SPA: Efficient User-Preference Alignment against Uncertainty in Medical Image Segmentation

Jiayuan Zhu *
University of Oxford
jiayuan.zhu@keble.ox.ac.uk

Junde Wu *†
University of Oxford
junde.wu@linacre.ox.ac.uk

Cheng Ouyang
University of Oxford
cheng.ouyang@eng.ox.ac.uk

Konstantinos Kamnitsas
University of Oxford
konstantinos.kamnitsas@eng.ox.ac.uk

Alison Noble
University of Oxford
alison.noble@eng.ox.ac.uk

Abstract

*Medical image segmentation data inherently contain uncertainty, often stemming from both imperfect image quality and variability in labeling preferences on ambiguous pixels, which depend on annotators' expertise and the clinical context of the annotations. For instance, a boundary pixel might be labeled as tumor in diagnosis to avoid under-assessment of severity, but as normal tissue in radiotherapy to prevent damage to sensitive structures. As segmentation preferences vary across downstream applications, it is often desirable for an image segmentation model to offer user-adaptable predictions rather than a fixed output. While prior uncertainty-aware and interactive methods offer adaptability, they are inefficient at test time: uncertainty-aware models require users to choose from numerous similar outputs, while interactive models demand significant user input through click or box prompts to refine segmentation. To address these challenges, we propose **SPA**, a segmentation framework that efficiently adapts to diverse test-time preferences with minimal human interaction. By presenting users a select few, distinct segmentation candidates that best capture uncertainties, it reduces clinician workload in reaching the preferred segmentation. To accommodate user preference, we introduce a probabilistic mechanism that leverages user feedback to adapt model's segmentation preference. The proposed framework is evaluated on a diverse range of medical image segmentation tasks: color fundus images, CT, and MRI. It demonstrates 1) a significant reduction in clinician time and effort compared with existing interactive segmentation approaches, 2) strong adaptability based on human feedback, and 3) state-of-the-art image segmentation performance across diverse modalities and semantic labels.*

---

*equal contribution
†project lead

1. Introduction

Deep learning-based medical image segmentation has achieved remarkable advancements over the past decade [6, 28, 32]. However, existing approaches often fall short when applied to real-world dynamic clinical environments. A critical challenge is to handle the inherent uncertainties in medical images [13]. A single medical image may have multiple different valid segmentations, depending on the labeling criteria under a specific medical context. For instance, in glioma detection from brain CT scans, it is often preferred to include surrounding tissue than risk missing part of the tumor [2]. While in radiation therapy for low-grade glioma, undersegmentation is used to protect sensitive brain tissue from excess radiation damage [27]. Therefore, developing adaptive methodologies that align segmentation uncertainties with specific labeling preferences for different clinical needs is essential.

Existing uncertainty-aware approaches [3, 17, 22, 29] represent segmentation uncertainty to users by generating numerous stochastic predictions (Fig. 1). However, these approaches require clinicians to painstakingly choose from countless similar-looking candidates, making the process overwhelming and time-consuming. Furthermore, since these models cannot incorporate human feedback for adjustments, there is a risk that none of the stochastic predictions may provide a satisfactory segmentation.

Incorporating human feedbacks, often in the form of visual prompts, into the segmentation has achieved significant progress [7, 18, 24, 31]. However, most of such interactive segmentation approaches are unable to incorporate uncertainty. Additionally, existing visual prompts like clicks and bounding boxes are often cumbersome in real world, as they demand pixel-level interaction input from the users (Fig. 1).

To address the above challenges, we believe the most effective solution is to use a small number of distinct candi-
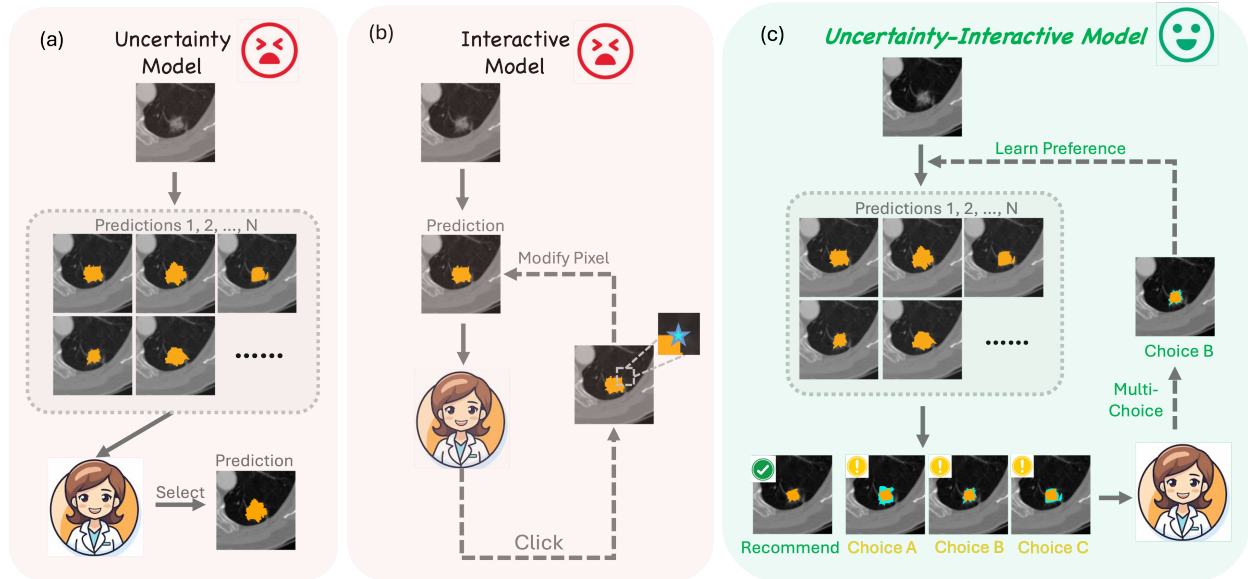
Figure 1. (a) Existing uncertainty-aware models suffer from choosing among numerous similar-looking candidates, making the process labor-intensive and time-consuming. (b) Interactive segmentation models lack the ability to incorporate image uncertainty and rely heavily on pixel-level interactions, which require a substantial amount of time and effort. (c) Our uncertainty-aware interactive segmentation model, SPA, efficiently achieves preference-aligned segmentation by incorporating medical image uncertainties and human interactions. Clinicians are presented with one recommended prediction and a few distinct segmentation candidates that capture uncertainty, allowing them to select the one best aligned with their clinical needs. If the user is unsatisfied with the recommended prediction, the model learns from the user selection, adapts itself, and presents users a new set of candidates. Our approach minimizes user interactions and eliminates the need for painstaking pixel-wise adjustments compared to conventional interactive segmentation models.

