Gaze-Assisted Medical Image Segmentation

Leila Khaertdinova[1] **Ilya Pershin**[1,2] **Tatiana Shmykova**[1,2] **Bulat Ibragimov**[3]
[1]Research Center for AI, Innopolis University
[2]Kazan Federal University [3]University of Copenhagen
l.khaertdinova@innopolis.university
{i.pershin,t.shmykova}@innopolis.ru
bulat@di.ku.dk

Abstract

The annotation of patient organs is a crucial part of various diagnostic and treatment procedures, such as radiotherapy planning. Manual annotation is extremely time-consuming, while its automation using modern image analysis techniques has not yet reached levels sufficient for clinical adoption. This paper investigates the idea of semi-supervised medical image segmentation using human gaze as interactive input for segmentation correction. In particular, we fine-tuned the Segment Anything Model in Medical Images (MedSAM), a public solution that uses various prompt types as additional input for semi-automated segmentation correction. We used human gaze data from reading abdominal images as a prompt for fine-tuning MedSAM. The model was validated on a public WORD database, which consists of 120 CT scans of 16 abdominal organs. The results of the gaze-assisted MedSAM were shown to be superior to the results of the state-of-the-art segmentation models. In particular, the average Dice coefficient for 16 abdominal organs was 85.8%, 86.7%, 81.7%, and 90.5% for nnUNetV2, ResUNet, original MedSAM, and our gaze-assisted MedSAM model, respectively. [1]

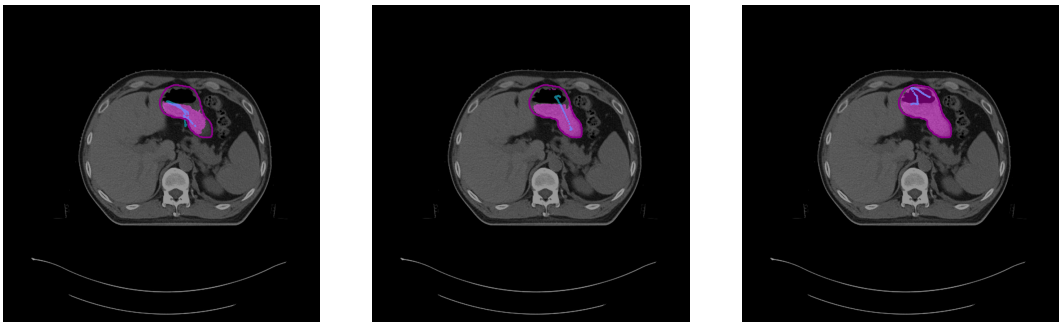
1 Introduction

The workload of radiologists has increased significantly over the past several decades, leading to concerns about the impact of fatigue on the accuracy of medical image interpretation [18, 6, 30, 26]. To mitigate the risk of decreased diagnostic quality, several strategies have been proposed, including limiting radiologists' workloads and adjusting the pace of image analysis [2]. Another viable approach involves optimizing radiologists' workflows, encompassing everything from the layout of reading rooms to team management practices [24]. Among these strategies, Artificial Intelligence (AI) emerges as a particularly promising solution.

Along with the potential prospects of AI, there are also significant drawbacks related to the explainability of AI models [27, 4] and the radiologists' bias towards AI [11, 8, 20]. These issues introduce challenges to the automation of the radiology workflow and the seamless integration of AI. Nonetheless, the emerging paradigm of hybrid intelligence, which combines human expertise with AI capabilities, provides a viable solution. Rather than relying on AI autonomously, hybrid intelligence emphasizes collaborative interactions between radiologists and AI to optimize their workflows. One promising method within this paradigm is the use of gaze analysis, a natural form of interaction given that medical visualization is central to radiology. Recently, multiple works in the research literature investigated the use of gaze data to improve deep learning models in medical image analysis [15, 32] and to extract patterns related to the increased workload [25].

[1]https://github.com/leiluk1/gaze-based-segmentation

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) The initial predicted mask by gaze data.

(b) The mask after half the correction time by gaze data.

(c) The mask at the end of the correction time by gaze data.

Figure 1: Steps to correct a stomach segmentation mask on a CT slice. Each subfigure shows the outline of reference segmentation contours, the predicted segmentation mask, and gaze points (blue) that are used for prediction.

