Uncertainty-Aware Bayesian Prompt Adaptation for Robust Cross-Modality Medical Segmentation

Sakang Hong[1,*], Jun-Young Kim[2,*], and Kyungsu Lee[1]

[1] Jeonbuk National University, Jeonju, South Korea
[2] Daegu Catholic University Hospital, Daegu, South Korea

Abstract. Segmentation foundation models such as SAM2 generalize well to natural images, yet adapting them to heterogeneous medical modalities remains challenging due to severe domain shifts and scarce annotations. We propose BayesPrompt, a Bayesian prompt adaptation framework for few-shot cross-modality medical segmentation. BayesPrompt combines Bayesian Meta-Prior Adaptation (BMPA), which regularizes lightweight decoder updates via source–target posterior alignment, with a Probabilistic Prompt Module (PPM) that encodes class-wise feature statistics and predictive uncertainty into attention-conditioned prompt tokens. This design supports both gradient-free fast prompting and efficient Bayesian fine-tuning while mitigating overfitting and improving calibration. We evaluate BayesPrompt on ultrasound and MRI rotator cuff tear segmentation under single- and cross-modality transfer with $k \in \{1, 3, 5, 10\}$ labeled target samples. BayesPrompt consistently improves Dice performance and robustness, demonstrating data-efficient adaptation of segmentation foundation models for medical imaging.

Keywords: Few-shot · Cross-modality adaptation · Bayesian prompt

1 Introduction

Medical image segmentation is a core enabling technology for quantitative diagnosis, treatment planning, and longitudinal assessment. Despite strong progress in task-specific supervised models, ranging from U-Net-style CNNs [20] to transformer-based designs such as SegFormer [26], robust generalization remains difficult when the test distribution deviates from the training distribution. In real-world clinical deployment, such distribution shifts arise from variations in imaging devices, acquisition protocols, patient populations, and, most prominently, imaging modalities. This limitation is especially pronounced in *cross-modality* settings (e.g., ultrasound (US) versus magnetic resonance imaging (MRI)), where acquisition physics, intensity statistics, spatial resolution, and noise characteristics differ substantially. Consequently, models trained on one modality often fail catastrophically when deployed on another, motivating cross-modality adaptation, representation alignment, and modality-specific normalization strategies [5, 7, 11, 24, 29].

[*] These authors contributed equally.

Recently, foundation models have reshaped computer vision by enabling promptable, general-purpose segmentation. Segment Anything Model (SAM) [13] and its successor SAM2 [19] demonstrate remarkable generalization in natural imagery and provide a convenient interface for interactive/automatic segmentation. Early attempts to transfer such models to medical imaging report promising potential but also highlight non-trivial domain gaps and reliability concerns [16]. Parameter-efficient adaptation modules (e.g., adapters) have therefore been explored to bridge performance gaps without full fine-tuning [3, 25]. In parallel, few-shot SAM adaptation for anatomical segmentation has been studied [27], and cross-modality SAM fine-tuning variants have been proposed for multimodal MR settings [21]. Nevertheless, a critical challenge persists: *few-shot cross-modality adaptation* must simultaneously avoid overfitting (due to scarce labels) and remain reliable under severe modality shift. Test-time adaptation (TTA) provides another practical direction by adapting models online at inference time, but medical-domain TTA methods can be sensitive to distribution shift and optimization instability, particularly in cross-domain/cross-modality scenarios [23, 28]. Cache-driven and spatial TTA mechanisms further improve stability, yet they do not directly address predictive uncertainty or calibration, both of which are central to clinical trust [15]. Prior work in Bayesian deep learning and uncertainty quantification has established that separating epistemic and aleatoric uncertainty is essential for trustworthy predictions [8]. Practical approaches include dropout-based Bayesian approximations [6] and Bayesian layers [22]. In medical image segmentation, calibration and uncertainty estimation have been studied extensively [14, 17], yet these ideas are not commonly integrated into prompt-based adaptation of segmentation foundation models under cross-modality few-shot constraints.