dates to represent segmentation uncertainty, allowing users to make a straightforward multi-choice selection. In this paper, we introduce **SPA**, a new paradigm for efficient **S**egmentation **P**reference **A**lignment with uncertainty in medical image segmentation (Fig. 1). SPA presents image uncertainties by generating multiple segmentations. Instead of overwhelming clinicians with numerous similar-looking predictions as in conventional uncertainty-aware segmentations, our model offers one recommended prediction and four representative correction proposals per iteration. Once the user selects an option, SPA rapidly adapts to the corresponding labeling preference, allowing convergence to the desired segmentation within only a few iterations. Experimental validation demonstrates that clinicians can segment 35% more images with 39% fewer iterations compared to previous interactive models, highlighting SPA's potential for real-world clinical applications. In summary, our contributions are:

- We introduce SPA, the first framework specifically designed to align user preferences with segmentation uncertainty, minimizing both interaction time and clinician effort.
- We propose a preference distribution using mixture Gaussians to represent image uncertainties, enabling rapid adaptation to user preferences with minimal iterations.
- We develop a multi-choice interaction mechanism that cap-

tures user preferences from modeled uncertainties, achieving preference-aligned predictions while reducing clinician interaction effort.
- We compare SPA with deterministic, uncertainty-aware, and interactive SOTA methods. SPA achieves superior results measured in Dice Score across multi-clinician annotated datasets such as REFUGE2, LIDC-IDRI and QUBIQ. It also consistently outperforms interactive methods within fewer rounds of user interactions.

2. Related Work

Medical Image Segmentation Medical image segmentation involves predicting a dense pixel-wise map to delineate structures such as lesions, tumors, and anatomical regions. It plays an indispensable role in disease diagnosis and treatment planning [26, 35, 39]. Deep learning has significantly enhanced segmentation accuracy and efficiency across various clinical applications [5, 34], with many models demonstrating remarkable success [6, 28, 32]. However, these models often fail to address the inherent uncertainty in medical images, as multiple plausible segmentations may exist for a single image [13]. To improve medical image segmentation, it is crucial to develop methods that address these uncertainties and identify the most suitable segmentation for the specific clinical context.
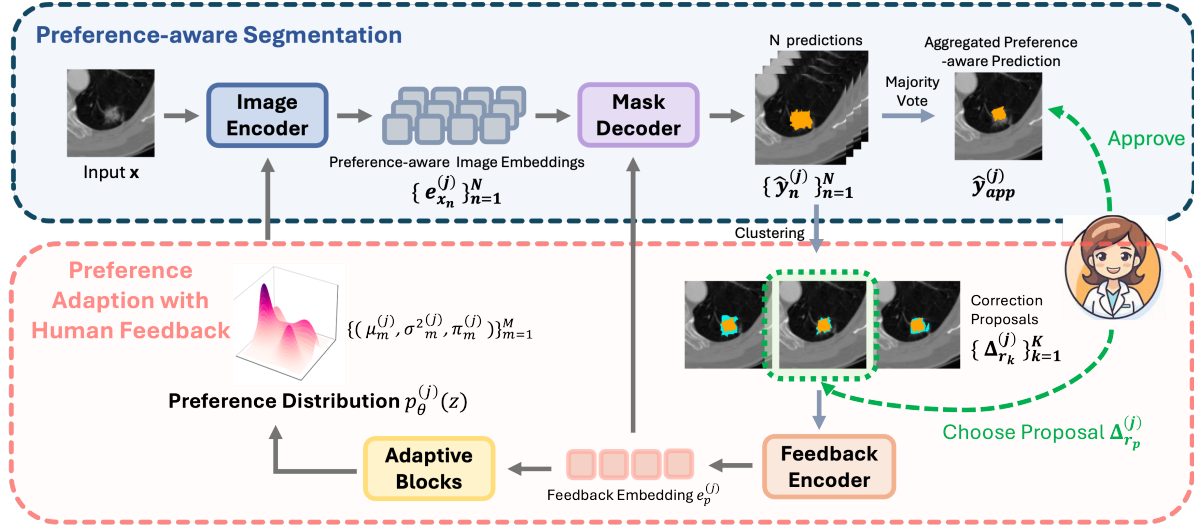
Figure 2. **Overall framework of SPA**. The inference process comprises of two steps: *Preference-aware Segmentation* and *Preference Adaptation with Human Feedback*. At timestamp j, SPA takes the input image $\mathbf{x}$, an interaction embedding $e_p^{(j)}$, and samples $\{\mathbf{z_n}^{(j)}\}_{n=1}^N$ drawn from the preference distribution $p_\theta^{(j)}(z)$ to generate N segmentation predictions. These predictions are combined into a aggregated preference-aware prediction $\hat{\mathbf{y}}_{app}^{(j)}$ via majority vote. If the clinician does not approve $\hat{\mathbf{y}}_{app}^{(j)}$, SPA generates K correction proposals $\{\mathbf{\Delta_{r_k}}^{(j)}\}_{k=1}^K$ in the adaptation step. The clinician selects the preferred proposal and the preference distribution is updated to $p_\theta^{(j+1)}(z)$ based on this choice. This process iterates until the segmentation meets clinical satisfaction.

Uncertainty-aware Segmentation Uncertainty in medical images cannot be reduced by adding more data or using more complex models [19]. Techniques such as model ensembling, label sampling [15], and multi-head strategies [11] attempt to address this uncertainty by generating a range of potential predictions that reflect different user preferences [14, 40]. Probabilistic segmentation methods, including ProbUNet [20], PhiSeg [3], CM-Net [37], and MRNet [16] explicitly model the posterior distribution of parameters or predictions to capture uncertainty. However, these approaches often generate multiple predictions, requiring clinicians to review each one individually. Moreover, the predictions may not perfectly match the specific clinical context and some techniques rely on prior knowledge of the clinician's expertise, complicating their adoption in clinical practice. In contrast, a more efficient solution is to incorporate human interactions, allowing clinicians to refine the segmentations through direct interaction to better align with the clinical context.

Interactive segmentation Interactive segmentation is an iterative process where automated segmentation results are refined through user input until they meet the desired output. Previous methods [18, 30, 33] have effectively modified predictions based on user interaction at the pixel level, with many achieving notable success in the medical domain [9, 25, 36, 38]. However, these interactive models often fail to address the inherent uncertainty in medical images. Additionally, they may require numerous iterations to align with the user preference for specific clinical context, given the existence of multiple valid segmentations. An interac-

tive model capable of incorporating image uncertainties and learning from user interactions would reduce the number of iterations required to achieve the desired outcomes.

