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

Supplementary Material

1. Motivation Details

Table 1. A preliminary experiment testing the impact of individual clinicians, conducted for the optic cup segmentation on REFUGE2 test set under U-Net's structure with Dice Score (%). The results indicate that the segmentation performance is consistent for individual clinician but varies significantly across different clinicians.

	Clinician 1	Clinician 2	Clinician 3	Clinician 4
Model 1	**71.28**	60.28	47.67	52.25
Model 2	61.06	**66.46**	63.30	63.84
Model 3	52.30	62.29	**69.30**	64.06
Model 4	53.05	62.72	65.45	**67.19**

Medical image uncertainty is often reflected in the varying user preferences, causing inconsistent annotations between different clinicians. To further explore this, we conducted a preliminary experiment to quantitatively demonstrate that individual clinicians exhibit consistent segmentation patterns, while significant variation exists between different clinicians. In this experiment, we trained U-Net models [5] using each clinician's annotations for optic cup segmentation with a subset of the REFUGE2 dataset. This resulted in four distinct models (Model 1-4), each corresponding to a different clinician (Clinician 1-4). Table 1 shows the segmentation performance of each model when evaluated against different clinicians. Notably, each model performed best when trained and tested with the same clinician's annotations, but its performance dropped significantly when evaluated against other clinician's annotations.

These observations suggest that each clinician exhibits a distinct and consistent annotation pattern, which directly influences the performance of the segmentation model. For example, Model 1 achieved its best performance when tested on Clinician 1's annotations, with a Dice Score of 71.28%. However, its performance decreased substantially when evaluated against the annotations of Clinician 2, 3, or 4, with scores as low as 47.67%. This trend persisted across all models, indicating that segmentation performance declines significantly when a model is trained on one clinician's annotations and tested on another's.

As each clinician's annotation behavior is informed by their specific preference, this finding suggests that adapting to these preference-driven behaviors could improve preference-specific segmentation predictions. We further hypothesize that not only are annotation patterns consistent within individual clinicians, but their interaction behaviors during interactive segmentation are also likely to be consistent. This hypothesis motivated the development of the SPA model, which is designed to adaptively learn and adjust to each clinician's specific preference through human interactions. By dynamically incorporating clinician feedback, SPA refines the segmentation process in response to individual interactions, aligning the model's predictions with the individual clinician preference.

2. Theoretical Proof

Let $D = \{\mathbf{r_u}^{(j)}\}_{j=1}^{J}$ represent J interactions from user u, where each interaction $\mathbf{r_u}^{(j)}$ is generated i.i.d. from a specific component $\mathcal{N}(\mu_u, \sigma_u^2)$. The posterior probability that the samples (interactions) D comes from user (Gaussian component) p is given by:

$$P(U = u \mid D) = \frac{\pi_u \prod_{j=1}^{J} N(\mathbf{r_u}^{(j)} \mid \mu_u, \sigma_u^2)}{\sum_{m=1}^{M} \pi_m \prod_{j=1}^{J} N(\mathbf{r_u}^{(j)} \mid \mu_m, \sigma_m^2)} \tag{1}$$

The average log-likelihood of each interaction belonging to user u is:

$$\frac{1}{J} \log L_u(D) = \frac{1}{J} \log \pi_u + \frac{1}{J} \sum_{j=1}^{J} \log N(\mathbf{r_u}^{(j)} \mid \mu_u, \sigma_u^2) \tag{2}$$

As $J \to \infty$, the empirical average converges to the expected value under the true distribution $\mathcal{N}(\mu_u, \sigma_u^2)$:

$$\lim_{J\to\infty} \tfrac{1}{J} \textstyle\sum_{j=1}^{J} \log N(\mathbf{r_u}^{(j)} \mid \mu_u, \sigma_u^2) = \mathbb{E}_{\mathbf{r_u} \sim N(\mu_u, \sigma_u^2)} \left[\log N(\mathbf{r_u} \mid \mu_u, \sigma_u^2)\right] \tag{3}$$