One of the tasks frequently addressed in surgical and radiotherapy planning is medical image segmentation [19]. Precise segmentation masks are crucial for providing essential information that aids in clinical diagnosis and patient monitoring. Furthermore, organ segmentation is vital in therapeutic interventions, particularly in cancer treatments that rely on radiation therapy [28, 9]. The advent of deep learning has enabled the automatic segmentation of medical images, significantly simplifying the annotation process. These methods alleviate the challenges posed by the labour-intensive nature of manual segmentation, which often demands several hours of effort from annotators for each case [21]. However, segmentation inaccuracies often necessitate accurate manual corrections. Recent advances in interactive segmentation have made this correction process more efficient across various domains, including medicine, reducing the effort to just a few mouse clicks [33, 22, 34].

In this paper, we aim to incorporate eye gaze data into the interactive segmentation workflow. Specifically, we propose lightweight fine-tuning strategies of the existing interactive segmentation model, namely Segment Anything Model, or SAM [17, 22], using simulated eye gaze data. At inference time, utilizing streams of real gaze data instead of mouse clicks could provide a more intuitive and faster way for annotating 3D medical images, such as CT (Computed Tomography) and Magnetic Resonance Imaging (MRI) scans, which comprise a large number of slices [1]. Figure 1 illustrates an example of the gaze-based process for correcting a stomach segmentation mask on a single CT slice.

Overall, the main contributions of this study are listed as follows:

- We present a novel approach for the real-time correction of segmentation in CT scans based on sequential gaze information. More specifically, we suggest adapting existing interactive segmentation frameworks to incorporate gaze data as a prompt.
- We introduce and evaluate various fine-tuning options for the medical interactive segmentation model, namely MedSAM [22], by utilizing synthetic prompts that mimic eye gaze streams.
- Extensive experiments and ablations conducted by both human participants and medical experts demonstrate the effectiveness of the proposed method, compared to the existing methods, such as mouse clicks and bounding boxes. Additionally, we analyze the radiologist's behaviour during the segmentation process, focusing on eye movements and the time spent on experiments using eye-tracking technology.

2 Related work

Interactive segmentation. Interactive segmentation is a specific type of segmentation that leverages user interaction to enhance the segmentation process. Usually, users provide information for initial segmentation, such as bounding boxes or textual descriptions of the region of interest, and, if necessary, subsequently refine the quality of the predicted segmentation. With its history in the

pre-deep learning era, interactive segmentation has evolved from graph represented [5] and random walk [12] based approaches to contemporary foundation models [17, 35].

SAM [17] is the state-of-the-art interactive segmentation model, which provides impressive performance across diverse segmentation tasks that involve human interaction. The architecture of the model consists of a Vision Transformer-based image encoder, a prompt (mouse clicks coordinates, bounding boxes, textual descriptions) encoder, and a mask decoder. While SAM demonstrated impressive performance on natural images, the performance in clinical settings is weaker [23]. To address this, MedSAM [22] was fine-tuned on the large-scale medical dataset and adapted to medical image segmentation using bounding boxes.

Eye tracking technology in the medical field. Eye-tracking technology has emerged as an effective tool in the medical field, providing diverse applications across various disciplines [29], particularly in image segmentation. Sadeghi et al. [1] introduced a method that uses eye gaze information for interactive user-guided segmentation. Their approach involves three main steps: tracking eye gaze for seed pixel selection, using an optimization technique for image labeling based on user input, and integrating these elements to provide real-time visual feedback. This method has demonstrated significant speed-up in the process compared to traditional mouse controls. Besides, Gaze2Segment [7] combined biological and computer vision techniques to improve radiologists' diagnostic workflows, highlighting the potential of eye tracking in medical image analysis.

Medical segmentation. Medical segmentation is an essential and critical process for the identification, delineation, and visualization of regions of interest in medical images [13]. This procedure is vital for accurate diagnosis, treatment planning, and monitoring of diseases. Recent advancements in deep learning methods, along with user-guided approaches such as eye tracking, have significantly improved the interactivity and responsiveness of segmentation tools in the medical field [16, 1, 7].

To the best of our knowledge, there is currently only one gaze-assisted interactive medical segmentation model, which was presented recently [31]. In this study, the authors used the original SAM to predict a segmentation mask based on a single gaze point. The authors compared gaze-based and mouse-click segmentation approaches and demonstrated that gaze-based segmentation was approximately two times faster and had slightly worse quality than click-based segmentation. However, they did not explore the implications of using multiple gaze points. So, it is quite unclear what happens if the gaze leaves the region of interest. This phenomenon may occur due to natural gaze variance. Furthermore, the study did not address how users should proceed if they need to remove a region from a segmentation mask rather than add one. By highlighting these gaps, our study aims to further investigate the potential of gaze data in the interactive segmentation of medical images.