3. Methodology

3.1. Framework Overview

Medical image segmentation faces challenges due to variations in labeling preferences, often influenced by clinical guidelines for downstream analysis. These variations can result in multiple plausible annotations for a single image, highlighting the need for adaptable segmentation models that align with user preferences across different clinical contexts. To address this challenge, we propose SPA, a segmentation framework that adapts to user preferences. These preferences are modeled using a Gaussian Mixture distribution $p_\theta(z) = \sum_{m=1}^M \pi_m \mathcal{N}(z|\mu_m, \sigma_m^2)$, where parameters $\theta = \{(\mu_m, \sigma_m^2, \pi_m)\}_{m=1}^M$ define the *preference distribution* $p_\theta(z)$. Samples are drawn from this distribution as a conditioning signal to capture variations under user preferences [1]. During training, the preference distribution $p_\theta(z)$ captures preference-related segmentation variations. At inference, the model efficiently adapts to user preferences by estimating the maximum likelihood based on user's (clinician) feedback.

Our goal is to train a general function $f(\cdot, \cdot)$ that can adapt

---

[1]Preliminary experiments with MRNet [16] and findings in Appendix suggest that while segmentation performance varies across clinicians, each clinician demonstrates consistent results. These observations indicate that modeling interaction behavior can effectively learn user preferences, thereby improving model efficiency.

to different user preferences p's through a multi-choice interaction process: $\mathbf{y} = f(\mathbf{x}, \mathbf{\Delta_{r_p}})$, where $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is the input image, and $\mathbf{\Delta_{r_p}}$ reflects user interaction, in the form of a selection among multiple ways for correcting the current segmentation. Specifically, the inference of our SPA framework consists of two main steps (Fig. 2): *Preference-aware Segmentation* (Section 3.2) and *Preference Adaption with Human Feedback* (Section 3.3). The *Preference-aware Segmentation* generates multiple valid segmentations to represent image uncertainty and the *Preference Adaption with Human Feedback* refines these segmentations iteratively to align with specific user preferences. As shown in the upper part of Fig. 2, the model function $g(\cdot, \cdot, \cdot)$ processes three components in the *Preference-aware Segmentation*: the input image $\mathbf{x}$, interaction embedding $e_p^{(j)}$, and $\{\mathbf{z_n}^{(j)}\}_{n=1}^N$ sampled from the preference distribution $p_\theta^{(j)}(z)$ at timestamp j. Then, model generates N predictions $\{\hat{\mathbf{y}}_{\mathbf{n}}^{(j)}\}_{n=1}^N$ to represent uncertainty across individual preferences, where $\mathbf{z_n}^{(j)}$'s control the subtle shape based on the user feedback. These predictions are combined into a aggregated preference-aware prediction $\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}$ and clustered into K *correction proposals* $\{\mathbf{\Delta_{r_k}}^{(j)}\}_{k=1}^K$. The aggregated preference-aware prediction $\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}$ is generated through majority vote and then binarized. If the clinician approves with $\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}$, it will be used as the final segmentation and the iteration terminates. Otherwise, the framework uses *Preference Adaption with Human Feedback* step for further adaption, as shown in the lower part of Fig. 2: The clinician selects the proposal $\mathbf{\Delta_{r_p}}^{(j)}$ that best aligns with their preference p. The preference distribution $p_\theta^{(j)}(z)$ is then updated to $p_\theta^{(j+1)}(z)$ based on this feedback. This iterative process continues until the clinician satisfies with the prediction.

By doing so, SPA gain three advantages over previous methods: First, it explicitly models diverse human preferences incorporated in multi-annotator datasets with a preference distribution $p_\theta(z)$ in training time, increasing the chance for a preferred segmentation to be reachable at inference time. Second, it improves efficiency by adapting predictions with fewer interaction rounds than existing interactive segmentation methods. Third, it reduces clinician's effort at each user interaction by replacing pixel-level pinpoint with a simpler multi-choice selection approach.

3.2. Preference-aware Segmentation

As shown in the upper part of Fig. 2, our *Preference-aware Segmentation* step involves the synergy of an image encoder and a mask decoder. Given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, we first obtain a general image embedding $\mathbf{e_x}^{(j)} \in \mathbb{R}^{L \times \frac{H}{P} \times \frac{W}{P}}$ at time j, where L represents the output channels and P is the Vision Transformer (ViT) patch size. This embedding is generated using a pre-trained ViT, following the architecture described in MAE [12]. Additionally,

we sample N preference conditions $\{\mathbf{z_n}^{(j)}\}_{n=1}^N$ from the preference distribution $p_\theta^{(j)}(z)$. Each $\mathbf{z_n}^{(j)} \in \mathbb{R}^{L \times \frac{H}{P} \times \frac{W}{P}}$ is concatenated with the general image embedding $\mathbf{e_x}^{(j)}$, and then processed through convolutional layers and ReLU activations to produce a set of preference-aware image embeddings $\{\mathbf{e_{x_n}}^{(j)}\}_{n=1}^N$, where each $\mathbf{e_{x_n}}^{(j)} \in \mathbb{R}^{L \times \frac{H}{P} \times \frac{W}{P}}$.

$$\mathbf{e_{x_n}}^{(j)} = ReLU(Conv(\mathbf{e_x}^{(j)} \oplus \mathbf{z_n}^{(j)})), n = 1, 2, ..., N \tag{1}$$

To predict dense semantic masks from the preference-aware image embeddings $\{\mathbf{e_{x_n}}^{(j)}\}_{n=1}^N$ and the interaction embedding $e_p^{(j)}$ (Section 3.3), we employ a mask decoder that follows the same architecture as SAM [18]. This decoder predicts N segmentation masks $\{\hat{\mathbf{y}}_{\mathbf{n}}^{(j)}\}_{n=1}^N$ for the input image $\mathbf{x}$ at timestamp j. The aggregated preference-aware prediction $\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}$ is generated by applying a majority vote on the predictions $\{\hat{\mathbf{y}}_{\mathbf{n}}^{(j)}\}_{n=1}^N$, followed by binarization.

3.3. Preference Adaption with Human Feedback

As shown in the lower part of Fig. 2, $\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}$, our model also provides K correction proposals $\{\mathbf{\Delta_{r_k}}^{(j)}\}_{k=1}^K$ for further selection. These correction proposals reflect the distinctive possible ways for adapting the prediction to the desired preference. To create $\{\mathbf{\Delta_{r_k}}^{(j)}\}_{k=1}^K$, K-means clustering is applied to the predictions $\{\hat{\mathbf{y}}_{\mathbf{n}}^{(j)}\}_{n=1}^N$. Each cluster produces a representative mask $\mathbf{r_k}^{(j)} \in \{\mathbf{r_k}^{(j)}\}_{k=1}^K$ through inner-cluster majority vote. The pixel-wise differences from $\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}$ highlight areas for correction, forming the K correction proposals $\{\mathbf{\Delta_{r_k}}^{(j)}\}_{k=1}^K$. Once the clinician selects a preferred correction proposal $\mathbf{\Delta_{r_p}}^{(j)}$, the corresponding representative mask $\mathbf{r_p}^{(j)}$ is compared with $\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}$ (both in their pre-binarized forms) to identify the locations with the greatest divergence, denoted as $P_{diff} = \{(h, w)\}, h \in [0, H-1], w \in [0, W-1]$. A random location (point prompt) sampled from P_{diff}, along with its semantic label, is then chosen as the feedback signal. This signal is mapped into a L-dimensional vector interaction embedding e_p by the feedback encoder, as shown in Fig. 2. The feedback signal for the input image $\mathbf{x}$ is initialized by randomly selecting a pixel, as no correction proposals are available at this stage.