The expected log-likelihood difference between user u and any other user $i \neq u$ is expressed as the negative Kullback-Leibler (KL) divergence:

$$\mathbb{E}_{\mathbf{r_u} \sim N(\mu_u, \sigma_u^2)} \left[\log N(\mathbf{r_u} \mid \mu_u, \sigma_u^2) - \log N(\boldsymbol{\Delta}_{\mathbf{r_u}} \mid \mu_i, \sigma_i^2)\right] = -D_{\mathrm{KL}} \left(N(\mu_u, \sigma_u^2) \,\|\, N(\mu_i, \sigma_i^2)\right) \tag{4}$$

Combining this with Equation 2, we obtain the likelihood ratio between user u and any other user $i \neq u$:

$$\frac{L_u(D)}{L_i(D)} = e^{-QD_{\mathrm{KL}}(N(\mu_u, \sigma_u^2) \,\|\, N(\mu_i, \sigma_i^2))} \cdot \frac{\pi_u}{\pi_i} \tag{5}$$

Thus, the posterior distribution from Equation 1 for user u becomes:

$$P(U = u \mid D) = \left[1 + \textstyle\sum_{i\neq u} \left(\frac{\pi_u}{\pi_i} e^{-QD_{\mathrm{KL}}(N(\mu_u, \sigma_u^2) \,\|\, N(\mu_i, \sigma_i^2))}\right)\right]^{-1} \tag{6}$$

Since $\mu_i \neq \mu_u$ or $\sigma_i^2 \neq \sigma_u^2$ for $i \neq u$, we have $D_{\mathrm{KL}}(N(\mu_u, \sigma_u^2) \,\|\, N(\mu_i, \sigma_i^2)) > 0$ [1]. As $J \to \infty$,

[1] KL divergence property: $D_{\mathrm{KL}}(A \,\|\, B) \geq 0, D_{\mathrm{KL}}(A \,\|\, B) = 0$ iff $A = B$

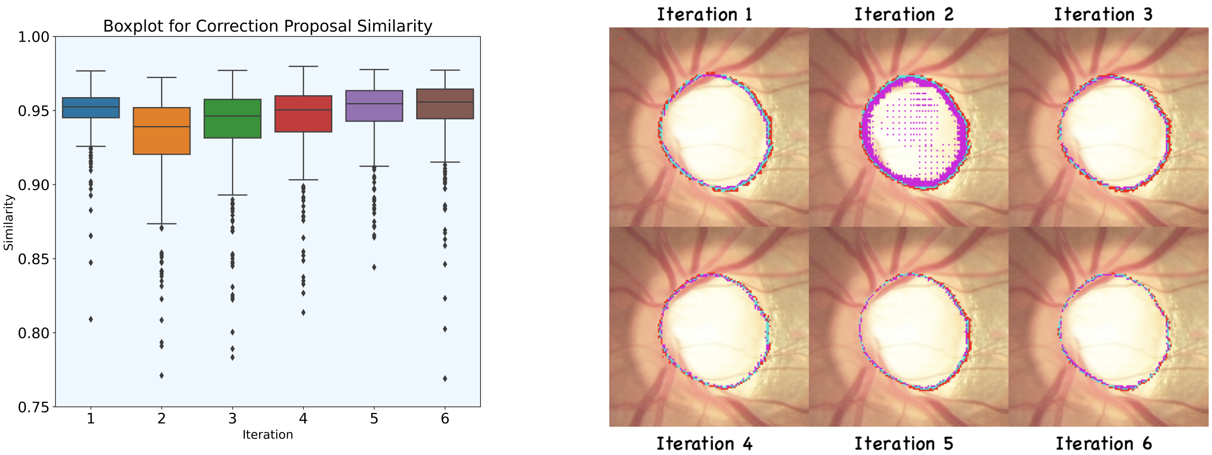

Figure 1. **Representative Segmentation Candidates Converge Over Iterations.** (a) Boxplot illustrating the similarity (measured by Dice Score) between representative segmentation candidates across multiple iterations. (b) Visualization showing the disagreement among four correction candidates at each iteration. Red regions indicate areas recognized by only one candidate as part of the target, light blue by two candidates, and purple by three candidates. As iterations progress, areas of disagreement shrink and alignment increases, reflecting the model's adaptation to user interactions and the reduction of uncertainty over time, especially from iterations 2 to 6.

