Supplementary Experiments

Combination of aleatoric uncertainty-aware data filter and dynamic uncertainty-aware optimization.

We present supplementary experiments on the joint implementation of Aleatoric Uncertainty-aware Data Filtering (AUDF) and Dynamic Uncertainty-aware Optimization (DUO) strategies. In the experimental setup, nnU-Net serves as the baseline network, while AUDF utilizes MedSAM2 as its feature extractor, employing the mean score across all classes as the final Aleatoric Uncertainty Value (AUV). Based on AUDF's pruning of 5% and 10% of noisy data, we subsequently apply DUO to train and evaluate the baseline network (AUDF+DUO). This constitutes a two-stage strategy that incurs increased training overhead. As shown in Table 1, the performance comparison of the aforementioned approaches across five distinct segmentation datasets demonstrates that incorporating DUO yields consistent performance improvements under different data pruning ratios. Combined with Table 2 in the manuscript, the aforementioned results indicate that after noise pruning in the training data, the performance gains achieved by DUO become relatively modest. For example, incorporating DUO into the baseline network yields an improvement of up to 1.17% on the LiTS dataset, while adding DUO on top of AUDF leads to an enhancement of up to 0.83%. This behavior can be attributed to the fact that the filtered noisy samples contain instances with annotation errors. As the overall quality of the training data increases, the effectiveness of the DUO mechanism designed to mitigate the model's sensitivity to label noise naturally becomes less pronounced.