In this work, we study *few-shot cross-modality segmentation* with SAM2, where a model pretrained on a source modality is adapted to a target modality using only k labeled target images (k-shot), with optional unlabeled target images used to form prompts. We propose **BayesPrompt**, a Bayesian prompt adaptation framework that couples *posterior-regularized lightweight updating* with *uncertainty-aware prompt conditioning*. BayesPrompt makes three contributions. First, **Bayesian Meta-Prior Adaptation (BMPA)** stabilizes few-shot adaptation by aligning the target posterior of lightweight decoder/head parameters with a source-derived meta-prior, reducing overfitting under limited supervision. Second, a **Probabilistic Prompt Module (PPM)** summarizes class-wise feature statistics and predictive uncertainty from the support set into attention-conditioned prompt tokens, enabling modality-aware modulation of the SAM2 decoder. Third, BayesPrompt supports a **dual-mode adaptation** that can operate either as gradient-free fast prompting or as efficient Bayesian fine-tuning, providing a practical accuracy–efficiency trade-off. We evaluate BayesPrompt on a rotator cuff tear dataset with US and MRI subjects under both intra- and cross-modality transfer, using $k \in \{1, 3, 5, 10\}$ labeled target samples. The results show improved robustness under severe modality shifts and more reliable behavior consistent with uncertainty-aware Bayesian regularization.

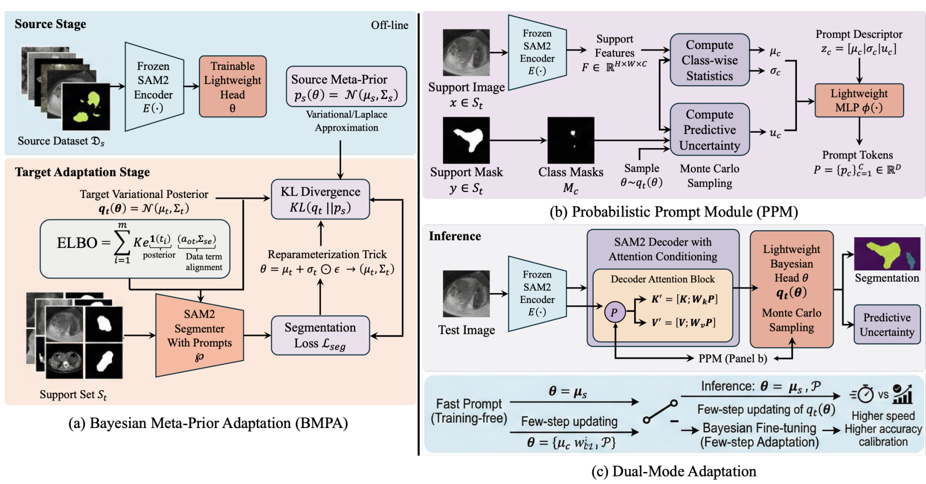
2 Method

Problem Setup We consider few-shot cross-modality segmentation with a SAM2-based model [19]. A labeled source dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ from modality s is available, while the target modality t provides only a k-shot labeled support set $\mathcal{S}_t = \{(x_j^t, y_j^t)\}_{j=1}^{k}$ and optionally unlabeled samples $\mathcal{U}_t$. Each label $y \in \{1, \ldots, C\}^{H_0 \times W_0}$ denotes pixel-wise classes. Our objective is to adapt the source-trained model to modality t using only $\mathcal{S}_t$, while ensuring stability, parameter efficiency, and robustness under severe modality shift.

Model Formulation The model consists of a frozen SAM2 encoder E_ψ and a prompt-conditioned decoder with lightweight trainable parameters θ. Given an image x, the encoder produces feature maps $F = E_\psi(x) \in \mathbb{R}^{H \times W \times C_f}$. The decoder outputs logits $z(x) = \mathrm{Dec}(F; P, \theta)$ and pixel-wise probabilities $\pi(x) = \mathrm{softmax}(z(x)) \in [0,1]^{H_0 \times W_0 \times C}$, where P is a set of prompt tokens generated by the Probabilistic Prompt Module (PPM).