$e^{-JD_{\text{KL}}} \to 0$ for each $i \neq u$, leading to $P(U = u \mid D) \to 1$. In other words, as $J \to \infty$, $p_\theta(z) \to N(\mu_u, \sigma_u^2)$. Therefore, the preference distribution $p_\theta(z)$ can adapt to specific user preferences based on new interactions. Since there is no closed-form solution for updating $\theta = \{(\mu_m, \sigma_m^2, \pi_m)\}_{m=1}^M$, we use MLP blocks with six forward layers and ReLU activations to adjust these parameters, refining the preference distribution $p_\theta(z)$ based on interactions. The preference distribution $p_\theta(z)$ thus adapts effectively to individual user preferences through interactions, enabling a personalized segmentation model that aligns closely with diverse clinical contexts and user expectations.

3. Implementation Details

To capture image uncertainty, we generate $N = 48$ predictions by sampling from the preference distribution $p_\theta^{(j)}(z)$. 48 predictions is experimentally the best to balance model performance and computational cost. The Dice Score for generating 24, 36, 48 predictions on REFUGE2 after six iterations are 83.05%, 84.70% and 86.22%. Therefore, we choose to sample 48 predictions as the hyperparameter. Additionally, we generate $K = 4$ representative segmentation candidates to allow users to make a multiple-choice selection. This hyperparameter is set to 4 because it is commonly used in high-stakes tests and is practical in various scenarios [1]. As a result, 4 is chosen as the number of representative segmentation candidates in our medical setting. In addition, we sample the most representative point from the representative segmentation candidate to prevent overfitting, although the segmentation candidates can be directly used. The initial interaction is uniformly selected from the area agreed upon by all annotators. It mimics real-world user behavior, where users first identify a rough target or shape and then refine the boundaries.

In this work, we face the challenge of medical image uncertainty, particularly when using multi-user annotated datasets such as REFUGE2, LIDC-IDRI, and QUBIQ. Each user's annotation reflects their unique interpretations, leading to inherent uncertainty in segmentations. To establish a reliable ground truth for evaluating our SPA method, we adopt a strategy that combines annotations from multiple users. This approach allows us to capture the medical image uncertainty associated with varying human preferences while minimizing potential biases. It ensures model robustness and generalisability.

The use of multi-user datasets is central to modeling medical image uncertainty. While individual user annotations are valuable, they can introduce personal biases or errors and may not fully represent the range of clinical contexts or human preferences. By combining multiple user annotations, we are able to account for the diverse preferences that contribute to uncertainty in medical images. This combination reflects a more balanced and representative ground truth, reducing the impact of any single user's subjective interpretation. Additionally, when only a limited number of users (e.g., four in the LIDC dataset) are available, it becomes crucial to use their combined annotations to better simulate the diversity of clinical contexts. By integrating multiple user annotations, we more accurately represent shifts in annotation conventions and clinical decision-making processes that occur across different medical environments, thereby capturing the uncertainty inherent in medical image analysis.

During the testing stage, for each image, we randomly generated combinations of user annotations from the multi-user annotated datasets. When four users provided annotations, all possible combinations (individual, pairs, triplets, and full set) were considered, resulting in 15 combinations (4 individual, 6 pairs, 4 triplets, and 1 full set). These combinations were uniformly selected, ensuring equal representation of all potential groupings. The selected annotations were then fused to create a consensus segmentation. This fusion involved averaging the chosen annotations to generate a probability map, which was subsequently binarized to form the final segmentation mask. By incorporating multiple user perspectives, this fused binary segmentation reflects the inherent uncertainty in the data and was used as the ground truth for evaluating the SPA model.

4. Efficiency Analysis on Different Interactive Models Details

Models that failed to reach the target Dice Score within the limit were assigned an iteration count of ten. We provide the failure rate statistics for REFUGE2 dataset reaching Dice 70% and 80%, for LIDC reaching Dice 60% and 70% in Table 4. It shows that our method, SPA, consistently achieves fewer failure cases across datasets and thresholds.