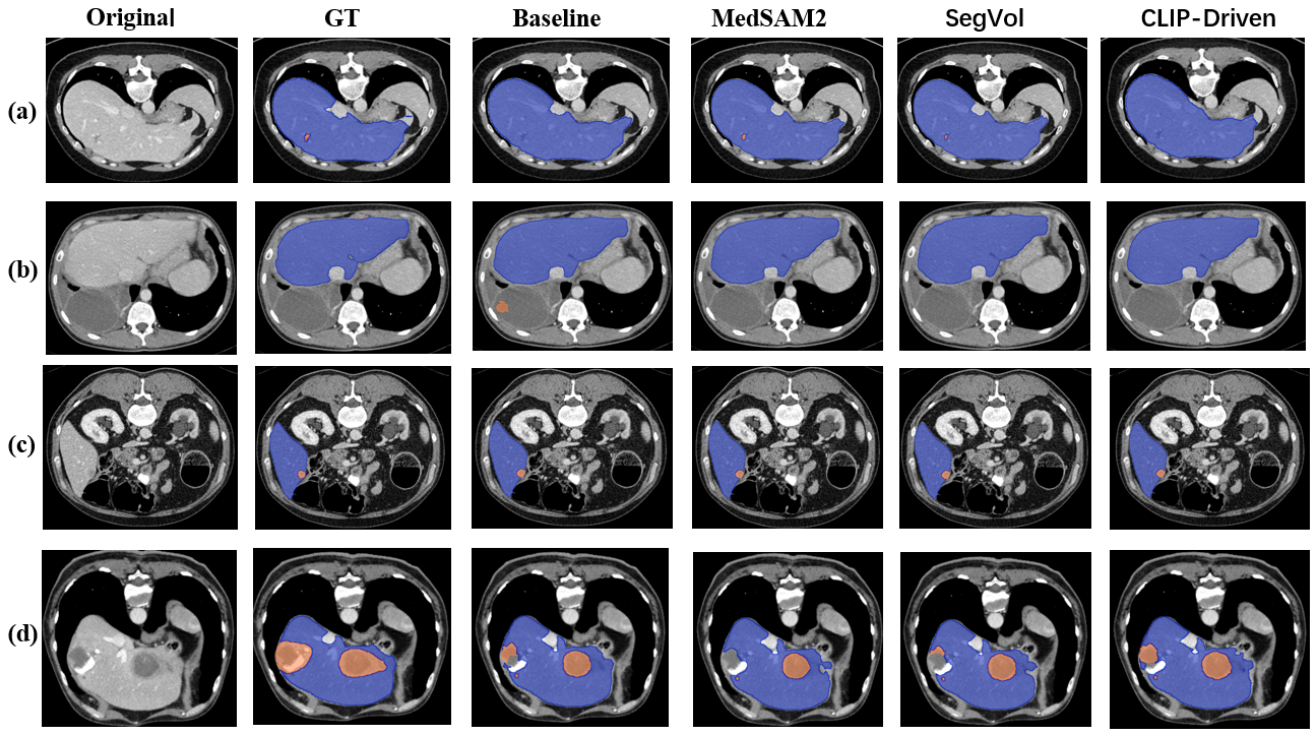

Figure 2. Qualitative results of the proposed aleatoric uncertainty-aware data filtering strategy on the LiTS dataset. Columns from left to right represent: original image, Ground Truth (GT), baseline model prediction, and predictions from models retrained after data filtering using different foundation models (MedSAM2, SegVol, CLIP-Driven).

Table 1. The Dice scores achieved by the joint aleatoric uncertainty-aware data filtering and dynamic uncertainty-aware optimization strategy on five datasets.

Method	LiTS	TotalSeg	WORD	FeTA	KiTS23
Baseline	82.96	80.34	79.03	84.30	70.15
Remove 5% of noisy samples.					
AUDF	83.31	81.31	80.27	84.72	70.43
AUDF + DUO	84.15	81.54	81.16	85.13	70.67
Remove 10% of noisy samples.					
AUDF	85.45	81.65	81.74	84.59	70.57
AUDF + DUO	85.67	81.84	81.76	84.69	70.76

Supplementary Analysis of Aleatoric Uncertainty-aware Data Filtering.

Qualitative Analysis of Aleatoric Uncertainty-Aware Data Filtering. In the experiments conducted on the LiTS dataset, we systematically evaluated the impact of aleatoric uncertainty-aware data filtering strategies based on various medical vision foundation models on segmentation performance. Qualitative results are shown in Figure 2. The visualizations demonstrate that the proposed method exhibits significant advantages across multiple typical scenarios.Figure 2(a) While the baseline model struggles to effectively segment small-sized tumors, models retrained after uncertainty-aware data filtering (e.g., MedSAM2, SegVol) clearly delineate the tumor regions, indicating that the filtering mechanism enhances the model's ability to perceive fine structures;Figure 2(b) The baseline model produces obvious false positive segmentations in certain areas, whereas the model trained with our data filtering strategy successfully avoids such mis-segmentation. This further confirms that difficult samples can misguide the model's attention toward tumor regions, and filtering out such samples helps improve the model's discriminative capability;Figure 2(c) For some typical cases, both the baseline and filtered models achieve relatively accurate segmentation, indicating that all methods perform well when sample quality is high;Figure 2(d) When segmenting large and structurally complex tumors, the baseline model and filtering strategies based on certain foundation models (e.g., MedSAM2, SegVol) fail to achieve satisfactory results. In contrast, the CLIP-Driven-based uncertainty-aware filtering approach performs better in such scenarios, demonstrating stronger robustness in handling highly uncertain samples.The experimental results show that the proposed aleatoric uncertainty-aware data filtering mechanism can effectively identify and remove potentially noisy samples in the training set, thereby improv-

Table 2. Direct Data Filtering Using Prediction Dice Scores.

Strategy	LiTS		WORD	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
Baseline	82.96	73.24	79.03	69.85
95%	82.68↓	72.99↓	80.70↑	71.29↑
90%	82.12↓	72.49↓	81.72↑	72.19↑

Table 3. The performance of different data filtering rates.

strategy	data	LiTS		WORD	
		Dice	mIoU	Dice	mIoU
Baseline	100 %	82.96	73.24	79.03	69.85
SegVol	95%	84.51	74.62	80.12	70.95
	90%	83.73	74.03	81.52	72.01
	85%	83.12	73.34	81.78	72.26
MedSAM2	95%	83.31	73.55	80.27	71.04
	90%	85.45	75.47	81.74	72.18
	85%	84.51	74.61	81.67	72.04

ing segmentation accuracy and generalization across samples of varying difficulty levels.

Limitations of Using Model Predictions as Aleatoric Uncertainty. Model predictions deeply integrate the dual influences of the model's current state and annotation quality during computational processing. Using Dice scores directly as filtering criteria will inevitably eliminate challenging samples that are crucial for enhancing model generalization, while retaining some defective samples with significant annotation noise that the model happens to fit. As shown in Table 2, the performance of this approach aligns with the theoretical analysis, demonstrating that it performs poorly compared to the three common aleatoric uncertainty quantification strategies discussed in the manuscript.