3 Methodology

In this section, we introduce a novel method for gaze-assisted interactive segmentation of medical images. We adapt MedSAM to 1) a gaze point sequence of arbitrary length, 2) the fixed label (foreground) for prompt points, and 3) the natural variance of the gaze data. Thus, during the segmentation process, gaze points are automatically transmitted to the selected model, allowing real-time corrections to the segmentation mask generated by the base model.

3.1 Problem definition

A 3D medical image refers to volumetric data that represents anatomical structures in three dimensions, typically acquired through imaging modalities such as CT or MRI. We validated the segmentation adjustment with gaze using 3D CT images, where each image $\boldsymbol{X} \in \mathbb{R}^{H \times W \times D}$ consists of D slices of shape $H \times W$. Thus, we transform CT scans to 2D slices for training pipeline and detailed analysis.

Medical segmentation aims to partition a medical image into meaningful regions, enabling the identification of anatomical structures and abnormalities. In the case of 3D CT, each image $\boldsymbol{X}$ is associated with mask $\boldsymbol{Y} \in \{0, 1\}^{H \times W \times D}$, where 1 corresponds to pixels containing the substance of interest, e.g. pancreas in abdominal CT, and 0 – to background points.

In this paper, we focus on interactive segmentation that leverages user interaction to refine initial segmentation masks. Specifically, we aim to utilize users' gaze to subsequently adjust the masks. Eye gaze data consists of gaze coordinates (x_t, y_t) for the currently viewed slice and are available for

each timestamp t. Our objective is to improve the segmentation mask in real-time using a sequence of gaze points $G_T = \{(x_t, y_t)\}_{t=0}^{T}$,

$$Q(M(X, G_T)) > Q(M(X, G_{T'})),$$

where Q is a quality metric (higher is better, to be specific), M is a gaze-assisted segmentation model generating mask $\boldsymbol{Y}$, and $G_{T'}$, where $T' < T$, is a previous sequence of gaze points. This formulation emphasizes the aim to improve segmentation quality by incorporating real-time gaze data.

3.2 Model training with eye gaze coordinates

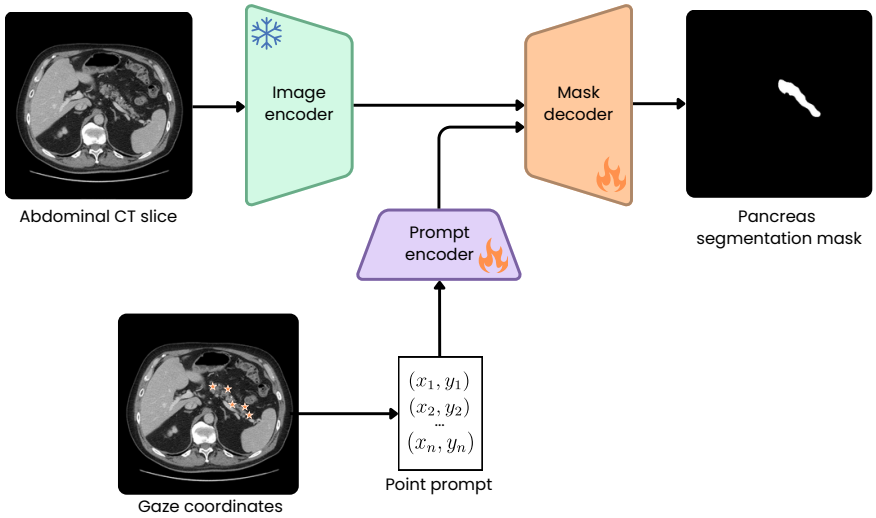


Figure 2: The proposed framework for gaze-assisted interactive segmentation of medical images. An illustrative example demonstrates the segmentation mask for the pancreas organ, which is predicted based on input gaze coordinates serving as a point prompt for the MedSAM model.

We introduce an approach for gaze-assisted interactive segmentation of medical images, which is depicted in Fig. 2. The proposed framework is designed to integrate eye gaze coordinates into the segmentation process. To achieve this, we trained the MedSAM model, specifically the mask decoder and the prompt encoder components. We generate synthetic gaze points, simulating eye gaze coordinates input prompt of arbitrary length. In this setup, the input gaze data acts as a prompt sent to the MedSAM model's prompt encoder, guiding the segmentation process.