Using the interaction embedding $e_p^{(j)}$, the preference distribution $p_\theta^{(j)}(z)$'s parameters $\theta = \{(\mu_m^{(j)}, \sigma_m^{2\,(j)}, \pi_m^{(j)})\}_{m=1}^M$ are updated, resulting preference distribution $p_\theta^{(j+1)}(z)$. This adjustment is done through adaptive blocks consisting of six fully connected layers and ReLU activations, resulting in a updated preference distribution $p_\theta^{(j+1)}(z)$, as shown in the lower part of Fig. 2. In this step, the overall framework f adapts its output to the specific human preference through interactions, expressed as: $f^{(j)}(\mathbf{x}, \mathbf{\Delta_{r_p}}^{(j)}) = f^{(j)}(\mathbf{x}, e_p^{(j)})$. This adaptive update process allows the preference distribution $p_\theta^{(j+1)}(z)$ to progressively align with human preferences

4

through sufficient interactions, as described in Section 3.4.

3.4. Preference Distribution

Our approach theoretically ensures that SPA gradually adapts to each clinician's clinical contexts, aligning with personalized preferences through iterative feedback. We assume that each clinician's interactions, based on their preference and stemming from different medical contexts, are independent and identically distributed (i.i.d.). To represent all possible user preferences for different clinical needs, we model the preference distribution $p_\theta(z)$ as a mixture of M Gaussian distributions:

$$p_\theta(z) = \sum_{m=1}^{M} \pi_m \mathcal{N}(z \mid \mu_m, \sigma_m^2) \qquad (2)$$

where $\mathcal{N}(z \mid \mu_m, \sigma_m^2)$ denotes the Gaussian component for clinician m, with mean μ_m and variance σ_m^2. The mixture weights π_m satisfy $\pi_m > 0$ and $\sum_{m=1}^{M} \pi_m = 1$.

Each component in this mixture model reflects a clinician's specific interaction behavior, allowing the framework to account for multiple plausible annotations. Through iterative interactions, SPA adapts this preference distribution $p_\theta(z)$, aligning it with the clinician's unique segmentation preferences and adapting to their personalized clinical needs. This process ensures that the updated distribution $p_\theta^{(j+1)}(z)$ effectively captures clinician preferences, thus improving future predictions $\hat{\mathbf{y}}_{\mathbf{app}}^{(j+1)}$. We provide a theoretical proof in Appendix demonstrating that, given sufficient interactions from a specific clinician p, the preference distribution $p_\theta(z)$ could converge to the clinician's personalized distribution $\mathcal{N}(\mu_p, \sigma_p^2)$ in theory. It theoretically guarantees that SPA can ultimately reflect each clinician's requirements, aligning closely with diverse clinical contexts and user preferences through multi-round feedback alignment.

3.5. Training Details

Our training process simulating the actual inference, by constructing interactive segmentation tasks for each batch, where *Preference-aware Segmentation* and *Preference Adaption with Human Feedback* stages alternates.

Specifically, for each training iteration, at the *Preference-aware Segmentation* stage, parameters of image encoder and mask decoder, denoted as $\theta_{PSeg}^{(j)}$ are updated using the cross-entropy loss: $\theta_{PSeg}^{(j+1)} = \theta_{PSeg}^{(j)} + \alpha \nabla_{\theta_{PSeg}} \mathcal{L}_{ce}(\hat{\mathbf{y}}_{\mathbf{app}}^{(j)}, \mathbf{y})$. To capture image uncertainty, we generate $N = 48$ predictions by sampling from the preference distribution $p_\theta^{(j)}(z)$, which is modeled with of $M = 16$ Gaussian components. The means, variances, and mixture weights $\{(\mu_m^{(0)}, \sigma_m^{2\,(0)}, \pi_m^{(0)})\}_{m=1}^{16}$ are uniformly initialized. Additionally, we generate $K = 4$ correction proposals to allow clinicians for multi-choice selection.

For the *Preference Adaption with Human Feedback* step, all parameters from the *Preference-aware Segmentation* step remain frozen. A new prediction $\hat{\mathbf{y}}_{\mathbf{app}}^{(j+1)}$ is generated by sampling from the updated preference distribution $p_\theta^{(j+1)}(z)$. We denote all parameters in feedback encoder and Gaussian parameters $\{(\mu_m^{(j)}, \sigma_m^{2\,(j)})\}_{m=1}^{M}$ as $\theta_{PAF}^{(j)}$. $\theta_{PAF}^{(j)}$ are updated using the cross-entropy loss: $\theta_{PAF}^{(j+1)} = \theta_{PAF}^{(j)} + \alpha \nabla_{\theta_{PAF}} \mathcal{L}_{ce}(\hat{\mathbf{y}}_{\mathbf{app}}^{(j+1)}, \mathbf{y})$. Additionally, the Gaussian mixture weights $\{(\pi_m^{(j)})\}_{m=1}^{M}$, denoted as $\theta_\pi^{(j)}$, are updated with a mean squared error loss $\mathcal{L}_{mse}$ conditioned on $\hat{\mathbf{y}}_{\mathbf{app}}^{(j+1)}$ and $\mathbf{y}$, respectively: $\theta_\pi^{(j+1)} = \theta_\pi^{(j)} + \alpha \nabla_{\theta_\pi} \mathcal{L}_{mse}(P(I = i \mid \hat{\mathbf{y}}_{\mathbf{app}}^{(j+1)}, P(I = i \mid \mathbf{y}))$, where I stands for the Gaussian component in the preference distribution.

4. Experiment

We conducted extensive experiments to validate the effectiveness of SPA across seven uncertainty segmentation tasks represented by multi-clinician, using data from various imaging modalities, including color fundus images, CT, and MRI scans. SPA consistently achieves SOTA performance compared to deterministic, uncertainty-aware, and interactive models. Notably, SPA outperforms interactive models with significantly fewer iterations and it demonstrates strong generalization on unseen clinicians. Human evaluation further shows that SPA requires less time and fewer interactions to meet human expectations. Our results also show that predictions increasingly align with selected clinicians while diverging from those excluded. Ablation studies highlight the necessity of each component for optimal segmentation. Additional analysis of correction proposal similarity across iterations and prediction changes after each interaction is detailed in Appendix.