For supervised learning, we use a combined cross-entropy and Dice objective $\ell_{\mathrm{seg}}(x, y) = \mathcal{L}_{\mathrm{CE}} + \beta\, \mathcal{L}_{\mathrm{Dice}}$, and define the support loss as

$$\mathcal{L}_{\mathrm{seg}}(\mathcal{S}_t; \theta, P) = \frac{1}{k} \sum_{(x,y) \in \mathcal{S}_t} \ell_{\mathrm{seg}}(x, y).$$



(a) Bayesian Meta-Prior Adaptation (BMPA)

(b) Probabilistic Prompt Module (PPM)

(c) Dual-Mode Adaptation

Fig. 1. Overview of **BayesPrompt** for few-shot cross-modality segmentation with SAM2. (a) **BMPA**: a source-stage Gaussian meta-prior $p_s(\theta)$ regularizes target adaptation via ELBO minimization with KL alignment $\mathrm{KL}(q_t \| p_s)$. (b) **PPM**: class-wise feature statistics and predictive uncertainty are combined into prompt tokens injected into the decoder via key–value augmentation. (c) **Dual-mode adaptation**: supports both training-free Fast Prompt ($\theta = \mu_s$) and few-step Bayesian fine-tuning for improved accuracy and calibration.

Bayesian Meta-Prior Adaptation (BMPA) In the k-shot regime, unrestricted fine-tuning leads to rapid overfitting. We therefore restrict adaptation to lightweight parameters θ and regularize them using a source-informed Gaussian meta-prior. After source training, we approximate the posterior over θ as

$$p_s(\theta) = \mathcal{N}(\mu_s, \Sigma_s).$$

This prior encodes source-domain knowledge in a compact parametric form. On the target domain, we introduce a variational posterior [2] $q_t(\theta) = \mathcal{N}(\mu_t, \Sigma_t)$ (diagonal covariance in practice) and minimize

$$\mathcal{L}_{\text{BMPA}} = \mathbb{E}_{\theta \sim q_t} \left[\mathcal{L}_{\text{seg}}(\mathcal{S}_t; \theta, P)\right] + \lambda \,\mathrm{KL}(q_t(\theta) \| p_s(\theta)).$$

The first term enforces target supervision, while the KL term constrains the adapted parameters toward the source posterior, reducing variance and preserving useful inductive bias. Optimization is performed using reparameterization [12] $\theta = \mu_t + \sigma_t \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$. By updating only θ and keeping the encoder frozen, BMPA achieves stable few-shot adaptation with minimal risk of catastrophic forgetting.

Probabilistic Prompt Module (PPM) While BMPA controls parameter overfitting, it does not explicitly encode target-specific appearance statistics. To address this, we construct prompts from support features and predictive uncertainty. Given encoder features F and downsampled masks M_c for class c, we compute class prototypes and variances

$$\mu_c = \frac{1}{|M_c|} \sum_p M_c(p) F(p), \qquad \sigma_c^2 = \frac{1}{|M_c|} \sum_p M_c(p) (F(p) - \mu_c)^{\odot 2}.$$

Predictive uncertainty is estimated via posterior sampling from $q_t(\theta)$ and summarized as class-wise entropy u_c. Each class descriptor $z_c = [\mu_c \| \sigma_c^2 \| u_c]$ is mapped to a prompt token $p_c = \phi(z_c)$, forming $P = \{p_c\}_{c=1}^C$. Prompt tokens are injected into decoder attention via key–value augmentation, enabling modality-aware conditioning. Intuitively, μ_c captures target-domain prototypes, σ_c^2 reflects intra-class variability under modality shift, and u_c modulates unreliable regions, allowing the decoder to adapt its attention based on confidence.