Prompt generation from gaze data. Initially, we attempted to fine-tune the MedSAM model using point prompts generated only from the reference segmentation mask. Nevertheless, this method led to poor accuracy during inference. This can be explained by the fact that training with points from the ground truth mask made the model overly sensitive to areas outside the region of interest, which is not plausible for eye gaze streams. To address this issue, we utilize synthetic eye gaze coordinates to create a more realistic representation of gaze behaviour. In such a method, 80% of the generated points are chosen within the selected organ structure, specifically within the reference segmentation mask, while the remaining 20% are positioned outside of this mask. We find it quite reasonable to assume that, during the segmentation process, the majority of gaze points will accurately fall within the region of interest (the anatomical structure). However, considering that some gaze points may accidentally fall outside the target area, we allocate the remaining 20% of points outside the ground truth mask.

Secondly, we examine the dynamics of eye gaze movements, which are characterized by a series of rapid jumps or high-velocity movements, known as saccades, followed by fixations, which are located at periods in which the eye is stationary and remains relatively still [3]. By simulating these movements, we can generate random points that accurately reflect realistic gaze patterns. Alternatively, we can select random gaze coordinates, introducing some variability in the generated points. The latter approach is selected for generating point coordinates in prompts, as we believe it better reflects the natural fluctuations in gaze behaviour during real-world interactions.

Additionally, we propose another approach to generating point prompts based on mask differences. By analyzing the initial or base prediction, we create synthetic points that reflect the difference between the base prediction and the ground truth mask. In this approach, we allocate 70% of the points within the mask difference, 20% within the ground truth mask, and 10% outside the ground truth. However, these proportions warrant a more detailed exploration in future work to optimize the strategy for point allocation. We subsequently train the model using these generated points (refer to Fig. 3). These varied strategies enable the model to adapt to diverse gaze behaviours, thereby enhancing its segmentation capabilities through the incorporation of different input scenarios.

Fine-tuning MedSAM. As shown in Fig. 2, we use a typical encoder-decoder architecture of MedSAM for interactive segmentation with a prompt encoder in the bottleneck. The difference between gaze and other prompt sources (for example, mouse clicks or bounding boxes) is 1) the inability to manually indicate which gaze points correspond to the user's desire to add/remove/keep unchanged the segmentation masks, 2) the inability to decline segmentation changes, 3) imprecision of the gaze data.

While eye gaze data is generated as a sequence of coordinates for the fine-tuning pipeline, the SAM model requires the assignment of labels to input point coordinates, categorizing them as foreground (label 1), background (label 0), or not a point (label -1). However, justifying the inclusion or exclusion of specific points during experiments with an eye tracker poses challenges. To address this, we explored several labeling strategies for the point prompts.

Our strategies included: 1) randomly assigning points as either foreground or background, 2) removing point embedding and passing gaze coordinates to the prompt encoder without any associated label, and 3) assigning fixed labels (ones) to points and treating all gaze coordinates as foreground points. Ultimately, we found that all considered labeling strategies resulted in comparable performance. Following this, we choose to implement the latter, as it provides consistency for our experiments. We fine-tune the MedSAM model partially, specifically the prompt encoder and mask decoder, while keeping the other model's components frozen.

Interactive fine-tuning. The inspiration for the suggested approach is based on two facts: 1) network probability map can be used as a confidence of prediction; 2) gaze data provides the viewer's region of interest. Combining these facts allows us to create an approach to interactive gaze-assisted segmentation. As illustrated in Fig. 3, this methodology involves training the gaze points prompt encoder through simulations of human gaze behaviour.