4.1. Dataset

REFUGE2 benchmark [10] is a publicly available fundus image dataset for glaucoma analysis, including the optic cup segmentation. It includes annotations from seven independent ophthalmologists, each with an average of eight years of experience. The dataset contains 400 training images and 400 testing images.

LIDC-IDRI benchmark [1, 8] originally consists of 3D lung CT scans with semantic segmentations of possible lung abnormalities, annotated by four radiologists. We use a preprocessed version from [21], which includes 15,096 2D CT images. After an 80-20 train-test split, the training and testing datasets contain 12,077 and 3,019 images, respectively.

QUBIQ benchmark [23] investigates inter-clinician variability in medical image segmentation tasks. It includes one MRI brain tumor task (three annotations, 28 training cases, 4 testing cases); two MRI prostate tasks (six annotations, 48 training cases, 7 testing cases); one MRI brain-growth task

Table 1. **SPA Outperforms SOTA in Dice Score (%).** Performance analysis comparing Dice Scores across deterministic, uncertainty-aware, and interactive models. SAM-series models use clicks for interaction, while SAM-U uses bounding boxes. SPA, with its multi-choice correction proposal setting, consistently outperforms all models across diverse datasets. 1-Iter and 3-Iter indicate performance after one and three iterations, respectively.

Methods	Category	1-Iter	3-Iter	REFUGE2	LIDC	BrainTumor	Prostate1	Prostate2	BrainGrowth	Kidney	Ave
UNet		✓		68.94	62.99	87.30	83.89	77.22	62.02	82.40	74.96
TransUNet	Det	✓		80.83	64.09	90.14	83.35	68.34	86.58	52.99	75.19
SwinUNet		✓		78.67	59.45	91.23	82.02	74.19	74.88	69.41	75.69
Ensemble UNet		✓		70.75	63.84	90.56	85.27	79.07	71.69	89.30	78.64
ProbUnet		✓		68.93	48.52	89.02	72.13	66.84	75.59	75.73	70.96
LS-Unet	Unc	✓		73.32	62.05	90.89	87.92	81.59	85.63	72.31	79.10
MH-Unet		✓		72.33	62.60	86.74	87.03	75.61	83.54	73.44	77.32
MRNet		✓		80.56	63.29	85.84	87.55	70.82	84.41	61.30	76.25
SAM		✓		82.59	66.68	91.55	92.82	77.04	86.63	85.72	83.29
MedSAM	Int	✓		82.34	68.42	92.67	89.69	74.70	85.91	78.02	81.68
MSA		✓		83.03	66.88	88.16	89.06	68.94	80.62	25.29	71.71
SAM-U V1		✓		82.45	62.24	92.67	81.46	66.56	87.79	89.50	80.38
SAM-U V2	Unc-Int	✓		80.66	64.82	93.11	91.89	72.91	87.51	90.74	83.09
SPA		✓		**83.47**	**88.07**	**94.29**	*93.12*	**83.34**	**88.14**	**94.08**	**89.22**
SAM			✓	82.61	66.71	92.14	92.72	77.54	86.58	90.43	84.10
MedSAM	Int		✓	82.13	68.45	93.26	90.05	73.81	86.09	79.88	81.95
MSA			✓	83.08	66.87	91.25	90.22	71.34	81.87	46.76	75.91
SAM-U V1			✓	82.10	62.84	92.31	81.79	66.74	87.84	89.24	80.40
SAM-U V2	Unc-Int		✓	80.54	65.44	92.40	90.00	73.17	87.87	91.35	82.96
SPA			✓	*85.42*	*88.56*	*94.31*	*92.97*	*84.05*	*88.18*	*94.26*	*89.68*

(seven annotations, 34 training cases, 5 testing cases); and one CT kidney task (three annotations, 20 training cases, 4 testing cases).

4.2. Experimental Setup

Our network was implemented using the PyTorch platform and trained/tested on an RTX A4000 with 16GB of memory. During training, we used the Adam optimizer with an initial learning rate of $1e^{-4}$ and adjusted it using StepLR strategy. To ensure a fair comparison, deterministic methods with multiple annotations were trained using majority vote. For the SAM-series interactive models, click or bounding box prompts were randomly generated based on the original model settings. During testing, a random set of annotations was selected, fused, and binarized. This fused binary segmentation was then used as the ground truth for evaluation. Further details are provided in Appendix. We consistently used ViT/B as the backbone when vision transformer was involved in the models.

4.3. Experimental Result

4.3.1. Performance Analysis with SOTA Methods

To demonstrate the advantages of SPA, we compared it with SOTA methods, classified into deterministic methods (UNet [28], TransUNet [6], SwinUNet [4]), uncertainty-aware methods (Ensemble UNet, ProbUnet [20], LS-Unet [15], MH-Unet [11], MRNet [16]), and interactive methods (SAM [18], MedSAM [25], MSA [36]). We also compared SPA with SAM-U [9], an uncertainty-interactive method

that simply introduces uncertainty by generating multiple prompts. SAM-U was evaluated using both SAM and Med-SAM backbones with bounding boxes as the interaction strategy, named SAM-U V1 and SAM-U V2, respectively. Other SAM-series methods relied on user clicks for interactions. Results for interactive models were reported after one and three iterations.

SPA consistently outperforms all methods, achieving an average Dice Score of 89.68% after three iterations: Table 1 provides a quantitative comparison of Dice Scores across datasets. The improvement is particularly notable in the LIDC segmentation task, where SPA exceeds current SOTA methods by 20%. Even after one iteration, SPA demonstrates superior performance than SOTA methods, highlighting its potential to generate high-quality segmentation with minimal interaction. Fig. 3 shows visual comparisons between SPA and other SOTA methods, presenting segmentations after six iterations for the interactive models. The results indicate that the segmentations predicted by SPA align more closely with the ground truth, especially at boundary regions.