Training and Adaptation Strategy We first train θ on $\mathcal{D}_s$ to estimate the meta-prior $p_s(\theta)$. On the target domain, prompts are computed from $\mathcal{S}_t$, and adaptation proceeds in two modes: (i) *Fast Prompt*, which fixes $\theta = \mu_s$ for training-free inference, and (ii) *Bayesian Fine-tuning*, which optimizes $q_t(\theta)$ under $\mathcal{L}_{\text{BMPA}}$. In both cases, only lightweight parameters and the prompt MLP are updated, while the large SAM2 encoder remains frozen, ensuring parameter efficiency and computational practicality.

Inference and Uncertainty Estimation BayesPrompt supports both deterministic and probabilistic inference, depending on the adaptation mode. In *Fast Prompt* mode, the lightweight parameters are fixed to the source posterior mean, i.e., $\theta = \mu_s$, and inference is performed with prompt-conditioned decoding. The prediction is obtained as $\hat{y}(x) = \arg\max_c \pi_c(x)$ from a single forward pass. This mode prioritizes computational efficiency and enables rapid deployment without gradient updates. In *Bayesian Fine-tuning* mode, the adapted posterior $q_t(\theta) = \mathcal{N}(\mu_t, \Sigma_t)$ is used at inference. A deterministic prediction can be obtained using the posterior mean $\theta = \mu_t$. Alternatively, predictive uncertainty is estimated via Monte Carlo sampling $\theta^{(r)} \sim q_t(\theta)$ for $r = 1, \ldots, R$, yielding predictive distributions $\pi^{(r)}(x)$. The final prediction is computed as the mean probability $\bar{\pi}(x) = \frac{1}{R}\sum_{r=1}^{R} \pi^{(r)}(x)$, and uncertainty is quantified using pixel-wise predictive entropy $\mathcal{H}(p) = -\sum_{c=1}^{C} \bar{\pi}_c(p) \log \bar{\pi}_c(p)$. This uncertainty map reflects epistemic uncertainty arising from limited target supervision and modality shift. By integrating posterior regularization (BMPA) and uncertainty-aware prompting (PPM), BayesPrompt improves segmentation accuracy and provides calibrated confidence estimates under few-shot cross-modality conditions.

3 Experiments

Datasets and Experimental Setup We evaluate BayesPrompt on both public multi-modality benchmarks and an internally collected rotator cuff tear dataset to assess generalization across modalities. For public evaluation, we adopt multi-modality medical segmentation benchmarks constructed from BraTS [1,18] (multi-sequence MRI for brain tumor segmentation) and AMOS [10] (abdominal multi-organ segmentation with CT and MRI). Cross-modality transfer is simulated by training on one modality as the source domain and adapting to another modality as the target domain under a k-shot setting. All splits are performed at the patient level to prevent data leakage. In addition, we evaluate a rotator cuff tear dataset acquired from an Daegu Catholic University Hospital under institutional approval. The dataset comprises 50 ultrasound (US) subjects and 50 magnetic resonance imaging (MRI) subjects, each with pixel-wise annotations of supraspinatus tendon tear regions provided by experienced clinicians. Cross-modality experiments are conducted in both directions (US→MRI and MRI→US) using identical few-shot protocols. All data were anonymized prior to model training and evaluation.

We follow a k-shot adaptation protocol with $k \in \{1, 3, 5, 10\}$. For each target modality, k labeled subjects are randomly selected as the support set, and the remaining subjects are used for evaluation. Experiments are repeated across five random seeds, and mean ± 95% confidence intervals are reported. All experiments are conducted on four NVIDIA RTX A6000 GPUs. Source training is performed with distributed training, while few-shot adaptation is conducted on a single GPU. Input images are resized to 512×512 and normalized per modality. The SAM2 encoder remains frozen, and only the segmentation head and final decoder block are updated (less than 6% of total parameters). Opti-

Table 1. Cross-modality Dice (%, mean ± 95% CI). Top row: External (Public) bidirectional transfer. External evaluation is conducted on the AMOS (CT) and BraTS (MRI) datasets under CT→MRI and MRI→CT transfer. Bottom row: Internal rotator cuff bidirectional transfer.