Table 2. Failure rate for REFUGE2 reaching Dice 70% and 80%, for LIDC reaching Dice 60% and 70%.

	SAM		MedSAM		MSA		SAM-U V1		SAM-U V2		SPA	
REFUGE2	10.2%	28.2%	11.7%	29.2%	9.0%	27.2%	10.5%	33.0%	15.0%	38.7%	**4.7%**	**17.0%**
LIDC	29.4%	32.5%	27.2%	29.8%	29.1%	31.9%	31.9%	34.9%	27.9%	31.3%	**7.0%**	**13.0%**

In addition, to further verify the generalization and robustness, we further demonstrate the number of iterations required to reach specific Dice Score for different unseen annotators in Table 3 . It shows that our approach, SPA, consistently requires less iterations than other interactive models to reach specific Dice Score on the REFUGE2 dataset, regardless of the annotator.

Table 3. Number of iterations required to reach Dice 75% and 84% toward different unseen annotators' (A, B, C).

	SAM		MedSAM		MSA		SAM-U V1		SAM-U V2		SPA	
A	6.85	8.70	7.39	9.31	6.78	8.76	6.90	8.64	6.25	8.33	**5.64**	**7.69**
B	2.76	5.87	2.09	5.03	2.49	5.57	2.70	5.42	3.22	6.79	**1.89**	**4.38**
C	1.72	4.08	1.66	3.58	1.56	3.64	1.66	3.99	1.53	3.57	**1.50**	**3.01**

5. Representative Segmentation Candidate Similarity across Interactions

In this section, we explore how the similarity between representative segmentation candidates evolves over multiple iterations of human interaction. Our multi-choice approach generates distinct segmentation candidates after each iteration, enabling users to guide the refinement process by selecting the segmentation candidate that they find the most appropriate based on their preference. This analysis investigates how these representative segmentation candidates change as the model adapts to user feedback over time. For illustration, we use the optic cup segmentation task from the REFUGE2 dataset. To facilitate comparison, we directly used the raw K-means clustering results applied to the N predictions $\{\hat{y}_n^{(j)}\}_{n=1}^N$. This approach allows us to evaluate the inherent divergence or convergence of the candidates without any post-processing adjustments. To quantify the similarity between candidates, we computed the Dice Score as the similarity matrix between each pair of K-means-generated segmentation candidates for every image at each iteration.

Fig. 1 illustrates the evolution of representative segmentation candidate similarity across six iterations. The boxplot on the left displays the range of similarity scores across all images, while the plot on the right provides a visual representation of areas where the four correction candidates disagree. In this visualization, red areas correspond to regions identified by only one candidate as part of the target, light blue by two candidates, and purple by three candidates.

In the first iteration, where no user feedback is involved, the segmentation candidates are relatively similar, resulting in a median similarity of approximately 0.952. This is because the model generates predictions based on the same input features, leading to only minor variations among the representative segmentation candidates. The visualization (Fig. 1) confirms this, showing minimal divergence between candidates, with areas of disagreement being small and concentrated mainly at the optic cup boundaries. After the first human interaction (Iteration 2), the similarity between candidates decreases significantly as the user's feedback introduces new corrections based on their preference. This feedback causes the representative segmentation candidates to diverge, contributing to different plausible segmentations. The median similarity score drops to 0.939, as shown by the noticeable increase in areas of disagreement (Fig. 1), particularly at the optic cup boundary. The diversity in candidates at this stage reflects the model's flexibility in generating a range of segmentations in response to user interactions.

As the human interaction process continues, our model refines its predictions based on user feedback. The correction candidates gradually converge as the model learns from the user's interaction and moves toward a more specific, preference-aligned segmentation. This steady convergence is evident in iterations 3 through 6 in Fig. 1, where the similarity scores gradually increase. By the final iteration, the candidates exhibit a high median similarity of 0.956, surpassing that of the initial iteration. The visualization on the right (Fig. 1) demonstrates this convergence as the areas of disagreement shrink significantly, particularly at the boundary areas. This indicates that the model has effectively adapted to the user's interaction, reducing uncertainty in its predictions and converging toward a consistent segmentation aligned with the human preference.