Analysis of Data Filtering Rates. Additional ablation studies were conducted on the data filtering ratio for both tumor and multi-organ segmentation tasks. Under a consistent experimental setup utilizing foreground-normalized images as input and the mean score of all classes as the final AUV, Table 3 presents a comparative analysis of segmentation performance when retaining 95%, 90%, and 80% of the training data. The results indicate that a filtering ratio of 90% leads to more stable performance improvements; however, the optimal data retention ratio may vary depending on the specific requirements of different segmentation datasets.

Analysis of inference resources consumed by different visual foundation models. To evaluate the practical efficiency of different visual foundation models, we compared the parameter count, GPU memory usage during single-sample inference, and inference time of MedSAM2, SegVol, and CLIP-Driven under the same hardware environment, as shown in Table 4. From an efficiency perspective, MedSAM2 demonstrates significant advantages, it has the lowest parameter count and consumes considerably less GPU memory during inference compared to SegVol and CLIP-Driven. This indicates that MedSAM2 is a more lightweight and memory-friendly model, with strong potential for deployment. In contrast, CLIP-Driven is a category-dependent model, and its memory usage increases with the number of categories.

Table 4. Comparison of three models' parameters, GPU memory usage, number of classes, and inference time on the same sample.

Method	**Classes**	**Param (M)**	**Mem (MiB)**	**Time (s)**
MedSAM2	2	38.96	1559 / 24564	7.89
SegVol	2	180.89	3517 / 24564	7.67
CLIP-Driven	2	62.60	5607 / 24564	5.98
CLIP-Driven	7	62.60	10344 / 24564	12.43

Effectiveness of AUDF in Different Backbone Networks. The manuscript compares performance improvements using nnU-Net as the baseline network. To validate the applicability of the Aleatoric Uncertainty-aware Data Filtering (AUDF) strategy to other baseline networks, we conducted additional experiments on Swin-UNETR and U-Mamba. To ensure a fair experimental comparison, we uniformly selected MedSAM2 as the feature extractor, used foreground-normalized images as input, adopted the mean score of all classes as the final AUV, and excluded 10% of noisy training data. As shown in Table 5, the AUDF can be seamlessly transferred to the training of other baseline networks, providing consistent performance improvements.

Supplementary Analysis of Dynamic Uncertainty-aware Optimization.

Comparison between DUO and Static Re-weighting Methods. Under a unified experimental setup, MedSAM2 is selected as the feature extractor, with foreground-normalized images as input, and the mean score of all classes served as the final AUV. The experiment statically re-weighting the training loss using precomputed uncertainty scores for each sample [2, 3]. Specifically, uncertainty can be computed at both the class level and the sample level, corresponding to inter-class and inter-sample reweighting of the loss, respectively. As illustrated in Table 6, comprehensive comparative experiments conducted on both tumor and multi-organ segmentation datasets demonstrate the effectiveness of the static reweighting strategy based on uncertainty scores. However, the performance improvements achieved by DUO are more pronounced and stable. This can be attributed to DUO's design, which enables dynamic adjustment of per-class weights for each sample during every training iteration. Such dynamic and fine-grained modulation substantially enhances the model's learning efficiency.

Table 5. Performance of the Aleatoric Uncertainty-aware Data Filtering (AUDF) strategy on three baseline networks.

Method	AUDF	LiTS		TotalSeg		WORD		FeTA		KiTS23	
		Dice(%)	mIoU(%)	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
nnU-Net [6]	×	82.96	73.24	80.34	72.71	79.03	69.85	84.30	74.63	70.15	63.76
	✓	85.45↑	75.47↑	81.65↑	73.99↑	81.74↑	72.18↑	84.59↑	74.82↑	70.57↑	63.99↑
Swin-UNETR [4]	×	80.54	71.10	77.38	70.03	77.26	68.26	84.15	74.48	66.89	60.79
	✓	83.52 ↑	73.73 ↑	78.13↑	70.71↑	78.35↑	69.24↑	84.33↑	74.66 ↑	67.78↑	61.60 ↑
U-Mamba [12]	×	81.24	71.72	79.27	71.74	78.93	69.76	84.42	74.72	67.85	61.67
	✓	84.89↑	74.94 ↑	81.25↑	73.52↑	79.57↑	70.34 ↑	84.87↑	75.13↑	68.88 ↑	62.61 ↑