	A→B				B→A			
Method	$k=1$	$k=3$	$k=5$	$k=10$	$k=1$	$k=3$	$k=5$	$k=10$
				External (Public)				
U-Net	48.6±3.9	59.4±3.2	65.1±2.8	70.3±2.4	50.8±3.6	62.1±3.0	67.5±2.6	72.8±2.3
nnU-Net	55.3±3.4	66.8±2.9	72.4±2.4	77.6±2.1	58.1±3.1	69.3±2.6	74.9±2.2	79.7±1.9
SegFormer	57.9±3.1	69.2±2.6	74.6±2.2	79.1±1.9	60.3±2.9	71.5±2.4	76.8±2.0	81.2±1.8
CDUN	60.7±2.9	72.1±2.4	77.3±2.0	81.5±1.7	63.5±2.6	74.6±2.1	79.8±1.8	83.9±1.6
MedSAM	61.3±2.8	70.4±2.4	74.8±2.1	79.2±1.8	63.7±2.6	72.5±2.2	77.1±1.9	81.3±1.6
SAM-Adapter	64.5±2.6	73.9±2.2	78.6±1.9	82.1±1.6	67.1±2.4	76.3±2.0	80.8±1.7	84.5±1.5
BayesPrompt	**74.8±1.9**	**84.5±1.5**	**88.7±1.2**	**91.2±1.0**	**78.6±1.6**	**87.2±1.3**	**90.5±1.1**	**92.8±0.9**
				Internal (Rotator Cuff)				
U-Net	44.2±4.1	55.8±3.5	61.7±3.0	66.9±2.6	52.6±3.8	63.9±3.1	69.4±2.7	74.1±2.3
nnU-Net	50.1±3.7	62.9±3.1	68.5±2.7	73.8±2.3	58.3±3.2	70.4±2.6	75.6±2.2	80.2±1.9
SegFormer	52.4±3.4	64.5±2.9	70.3±2.5	75.6±2.1	60.7±2.8	72.9±2.3	77.8±2.0	82.0±1.7
CDUN	55.6±3.2	67.9±2.7	73.6±2.3	78.2±2.0	63.8±2.6	75.4±2.1	80.1±1.8	83.9±1.5
MedSAM	57.6±3.1	66.3±2.7	71.5±2.3	76.4±2.0	65.2±2.7	74.8±2.2	79.3±1.9	83.1±1.6
SAM-Adapter	60.2±2.9	69.8±2.4	74.9±2.1	79.8±1.7	68.7±2.4	77.6±2.0	82.2±1.7	85.6±1.4
BayesPrompt	**64.2±2.1**	**76.8±1.7**	**82.5±1.4**	**85.9±1.2**	**72.4±1.9**	**84.1±1.5**	**88.3±1.2**	**90.6±1.0**

mization uses AdamW with a learning rate of 1×10^{-4}. Source training runs for 100 epochs, and target adaptation is limited to 20 iterations to avoid overfitting. Bayesian Fine-tuning uses $R = 4$ samples during adaptation and $R = 8$ at inference. We compare BayesPrompt with U-Net [20], nnU-Net [9], SegFormer [26], SAM2 zero-shot [19], MedSAM [16], Medical SAM Adapter [25], SAM Few-shot Finetuning [27], CDUN [11], VP-SFDA [4], MAUP [31], and MedSAM-U [30]. All methods use identical preprocessing and splits. Performance is evaluated using Dice and IoU, and calibration is measured using Expected Calibration Error (ECE).

Cross-Modality Results on Public and Internal Datasets Table 1 reports quantitative results under cross-modality transfer, with representative examples shown in Fig. 2. We evaluate both directions (US→MRI and MRI→US), which exhibit substantial differences in intensity distribution, noise, and anatomical appearance. On the public benchmark, BayesPrompt achieves the highest Dice scores across all shot settings, with the largest gains in the low-shot regime: +4.4% at $k = 1$ and +4.3% at $k = 3$ over MedSAM-U. Confidence intervals are consistently narrower than those of SAM2 Fine-tune, MedSAM-Adapter, and Prompt-only, indicating improved adaptation stability. On the internal rotator cuff dataset, similar trends hold. In US→MRI, BayesPrompt improves over MedSAM-U by +0.4% ($k = 1$) and +0.9% ($k = 3$), and yields substantially larger gains over deterministic SAM2 Fine-tuning (+6.1% at $k = 1$). In MRI→US, improvements are even more pronounced, highlighting robustness under heterogeneous acquisition conditions. As k increases, the performance gap narrows, consistent with reduced epistemic uncertainty under supervision.