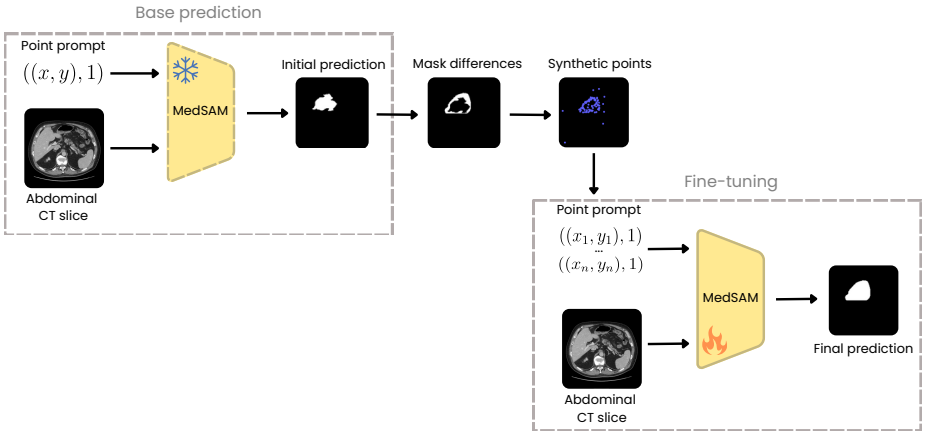


Figure 3: The provided pipeline outlines the steps for training a gaze-assisted segmentation model using synthetic gaze data points generated based on the mask difference approach. The process begins with the initial prediction from the frozen MedSAM. Using this initial prediction, we generate points that indicate the differences between the predicted mask and the ground truth, referred to as mask correction points. Next, the points are input into the same model to produce the final prediction.

Initially, a base prediction is generated using the point prompt within the ground truth mask. Subsequently, coordinates are derived based on the differences between the base predicted mask and the ground truth. These synthetic points are then utilized as prompts for the MedSAM prompt

encoder, with gaze coordinates assigned a label "foreground" passed into the encoder. This approach is predicated on the assumption that it effectively mimics authentic gaze behaviour. Additionally, we introduce a similar strategy that extends beyond the utilization of point prompts generated from mask differences. This involves incorporating the model's base prediction mask as an additional input to the prompt encoder, along with the generated points. During inference, the previous mask can be employed with the eye gaze points, thereby enhancing the robustness of the segmentation process.

3.3 Interactive segmentation

Our interactive segmentation method provides users with visualization of predicted segmentation masks that they can adjust base predictions using gaze. This iterative approach allows for precise corrections in areas where automated segmentation may have fallen short. In our experiments, we will evaluate the effectiveness of the proposed approach by asking an expert to correct segmentation masks for various anatomical structures, stopping the correction when the mask is close to ideal.

Radiology workstation. We developed an emulation of a radiologist workstation with eye-tracking capabilities. The workstation was equipped with a lightweight and easy-to-use eye-tracking device. The developed software sequentially showed medical images, recorded gaze points with timestamps, and communicated with an interactive gaze-assisted segmentation model to update the displayed segmentation mask.

4 Implementation details

4.1 Data

In this paper, we used the WORD dataset [21], which consists of 150 CT scans covering 16 abdominal organs. These CT scans were converted into 2D slices. Each slice was preprocessed to ensure that the corresponding ground truth mask contained a single distinct anatomical structure. Therefore, each organ can be represented as multiple structures when applicable. Similarly, to evaluate our models, we exploited the official test set of WORD.

4.2 Eye tracking hardware

We developed a framework that simulates radiological workstations for CT slice analysis, enhanced with eye-tracking capabilities. This workstation was set up in an isolated room at our institution and features an LG diagnostic 10-bit monitor with a resolution of 3840×2160 pixels and a pixel density of 7.21 px/mm, along with a Tobii Eye Tracker 4C that operates at a frequency of 90 Hz.

Calibration of the eye tracker is essential prior to experimentation to ensure accurate tracking of eye movements and alignment of gaze coordinates with the user's point of focus. Once calibrated, users can perform segmentation tasks using the software, with the entire process controlled via a single Enter button. The framework effectively captures and synchronizes eye movement data, predicted masks, and controller commands, enabling a thorough examination of gaze behaviour within a radiological context.

4.3 Training details

All our models were trained and tested using NVIDIA GTX 3090 Ti GPU. We employed the AdamW optimizer. We decreased the learning rate by 2 every 5 epochs with no improvement in validation loss. The initial learning rate was set at $5e^{-5}$, and the models were trained for 200 epochs. Additionally, we added early stopping with a patience of 10 epochs based on validation loss.

For the MedSAM model, we focused on training only the prompt encoder and mask decoder, while other components of the model architecture were kept frozen. Due to resource constraints, we fine-tuned our models on a random sample from the WORD dataset containing around 5% of all slices. Besides, the number of points in the point prompt was a critical hyperparameter. In this work, we report evaluations with 20 points, 50 points, and a random number of points ranging from 1 to 20 in the prompt. To incorporate a random number of points in the prompt for each sample during both training and testing, we padded some point prompts with a non-point label (-1).