This analysis provides valuable insights into how human

Table 4. **SPA Achieves Superior Dice Score Improvements Across Iterations.** Quantitative comparison of Dice Score improvements between consecutive iterations for different interactive models. The “Overall Diff” column shows the total Dice Score improvement from Iteration 1 to Iteration 6. SPA consistently achieves the highest performance gains, demonstrating its effectiveness in incorporating user interaction for segmentation refinement.

Model	Diff 1	Diff 2	Diff 3	Diff 4	Diff 5	Overall Diff
SAM [3]	-0.03	0.04	0.14	0.12	0.13	0.41
MedSAM [4]	0.43	**0.45**	0.04	0.06	0.04	1.02
MSA [6]	0.06	-0.01	0.01	0.05	-0.03	0.09
SAM-U V1 [2]	-0.09	-0.04	-0.04	0.05	0.05	-0.08
SAM-U V2 [2]	-0.11	0.00	0.06	-0.05	0.01	-0.09
SPA	**1.05**	0.34	**0.36**	**0.16**	**0.15**	**2.07**

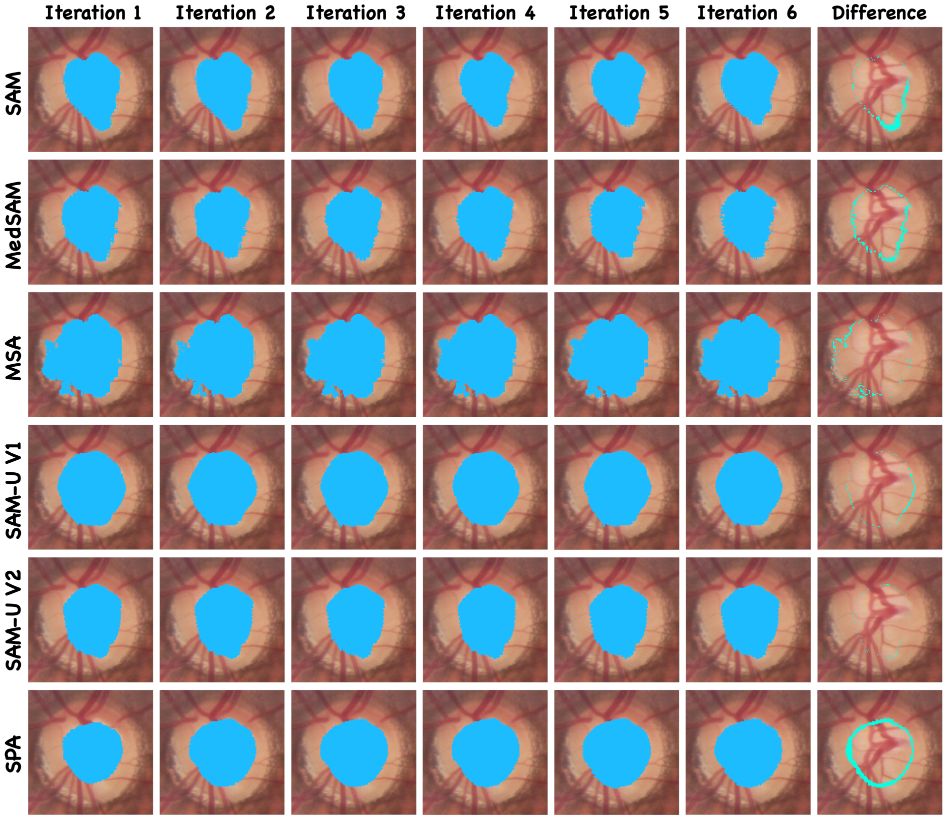

Figure 2. **SPA Outperforms Other Interactive Models in Prediction Refinement.** Visual comparison of predictions from SAM, MedSAM, MSA, SAM-U (V1, V2), and SPA models across six interaction iterations. The last column shows the difference between the final and initial predictions. SPA exhibits the most significant changes between iterations, indicating its greater sensitivity to user interaction and improved refinement of segmentation predictions.