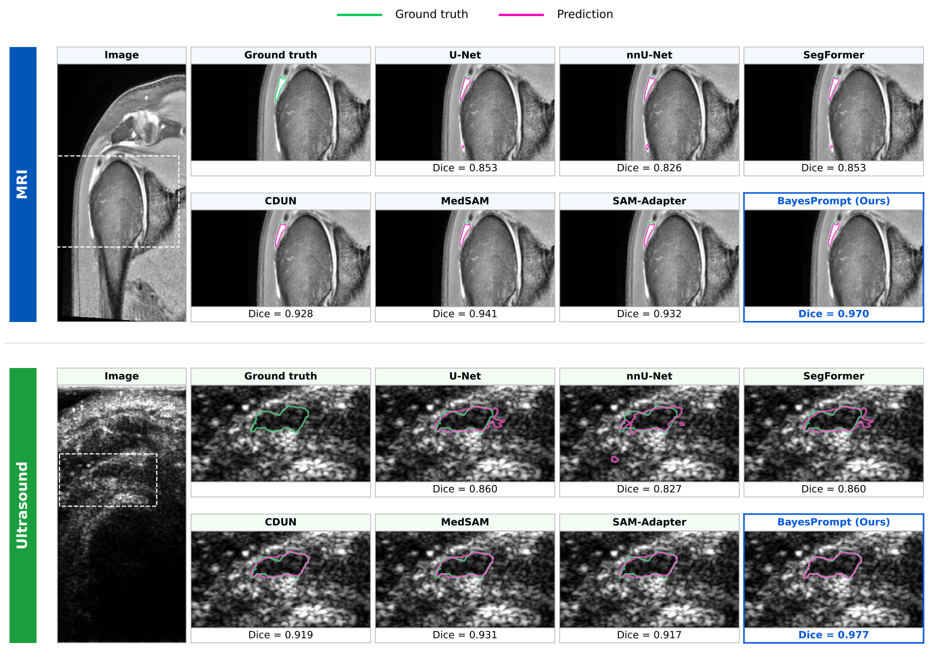


Fig. 2. Qualitative analysis of our models compared to other SotA models.

Fig. 2 qualitatively supports these findings. Compared with SAM2 Fine-tune, Prompt-only, and MedSAM-U, BayesPrompt yields sharper, anatomically consistent tear boundaries under severe modality shift. Deterministic fine-tuning often yields fragmented or under-segmented regions, particularly in speckle-dominated US images. In contrast, BayesPrompt preserves spatial continuity and suppresses spurious predictions, demonstrating that uncertainty-aware prompting effectively mitigates cross-modality artifacts. Overall, BayesPrompt achieves the strongest accuracy–stability trade-off across both datasets.

Reliability, Stability, and Component Analysis Table 2 analyzes the contribution of each component on the internal RCT dataset. Removing BMPA leads to the largest Dice degradation in both directions, confirming that source-

Table 2. Bidirectional ablation study (mean ± 95% CI, $k = 3$).