Table 1: Comparison of segmentation performance (DSC (%)) with state-of-the-art models.

Method	nnUNetV2 [14]	ResUNet [10]	MedSAM [22]	Ours MedSAM
Liver	96.19 ± 2.16	**96.55 ± 0.89**	92.15 ± 1.22	95.14 ± 0.65
Spleen	94.33 ± 7.72	95.26 ± 2.84	78.68 ± 0.00	**95.58 ± 0.27**
Kidney (L)	91.29 ± 18.15	95.63 ± 1.20	83.40 ± 0.00	**95.73 ± 0.03**
Kidney (R)	91.20 ± 17.22	95.84 ± 1.16	91.93 ± 0.00	**95.89 ± 0.01**
Stomach	91.12 ± 3.60	91.58 ± 2.86	91.61 ± 0.00	**94.70 ± 0.49**
Gallbladder	83.19 ± 12.22	82.83 ± 11.8	85.43 ± 0.00	**90.63 ± 0.02**
Esophagus	77.79 ± 13.51	77.17 ± 14.68	**88.94 ± 5.61**	87.11 ± 2.70
Pancreas	83.55 ± 5.87	83.56 ± 5.60	79.78 ± 0.00	**89.23 ± 0.34**
Duodenum	64.47 ± 15.87	66.67 ± 15.36	80.07 ± 0.00	**86.53 ± 0.34**
Colon	83.92 ± 8.45	83.57 ± 8.69	75.96 ± 0.00	**89.62 ± 2.43**
Intestine	86.83 ± 4.02	86.76 ± 3.56	84.00 ± 0.00	**88.28 ± 3.28**
Adrenal	70.0 ± 11.86	70.9 ± 10.12	62.74 ± 5.28	**80.28 ± 1.41**
Rectum	81.49 ± 7.37	82.16 ± 6.73	**92.54 ± 2.33**	90.34 ± 1.65
Bladder	90.15 ± 16.85	91.0 ± 13.5	86.74 ± 0.00	**94.16 ± 0.83**
Head of Femur (L)	93.28 ± 5.12	**93.39 ± 5.11**	70.17 ± 0.00	93.19 ± 1.49
Head of Femur (R)	**93.93 ± 4.29**	93.88 ± 4.30	63.28 ± 0.00	93.00 ± 1.60
Mean	85.80 ± 5.27	86.67 ± 4.81	81.71 ± 9.39	**90.54 ± 5.20**

5 Experiments and results

5.1 Proof of concept: synthetic gaze coordinates

We developed a system for conducting experiments on synthetic gaze data. Unlike model validation during training, this system allowed for iterative testing of the model, simulating human responses to the model's predictions. After each prediction of the segmentation mask, we randomly generated gaze coordinates within the reference segmentation mask, in the areas of discrepancy between the predicted and reference masks, and outside of these areas. The gaze points were generated from a uniform random sample within the targeted areas, ensuring that no point was selected more than once (without replacement).

Strategies for selecting gaze data. We examined three general strategies for selecting gaze data during model inference. The first strategy involves accumulating gaze data across all iterations. Since the model was trained with a fixed size of the gaze point prompt, at each iteration, we randomly selected N gaze points from the entire accumulated history. The second strategy is based on a linear combination of the nearest points between neighboring iterations. At each iteration, we selected N points. For each point in the current iteration, we identified the nearest point from the set of points used to predict the current segmentation mask. Subsequently, we computed a linear combination between them. The coefficients of the linear combination, α and $(\alpha - 1)$ were used to amplify or attenuate the significance of information from previous iterations. The initial gaze points were utilized during the first iteration. The last strategy involves using a model trained with a random number of points in the prompt. At each iteration, we randomly selected a fixed number of gaze points and data obtained during that iteration, thereby increasing the number of gaze points in the prompt by a fixed amount.

To test the first strategy, we generated 20 gaze points with varying distributions within the ground truth (gt) segmentation mask, mask differences, and beyond these areas. Table 4, provided in the Appendix, demonstrates the average Dice Similarity Coefficient (DSC) for strategies with different gaze distributions. The highest average DSC of 0.9 is achieved when all gaze points are located within the reference segmentation mask.