4.3.2. Efficiency Analysis on Different Interactive Models

We conducted an efficiency analysis to quantify the number of iterations required to reach specific Dice Scores, with a maximum of six iterations. Models that failed to reach the target Dice Score within the limit were assigned an iteration count of ten. Our proposed method, SPA, consistently required fewer iterations to achieve superior segmentation performance compared to other models.
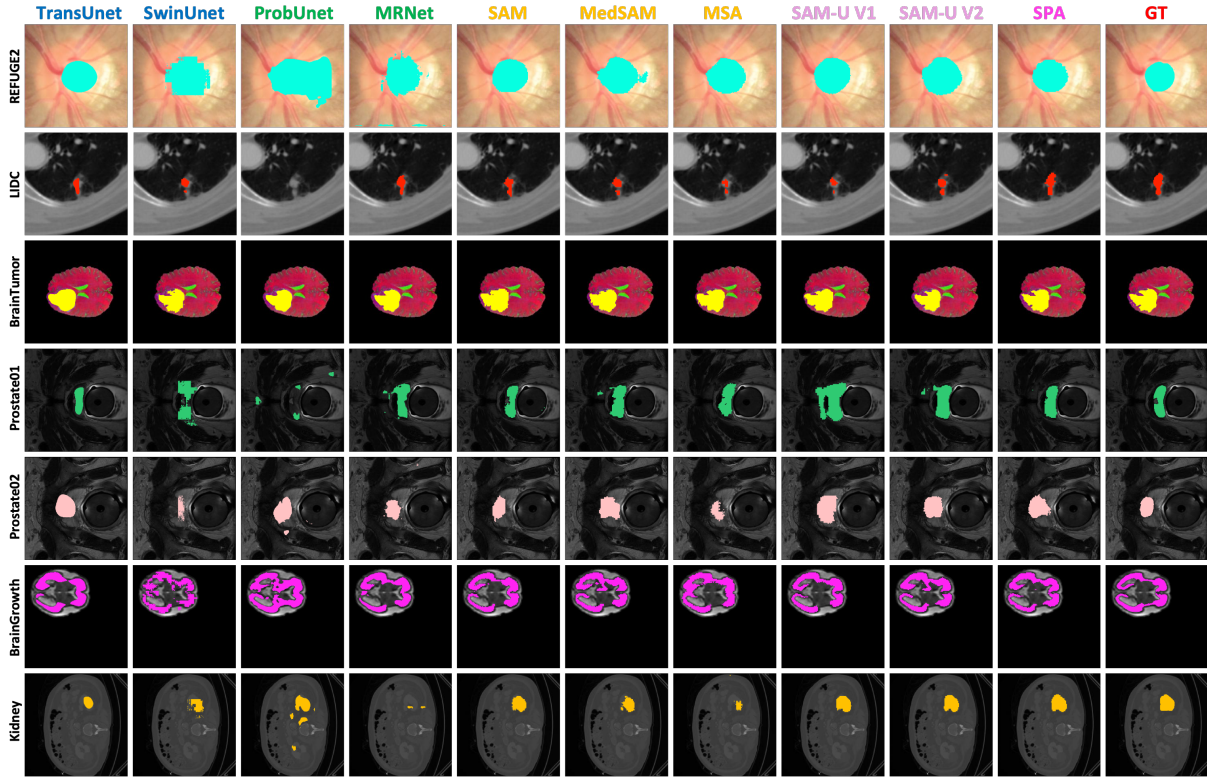
Figure 3. **SPA Shows Superior Segmentation Visualization.** Visual comparison of segmentation results with deterministic, uncertainty-aware, and interactive models after six iterations. SPA provides better adaptability, particularly at boundary regions.
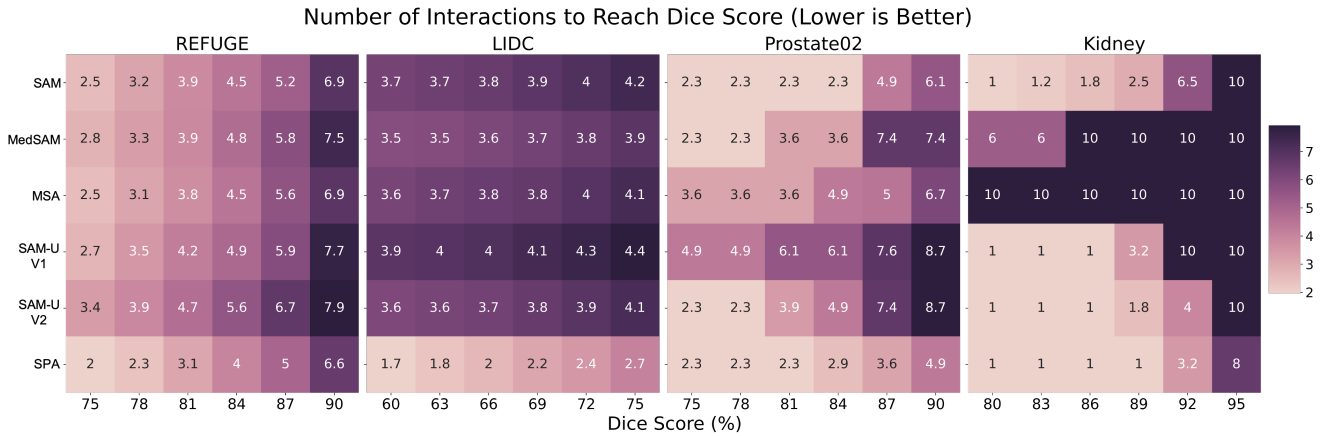

Figure 4. **SPA Demonstrates Extraordinary Efficiency.** Efficiency analysis comparing the average number of iterations required to reach specific Dice Scores across interactive models. Models that failed to reach the target Dice Score within six iterations are assigned an iteration count of ten. SPA consistently requires fewer iterations to achieve high-performance segmentation results.

SPA outperforms all other interactive models in terms of interaction efficiency, as shown in Fig. 4. On the REFUGE dataset, it requires fewer than five iterations to achieve a Dice Score of 87%, whereas SAM-U V2 needs 5.6 iterations to reach a Dice Score of just 84%. In the LIDC dataset, SPA achieves 75% Dice Score with only 2.7 iterations on average, while other models typically require four or more iterations.

Similarly, on the Kidney dataset, SPA reaches a Dice Score of 92% with 3.2 iterations, while other SAM-series models often need four to ten iterations. Notably, SPA reaches a Dice Score of 95% within eight iterations, but other models struggle to converge within the ten-iteration limit.

SPA demonstrates generalization and robustness to unseen preferences. To show this, we left one clinician's

7

Table 2. **SPA Demonstrates Strong Prediction Alignment with Individual Clinicians.** Alignment analysis by Dice Score (%) on the REFUGE2 optic cup segmentation task. The ground truth combines annotations from Clinicians 2, 4, 5, 6, and 7, excluding Clinicians 1 and 3. Predictions are compared with individual clinicians (columns) across interaction iterations (rows). The predictions show increasing alignment with included clinicians and divergence from excluded clinicians.