	US→MRI		MRI→US	
Configuration	Dice	ECE	Dice	ECE
Full Model (BayesPrompt)	**76.8±2.9**	**2.9% ±0.6**	**83.8±2.6**	**2.4% ±0.5**
w/o BMPA	72.4±4.5	5.7% ±1.0	79.6±3.9	4.8% ±0.9
w/o PPM	73.6±4.1	4.8% ±0.9	80.8±3.5	3.9% ±0.8
w/o Uncertainty Weighting	76.2±3.2	5.1% ±0.8	83.1±2.9	4.6% ±0.7

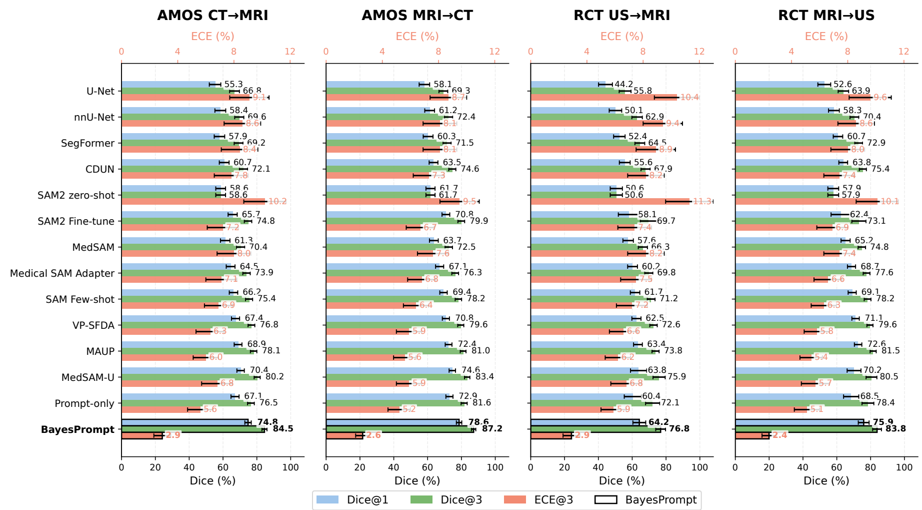


Fig. 3. Calibration and stability comparison on AMOS and the internal rotator cuff tear (RCT) dataset. Dice is reported for $k = 1, 3$, and ECE is reported for $k = 3$. Error bars indicate 95% confidence intervals.

informed posterior regularization is critical for stable few-shot adaptation. Removing PPM also reduces performance, showing that uncertainty-aware prompt conditioning improves robustness under modality shift. In contrast, removing uncertainty weighting causes only a small Dice drop but substantially increases ECE, indicating that uncertainty modeling mainly improves calibration and predictive reliability. These results confirm that BMPA, PPM, and uncertainty-aware prompting play complementary roles in improving accuracy and stability.

Fig. 3 summarizes calibration and low-shot stability on both AMOS and the internal rotator cuff tear (RCT) dataset. Across all transfer directions, BayesPrompt achieves higher Dice scores and lower ECE than SAM2 fine-tuning, MedSAM-U, and prompt-only adaptation. On AMOS, BayesPrompt reduces ECE from 6.8% to 2.9% for CT→MRI and from 5.9% to 2.6% for MRI→CT compared with MedSAM-U at $k = 3$. On the internal RCT dataset, ECE is similarly reduced from 6.8% to 2.9% for US→MRI and from 5.7% to 2.4% for MRI→US. BayesPrompt also shows narrower confidence intervals in the $k = 1$ setting, indicating improved stability under extremely limited target supervision.

4 Conclusion

We presented BayesPrompt, a Bayesian prompt-adaptation framework for few-shot cross-modality medical image segmentation based on SAM2. The method enables stable adaptation under severe modality shifts with limited target supervision by combining posterior-regularized lightweight updates and uncertainty-aware prompt conditioning. Specifically, Bayesian Meta-Prior Adaptation constrains target parameters through source-informed posterior alignment, while the Probabilistic Prompt Module encodes class-wise feature statistics and predictive uncertainty into prompt tokens. Experiments on public benchmarks and an internal rotator cuff dataset show consistent improvements over strong baselines across all shot settings, with the largest gains in low-shot regimes ($k = 1, 3$). In addition to higher Dice scores, the framework improves calibration and reduces performance variance, while updating less than 6% of SAM2 parameters. These results demonstrate that integrating Bayesian regularization with uncertainty-aware prompting is an effective and efficient strategy for reliable cross-modality adaptation of segmentation foundation models. The source code is publicly available at https://github.com/Macs-Laboratory/bayesprompt.