interactions influence the evolution of correction candidates. Initially, the candidates are similar because they are generated by the model alone, without any external corrections from user preferences. However, once the user’s feedback is introduced, the candidates diverge to reflect different potential segmentations, capturing the information introduced through the user’s interaction. Over time, the correction candidates begin to converge as the model refines its predictions, steadily narrowing down the range of plausible segmentations and aligning them more closely with user preferences. This steady convergence reflects the success of our multi-choice correction candidate approach, which effectively incorporates human feedback to refine the model’s predictions toward more preference-specific segmentations.

6. Prediction Change After Interactions

In this section, we highlight the effectiveness of our SPA model's multi-choice correction candidate interaction strategy, demonstrating how it outperforms other interactive models in terms of prediction refinement. The quantitative results and visual examples illustrate how the model's sensitivity to clinician interactions leads to more substantial improvements over iterations compared to other interactive models like SAM [3], MedSAM [4], MSA [6], and SAM-U variants [2].

Table 4 quantifies the changes in Dice Score (%) between consecutive iterations for each model for the REFUGE2 optic cup segmentation task. "Diff 1" refers to the improvement from Iteration 1 to Iteration 2, "Diff 2" from Iteration 2 to Iteration 3, and so on. The final column shows the overall Dice Score difference between the first and last iterations for each model, serving as a cumulative measure of how much each model's performance improved throughout the interactive process. SPA consistently achieves higher Dice Score improvements compared to all other models across almost every iteration. For example, the first interaction yields a substantial Dice Score increase of 1.05% for SPA, whereas SAM, SAM-U, and MSA models exhibit marginal or even negative changes in performance. The overall difference for SPA is 2.07%, significantly higher than the next-best-performing model (MedSAM with 1.02%).

Fig. 2 shows an visual example case comparing the REFUGE2 optic cup segmentation predictions of various interactive models. Each row corresponds to a different model, while the columns display the prediction results at each interaction iteration (from Iteration 1 to Iteration 6). The final column represents the difference between the last and first predictions, highlighting how much each model's prediction has evolved due to clinician interaction. As shown in the last column in 2, SPA exhibits the most substantial change between the first and last predictions, indicating its high sensitivity to clinician interactions. This responsiveness suggests that the SPA model is more adept at refining its predictions based on the provided feedback, leading to better alignment with the user preference. In contrast, models such as SAM, MedSAM, and MSA show more limited changes, indicating less responsiveness to the interactive corrections provided by the clinician.

The combined quantitative and qualitative evidence underscores the superior effectiveness of SPA's multi-choice correction candidate approach. SPA is more responsive to clinician interactions, leading to larger adjustments in its predictions and greater alignment with the user preference. This higher sensitivity to clinician feedback, compared to other interactive models, results in more meaningful improvements in segmentation performance over time. SPA's capacity to integrate multiple correction candidates allows it to dynamically adjust its predictions in response to clinician interaction, enabling more effective refinement of segmentations. This adaptability makes it particularly suited for clinical environments where human interactions are critical for achieving preference-specific medical image segmentations.

7. Human User Study Details

In order to evaluate the efficiency of our SPA model compared to the previous interactive model, MedSAM, we conducted a detailed human evaluation study simulating real-world medical image segmentation workflows. Five medical professionals, each with over five years of graduate-level expertise, participated in the study. Their task was to interact with the models to refine predictions until they met their clinical standards.

For the MedSAM model, participants were provided with two types of prompts: *Click* and *BBox*. These prompts could be used to include or exclude specific pixels in the target segmentation, offering flexibility for achieving desired results. Participants were allowed to select the prompt type that best suited their needs for each scenario, simulating the decision-making process in clinical practice. In contrast, the SPA model introduced a multi-choice interface designed to streamline interactions. Instead of requiring manual pixel inclusion or exclusion, SPA presented participants with several correction candidates during each iteration. This design allowed users to select the option that most closely aligned with their expectations, reducing the cognitive and manual effort involved in refining predictions.