Table 6. Performance comparison between DUO and static re-weighting method.

method	LiTS		WORD	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
Baseline	82.96	73.24	79.03	69.85
Class	83.26	73.50	78.77	69.62
Sample	83.57	73.78	79.28	70.07
DUO	**84.13**	**74.73**	**79.63**	**70.38**

Table 7. The necessity of conditional optimization in DUO.

method	LiTS		WORD	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
Baseline	82.96	73.24	79.03	69.85
RW	83.98	74.11	79.52	70.25
DUO−	83.87	74.04	79.41	70.18
DUO	**84.13**	**74.73**	**79.63**	**70.38**

Necessary Conditions for Dynamic Uncertainty-aware Optimization. Treating $\hat{\epsilon}$ solely as a learnable tensor yields suboptimal performance and lacks interpretability. Therefore, we conducted ablation studies by removing the constraints on $\hat{\epsilon}$ in the DUO (DUO−), with results summarized in Table 7. The performance deteriorates substantially after eliminating these constraints, falling below even the baseline strategy that only employs re-weighting (RW). This observation provides compelling evidence for the effectiveness of imposing constraints on the learnable $\hat{\epsilon}$ parameter.

To evaluate the practical efficiency of different visual foundation models, we compared the parameter count, GPU memory usage during single-sample inference, and inference time of MedSAM2, SegVol, and CLIP-Driven under the same hardware environment, as shown in Table 4. From an efficiency perspective, MedSAM2 demonstrates significant advantages: it has the lowest parameter count and consumes considerably less GPU memory during inference compared to SegVol and CLIP-Driven. This indicates that MedSAM2 is a more lightweight and memory-friendly model, exhibiting strong potential for deployment. In contrast, both CLIP-Driven and SegVol are category-dependent models, and their memory usage increases with the number of categories, with CLIP-Driven showing a particularly notable effect.

References

[1] Peng Cui, Wenbo Hu, and Jun Zhu. Calibrated reliable regression using maximum mean discrepancy. *Advances in Neural Information Processing Systems*, 33:17164–17175, 2020. 1

[2] Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. Learning sample difficulty from pre-trained models for reliable prediction. *Advances in Neural Information Processing Systems*, 36:25390–25408, 2023. 6

[3] Peng Cui, Guande He, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. Exploring aleatoric uncertainty in object detection via vision foundation models. *arXiv preprint arXiv:2411.17767*, 2024. 6

[4] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–426. Springer, 2023. 7

[5] Ling Huang, Su Ruan, Yucheng Xing, and Mengling Feng. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223, 2024. 2

[6] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 7

[7] Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019. 1

[8] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 1

[9] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using

bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020. 2

[10] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 1

[11] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017. 2

[12] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 7

[13] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019. 1

[14] Philipp Seeböck, José Ignacio Orlando, Thomas Schlegl, Sebastian M Waldstein, Hrvoje Bogunović, Sophie Klimscha, Georg Langs, and Ursula Schmidt-Erfurth. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE transactions on medical imaging*, 39(1):87–98, 2019. 2

[15] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022. 1

[16] Andrew Stirn and David A Knowles. Variational variance: Simple, reliable, calibrated heteroscedastic noise variance parameterization. *arXiv preprint arXiv:2006.04910*, 2020. 1

[17] Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516. IEEE, 2022. 1

[18] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 2

[19] Bing Yang, Xiaoqing Zhang, Huihong Zhang, Sanqian Li, Risa Higashita, and Jiang Liu. Structural uncertainty estimation for medical image segmentation. *Medical Image Analysis*, page 103602, 2025. 2

[20] Jiaxiang Yi and Miguel A Bessa. Cooperative bayesian and variance networks disentangle aleatoric and epistemic uncertainties. *arXiv preprint arXiv:2505.02743*, 2025. 1

[21] Wang Zhang, Ziwen Martin Ma, Subhro Das, Tsui-Wei Lily Weng, Alexandre Megretski, Luca Daniel, and Lam M Nguyen. One step closer to unbiased aleatoric uncertainty estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16857–16864, 2024. 1, 3