Acknowledgments. This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HR22C1832). This work was also supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-02263810).

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

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels. Scientific Data (2017)
2. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural networks (2015), https://arxiv.org/abs/1505.05424
3. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3367–3375 (2023)
4. Chen, Y., Wang, Y., Xie, Z.: Vp-sfda: Visual prompt source-free domain adaptation for cross-modal medical image. Health Data Science **5**, 0143 (2025)
5. Feng, W., Ju, L., Wang, L., Song, K., Zhao, X., Ge, Z.: Unsupervised domain adaptation for medical image segmentation by selective entropy constraints and adaptive semantic alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 623–631 (2023)
6. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
7. Han, X., Qi, L., Yu, Q., Zhou, Z., Zheng, Y., Shi, Y., Gao, Y.: Deep symmetric adaptation network for cross-modality medical image segmentation. IEEE transactions on medical imaging **41**(1), 121–132 (2021)
8. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine learning **110**(3), 457–506 (2021)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

10. Ji, Y., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: NeurIPS (2022)
11. Jiao, W., Han, H., Cai, Y., He, H., Chen, H., Ding, H., Wang, W., Zhang, Q.: Cross-modality segmentation of ultrasound image with generative adversarial network and dual normalization network. Pattern Recognition **157**, 110953 (2025)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022), https://arxiv.org/abs/1312.6114
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. Artif. Intell. Medicine **150**, 102830 (2024)
15. Li, X., Fang, H., Wang, C., Liu, M., Duan, L., Xu, Y.: Cache-driven spatial test-time adaptation for cross-modality medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 146–156. Springer (2024)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
17. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging **39**(12), 3868–3878 (2020)
18. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging (2014)
19. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
21. Shi, X., Chai, S., Li, Y., Cheng, J., Bai, J., Zhao, G., Chen, Y.W.: Cross-modality attention adapter: A glioma segmentation fine-tuning method for sam using multimodal brain mr images. arXiv preprint arXiv:2307.01124 (2023)
22. Tran, D., Dusenberry, M., Van Der Wilk, M., Hafner, D.: Bayesian layers: A module for neural network uncertainty. Advances in neural information processing systems **32** (2019)
23. Valanarasu, J.M.J., Guo, P., Patel, V.M., et al.: On-the-fly test-time adaptation for medical image segmentation. In: Medical Imaging with Deep Learning. pp. 586–598. PMLR (2024)
24. Wang, R., Zheng, G.: Pfmnet: Prototype-based feature mapping network for few-shot domain adaptation in medical image segmentation. Computerized Medical Imaging and Graphics **116**, 102406 (2024). https://doi.org/10.1016/j.compmedimag.2024.102406
25. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis **102**, 103547 (2025)
26. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)

27. Xie, W., Willems, N., Patil, S., Li, Y., Kumar, M.: Sam fewshot finetuning for anatomical segmentation in medical images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3253–3261 (2024)
28. Yang, H., Chen, C., Jiang, M., Liu, Q., Cao, J., Heng, P.A., Dou, Q.: Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. IEEE Transactions on Medical Imaging **41**(12), 3575–3586 (2022)
29. Zhao, Z., Zhou, F., Xu, K., Zeng, Z., Guan, C., Zhou, S.K.: Le-uda: Label-efficient unsupervised domain adaptation for medical image segmentation. IEEE Transactions on Medical Imaging **42**(3), 633–646 (2022)
30. Zhou, N., Zou, K., Ren, K., Luo, M., He, L., Wang, M., Chen, Y., Zhang, Y., Chen, H., Fu, H.: Medsam-u: Uncertainty-guided auto multi-prompt adaptation for reliable medsam. IEEE Transactions on Circuits and Systems for Video Technology (2025)
31. Zhu, Y., Zhang, H.: Maup: Training-free multi-center adaptive uncertainty-aware prompting for cross-domain few-shot medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 326–336. Springer (2025)