Participants interacted with each model iteratively, making adjustments until the predictions matched their desired criteria. Throughout the study, we recorded two key metrics for each model: (1) the total time required to achieve satisfactory results and (2) the number of interaction iterations needed to reach the final output. By averaging these metrics across multiple cases, we were able to quantify and compare the efficiency of SPA and MedSAM. This evaluation demonstrated the potential of SPA's uncertainty-aware, multi-choice framework to improve the user experience in medical image segmentation. The results suggest that SPA can significantly reduce the time and effort required to achieve accurate segmentation, while also offering greater adaptability to the needs of medical professionals.

8. Visualization for Prediction Alignment with Clinicians

Fig. 3 illustrates a visual comparison of the differences between SPA's segmentation predictions and individual clinicians' annotations over six iterations. The focus is on how the model's predictions evolve with human interaction, comparing included clinicians with excluded clinicians. In Iteration 1, the dark purple areas represent the initial differences between the model's predictions and the clinicians' anno-

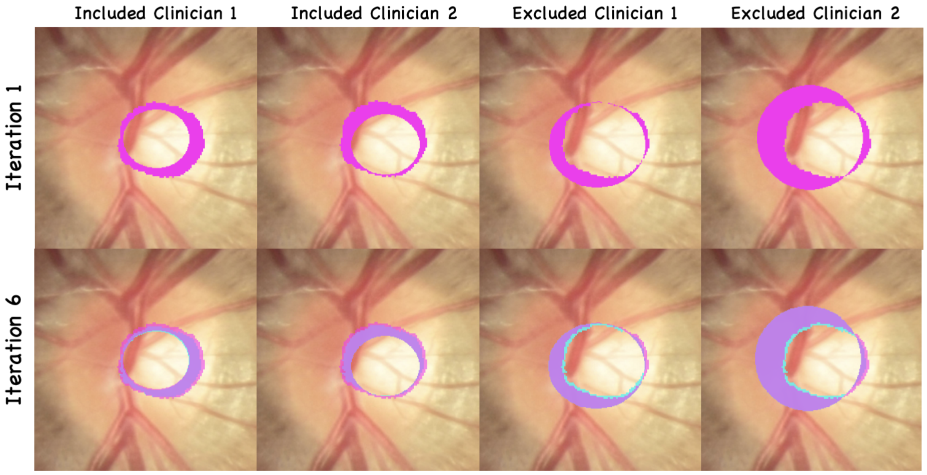

Figure 3. **SPA Demonstrates Visual Prediction Alignment with Clinician Annotations.** The figure illustrates the differences between SPA's optic cup segmentation predictions and individual clinicians' annotations in the REFUGE2 dataset across multiple interaction iterations. Dark purple represents the initial differences between the predictions and clinician annotations. Light purple indicates the overlap of differences between the first and sixth interactions, while blue shows new differences that emerge in the sixth iteration. The results demonstrate that the model increasingly aligns with included clinicians and diverges from excluded clinicians over time.

tations. By Iteration 6, the light purple regions show the overlap between the differences observed in Iteration 1 and the updated differences in Iteration 6, indicating which discrepancies remain consistent across iterations. The blue areas highlight new differences introduced in Iteration 6 that are not present in the first iteration. For included clinicians, the light purple areas shrink, signifying that the model's predictions are becoming more aligned with their annotations. In contrast, for excluded clinicians, the blue regions grow, showing that the model's predictions are diverging from their annotations. This visualization effectively demonstrates how the model refines its predictions over interactions, aligning more closely with the annotations of included clinicians while progressively moving away from those of excluded clinicians.

References

[1] Afsaneh Dehnad, Hayedeh Nasser, and Agha Fatemeh Hosseini. A comparison between three-and four-option multiple choice questions. *Procedia-Social and Behavioral Sciences*, 98:398–403, 2014. 2

[2] 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]. 4, 5

[3] 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]. 4, 5

[4] 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]. 4, 5

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

[6] 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]. 4, 5