Iteration	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Clinician 6	Clinician 7
1	75.02	78.92	83.98	78.25	80.81	84.86	59.35
2	74.03 (-1.32%)	80.58 (+2.10%)	84.03 (+0.05%)	78.88 (+0.81%)	81.92 (+1.37%)	85.62 (+0.90%)	60.43 (+1.82%)
3	73.61 (-1.88%)	80.88 (+2.48%)	83.68 (-0.36%)	79.16 (+1.16%)	82.17 (+1.68%)	85.87 (+1.19%)	60.79 (+2.43%)
4	73.35 (-2.23%)	81.29 (+3.00%)	83.40 (-0.69%)	79.19 (+1.20%)	82.32 (+1.87%)	**86.06 (+1.41%)**	61.06 (+2.88%)
5	73.27 (-2.33%)	81.35 (+3.08%)	83.32 (-0.79%)	79.17 (+1.18%)	82.48 (+2.07%)	85.90 (+1.23%)	61.14 (+3.02%)
6	**73.15 (-2.49%)**	**81.54 (+3.32%)**	**83.30 (-0.81%)**	**79.23 (+1.25%)**	**82.65 (+2.28%)**	86.03 (+1.38%)	**61.34 (+3.35%)**

annotations entirely out of the training phase and evaluated on the REFUGE2 optic cup segmentation task. SPA achieves a 75% Dice Score in 5.63 iterations on average, with other models requiring more than six iterations to reach similar accuracy. For a Dice Score of 84%, SPA requires an average of 7.69 iterations, surpassing the second-best model, SAM-U V2, by 0.64 iterations.

SPA's improved efficiency is further substantiated with human user study. This user study is based on the REFUGE dataset, conducted by two PhD students with medical expertise. One individual required an average of 6.62 seconds and 4.33 iterations per image with MedSAM, but only 4.34 seconds and 2.46 iterations with SPA. Similarly, the other individual took 7.40 seconds and 5.60 iterations per image with MedSAM, but just 4.77 seconds and 3.58 iterations with SPA. This significant reduction in both time and interactions demonstrates SPA's superior efficiency.

4.3.3. Prediction Alignment with Clinician Feedback in Model Adaptation

Table 2 shows prediction alignment results for optic cup segmentation on the REFUGE2 dataset, comparing SPA's predictions with individual clinicians' annotations after each iteration. In this case, the ground truth is defined as the weighted average of annotations from Clinicians 2, 4, 5, 6, and 7, excluding Clinicians 1 and 3. After each interaction, the Dice Scores for the included clinicians consistently improve, indicating that SPA is adapting to the desired context. For example, Clinician 2's score increases from 78.92% to 81.54% (+3.32%), with similar positive trends for Clinicians 4, 5, 6, and 7 (+ 1.25%, +2.28%, +1.38%, and +3.35%, respectively). In contrast, the Dice Scores for the excluded clinicians, such as Clinicians 1 and 3, show a consistent decline (-2.49% and -0.81%), indicating that the model is effectively excluding irrelevant clinicians. This multi-choice-based refinement demonstrates SPA's ability to align predictions with the included clinicians while excluding those not part of the ground truth. Visualization of this alignment are provided in Appendix.

4.3.4. Ablation Study

In this section, we conducted an ablation study on key components of SPA, including random sampling from the preference distribution (Random Gaussian), updating the mean and variance (Gaussian), and adjusting distribution weights (Mixture Gaussian). Table 3 shows the ablation results, with segmentation performance evaluated by Dice Score on the REFUGE2 and Kidney datasets after three iterations.

When only randomly sampling is used without calibrating the mean, variance, or weight of the preference distribution, the Dice Scores are 80.29% for REFUGE2 and 90.05% for the Kidney dataset. Training the mean and variance improves the scores to 84.12% and 92.06%, respectively. Additionally, training the distribution weights alone raises the scores to 82.87% for REFUGE2 and 92.29% for Kidney. Finally, combining all three components which results to calibrating distribution mean, variance, and weight to form the preference distribution yields the highest performance, with Dice Scores of 85.42% for REFUGE2 and 94.26% for the Kidney dataset. This highlights the complementary benefits of each component in achieving optimal segmentation performance.

Table 3. **Effectiveness of Network Modules in SPA.** Ablation study evaluating the impact of network components on segmentation performance for the REFUGE2 and Kidney datasets after three iterations. The table compares random sampling (Random Gaussian), updating the mean and variance (Gaussian), and updating distribution weights (Mixture Gaussian). Combining all modules achieves the highest Dice Scores, highlighting their complementary benefits in optimizing segmentation performance.

Random Gaussian	Gaussian	Mixture Gaussian	REFUGE2	Kidney
✓			80.29	90.05
✓	✓		84.12	92.06
✓		✓	82.87	92.29
✓	✓	✓	**85.42**	**94.26**

5. Conclusion

In this work, we introduce SPA, a novel segmentation framework that efficiently adapts to diverse user preferences with minimal human effort. By offering clinicians multi-choice

options based on image uncertainties at interactions, SPA reduces clinician workload and ensures preference-specific predictions. The proposed preference distribution allows the model to dynamically adapt to clinician feedback during inference, accelerating convergence and enhancing interaction efficiency. Extensive experiments show that SPA outperforms deterministic, uncertainty-aware, and interactive SOTA models, demonstrating strong adaptability across various clinical contexts while requiring significantly less time and effort. These results highlight SPA's potential to improve clinical workflows in real-world medical applications.

References

[1] Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J. R. van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.-Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical Physics*, 38(2):915–931, 2011. 5

[2] Byron Ballou, Gregory W. Fisher, Thomas R. Hakala, and Daniel L. Farkas. Tumor Detection and Visualization Using Cyanine Fluorochrome-Labeled Antibodies. *Biotechnology Progress*, 13(5):649–658, 1997. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1021/bp970088t. 1

[3] Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 119–127, Cham, 2019. Springer International Publishing. 1, 3

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation, 2021. arXiv:2105.05537 [cs, eess]. 6

[5] Heang-Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou. Deep Learning in Medical Image Analysis. In *Deep Learning in Medical Image Analysis : Challenges and Applications*, pages 3–21. Springer International Publishing, Cham, 2020. 2

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021. arXiv:2102.04306 [cs]. 1, 2, 6

[7] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. FocalClick: Towards Practical Interactive Image Segmentation, 2022. arXiv:2204.02574 [cs]. 1

[8] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013. 5

[9] Guoyao Deng, Ke Zou, Kai Ren, Meng Wang, Xuedong Yuan, Sancong Ying, and Huazhu Fu. SAM-U: Multi-box prompts triggered uncertainty estimation for reliable SAM in medical image, 2023. arXiv:2307.04973 [cs]. 3, 6

[10] Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, Baiying Lei, Benjian Zhao, Xinxing Xu, Shaohua Li, Francisco Fumero, José Sigut, Haidar Almubarak, Yakoub Bazi, Yuanhao Guo, Yating Zhou, Ujjwal Baid, Shubham Innani, Tianjiao Guo, Jie Yang, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. REFUGE2 Challenge: A Treasure Trove for Multi-Dimension Analysis and Evaluation in Glaucoma Screening, 2022. arXiv:2202.08994 [cs, eess]. 5

[11] Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. Who Said What: Modeling Individual Labelers Improves Classification, 2018. arXiv:1703.08774 [cs]. 3, 6

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, 2021. arXiv:2111.06377 [cs]. 4

[13] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, 32(4):582–596, 2019. 1, 2

[14] 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. 3

[15] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving Uncertainty Estimation in Convolutional Neural Networks Using Inter-rater Agreement. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 540–548, Cham, 2019. Springer International Publishing. 3, 6

[16] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12336–12346, 2021. ISSN: 2575-7075. 3, 6

[17] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, 2017. arXiv:1703.04977 [cs]. 1

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, 2023. arXiv:2304.02643 [cs]. 1, 3, 4, 6

[19] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009. 3

[20] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images, 2018. 3, 6

[21] Simon A. A. Kohl, Bernardino Romera-Paredes, Klaus H. Maier-Hein, Danilo Jimenez Rezende, S. M. Ali Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A Hierarchical Probabilistic U-Net for Modeling Multi-Scale Ambiguities, 2019. arXiv:1905.13077 [cs]. 5

[22] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images, 2019. arXiv:1806.05034 [cs, stat]. 1

[23] Hongwei Bran Li, Fernando Navarro, Ivan Ezhov, Amirhossein Bayat, Dhritiman Das, Florian Kofler, Suprosanna Shit, Diana Waldmannstetter, Johannes C. Paetzold, Xiaobin Hu, Benedikt Wiestler, Lucas Zimmer, Tamaz Amiranashvili, Chinmay Prabhakar, Christoph Berger, Jonas Weidner, Michelle Alonso-Basant, Arif Rashid, Ujjwal Baid, Wesam Adel, Deniz Ali, Bhakti Baheti, Yingbin Bai, Ishaan Bhatt, Sabri Can Cetindag, Wenting Chen, Li Cheng, Prasad Dutand, Lara Dular, Mustafa A. Elattar, Ming Feng, Shengbo Gao, Henkjan Huisman, Weifeng Hu, Shubham Innani, Wei Jiat, Davood Karimi, Hugo J. Kuijf, Jin Tae Kwak, Hoang Long Le, Xiang Lia, Huiyan Lin, Tongliang Liu, Jun Ma, Kai Ma, Ting Ma, Ilkay Oksuz, Robbie Holland, Arlindo L. Oliveira, Jimut Bahan Pal, Xuan Pei, Maoying Qiao, Anindo Saha, Raghavendra Selvan, Linlin Shen, Joao Lourenco Silva, Ziga Spiclin, Sanjay Talbar, Dadong Wang, Wei Wang, Xiong Wang, Yin Wang, Ruiling Xia, Kele Xu, Yanwu Yan, Mert Yergin, Shuang Yu, Lingxi Zeng, YingLin Zhang, Jiachen Zhao, Yefeng Zheng, Martin Zukovec, Richard Do, Anton Becker, Amber Simpson, Ender Konukoglu, Andras Jakab, Spyridon Bakas, Leo Joskowicz, and Bjoern Menze. QUBIQ: Uncertainty Quantification for Biomedical Image Segmentation Challenge, 2024. arXiv:2405.18435 [cs, eess]. 5

[24] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. SimpleClick: Interactive Image Segmentation with Simple Vision Transformers, 2023. arXiv:2210.11006 [cs]. 1

[25] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment Anything in Medical Images. *Nature Communications*, 15(1):654, 2024. arXiv:2304.12306 [cs, eess]. 3, 6

[26] Xueyan Mei, Hao-Chih Lee, Kai-yue Diao, Mingqian Huang, Bin Lin, Chenyu Liu, Zongyu Xie, Yixuan Ma, Philip M. Robson, Michael Chung, Adam Bernheim, Venkatesh Mani, Claudia Calcagno, Kunwei Li, Shaolin Li, Hong Shan, Jian Lv, Tongtong Zhao, Junli Xia, Qihua Long, Sharon Steinberger, Adam Jacobi, Timothy Deyer, Marta Luksza, Fang Liu, Brent P. Little, Zahi A. Fayad, and Yang Yang. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nature Medicine*, 26(8):1224–1228, 2020. Publisher: Nature Publishing Group. 2

[27] Pataje G.S. Prasanna, Helen B. Stone, Rosemary S. Wong, Jacek Capala, Eric J. Bernhard, Bhadrasain Vikram, and C. N. Coleman. Normal tissue protection for improving radiotherapy: Where are the Gaps? *Translational cancer research*, 1 (1):35–48, 2012. 1

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015. arXiv:1505.04597 [cs]. 1, 2, 6

[29] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses, 2017. arXiv:1612.00197 [cs]. 1

[30] Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J. Erickson. Interactive segmentation of medical images through fully convolutional neural networks, 2019. arXiv:1903.08205 [cs]. 3

[31] Konstantin Sofiiuk, Ilia A. Petrov, and Anton Konushin. Reviving Iterative Training with Mask Guidance for Interactive Segmentation, 2021. arXiv:2102.06583 [cs]. 1

[32] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016. arXiv:1706.00712 [cs]. 1, 2

[33] Guotai Wang, Wenqi Li, Maria A. Zuluaga, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Interactive Medical Image Segmentation using Deep Learning with Image-specific Fine-tuning. *IEEE Transactions on Medical Imaging*, 37(7):1562–1573, 2018. arXiv:1710.04043 [cs]. 3

[34] Jian Wang, Hengde Zhu, Shui-Hua Wang, and Yu-Dong Zhang. A Review of Deep Learning on Medical Image Analysis. *Mobile Networks and Applications*, 26(1):351–380, 2021. 2

[35] Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jiawei Chen, Kim-Han Thung, Toan Duc Bui, Jitae Shin, Guodong Zeng, Guoyan Zheng, Vladimir S. Fonov, Andrew Doyle, Yongchao Xu, Pim Moeskops, Josien P.W. Pluim, Christian Desrosiers, Ismail Ben Ayed, Gerard Sanroma, Oualid M. Benkarim, Adrià Casamitjana, Verónica Vilaplana, Weili Lin, Gang Li, and Dinggang Shen. Benchmark on Automatic 6-month-old Infant Brain Segmentation Algorithms: The iSeg-2017 Challenge. *IEEE transactions on medical imaging*, page 10.1109/TMI.2019.2901712, 2019. 2

[36] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation, 2023. arXiv:2304.12620 [cs]. 3, 6

[37] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. Disentangling Human Error from the Ground Truth in Segmentation of Medical Images, 2020. arXiv:2007.15963 [cs]. 3

[38] Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical SAM 2: Segment medical images as video via Segment Anything Model 2, 2024. arXiv:2408.00874 [cs]. 3

[39] Wentao Zhu, Yufang Huang, Liang Zeng, Xuming Chen, Yong Liu, Zhen Qian, Nan Du, Wei Fan, and Xiaohui

Xie. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical Physics*, 46(2):576–589, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13300. 2

[40] Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. A Review of Uncertainty Estimation and its Application in Medical Imaging, 2023. arXiv:2302.08119 [cs, eess]. 3