To test the second strategy with a linear combination of points, we fixed the distribution of points (20% in the gt mask, 70% in the region of differences between the predicted and gt masks, and 10% outside both regions) and changed the coefficient α from 0.1 to 0.9 in the linear combination of points. The greatest DSC of 0.89 was obtained at $\alpha = 0.6$, while the lowest value of 0.86 was obtained at $\alpha = 0.9$.

Table 5, provided in the Appendix, presents the results of synthetic experiments conducted over 5 iterations with a model trained with a random number of points in the prompt. Different distributions of gaze points were tested between the reference segmentation mask, mask differences, and the remaining part of the image. The highest average DSC was achieved with a distribution of 20% of points in the area of the reference segmentation mask, 70% in the area of mask differences, and 10% outside these areas, with two points selected at each iteration.

Benchmarking. In this section, we validate the performance of our proposed approach by comparing our best model strategy with 2D contemporary models benchmarked on WORD. Our model was trained using 20 points in the prompt, 80% of the point coordinates located within the reference mask and the remaining 20% positioned outside of it. The model was subsequently tested using the same approach, with 20 generated points in the input prompt. Besides, we include the original MedSAM in our comparison, testing it with bounding boxes. The bounding boxes are generated using Otsu's method based on the gt masks.

The results, as shown in Table 1, indicate that our method consistently outperforms the original MedSAM with bboxes and other state-of-the-art 2D models across various abdominal organs. Notably, our approach demonstrates significant improvements in segmentation performance (DSC), particularly in challenging cases such as the duodenum and adrenal organs.

5.2 Human evaluation

For our experiments, we engaged a proxy radiologist. The experiment consisted of two parts. In the first part, the radiologist annotated a set of X CT slices using the proposed gaze-based interactive segmentation method. In the second part, the same dataset was annotated using the bounding box approach introduced in the MedSAM study [22].

Table 2: DSC in human experiments. The point prompt consists of 20 gaze points. The strategies of gaze data selection with random gaze points (p), fixation points with replacement (fix rep), fixation points with accumulation (fix accum), a linear combination of gaze points (linear), accumulation of 2 gaze points at each iteration (accum 2 p) are demonstrated. Additionally, the approach with segmentation using generated bounding boxes (MedSAM) is shown.

organ	points	fix rep	fix accum	linear	accum 2 p	MedSAM
Liver	0.942	0.626	0.655	0.888	0.958	0.967
Spleen	0.931	0.685	0.871	0.642	0.915	0.972
Kidney (L)	0.959	0.796	0.961	0.930	0.958	0.957
Kidney (R)	0.848	0.545	0.884	0.895	0.758	0.878
Stomach	0.870	0.828	0.951	0.944	0.906	0.941
Gallbladder	0.909	0.896	0.906	0.903	0.882	0.908
Esophagus	0.913	0.906	0.920	0.766	0.854	0.881
Pancreas	0.877	0.528	0.901	0.882	0.915	0.916
Duodenum	0.839	0.851	0.840	0.730	0.846	0.623
Colon	0.885	0.869	0.873	0.773	0.851	0.916
Intestine	0.786	0.729	0.778	0.833	0.797	0.801
Adrenal	0.784	0.764	0.765	0.796	0.772	0.793
Rectum	0.901	0.744	0.795	0.666	0.935	0.948
Bladder	0.965	0.069	0.113	0.940	0.966	0.945
Head of Femur (L)	0.749	0.873	0.871	0.782	0.902	0.945
Head of Femur (R)	0.871	0.861	0.914	0.855	0.869	0.740
Mean	0.854	0.764	0.831	0.820	0.861	0.884

Furthermore, we conducted a series of experiments involving the proxy radiologist, employing various gaze data selection strategies and different fine-tuned models. For the model trained with a prompt consisting of 20 points, the results are presented in Table 2. The approach of selecting gaze data based on 20 randomly accumulated points achieved an average Dice coefficient of 0.854. In contrast, the method based on selecting 20 random fixation points with replacement demonstrated an average Dice coefficient of 0.764, whereas the method based on selecting random fixation points with accumulation achieved an average Dice coefficient of 0.831. The approach with a linear combination of random

Table 3: DSC in human experiments. The point prompt consists of 20 gaze points. The model with the segmentation mask from the previous iteration in the prompt was used. The strategies of gaze data selection with random gaze points (p), fixation points with replacement (fix rep), and fixation points with accumulation (fix accum) are shown.

organ	points	fix rep	fix accum
Liver	0.661	0.253	0.640
Spleen	0.863	0.644	0.919
Kidney (L)	0.852	0.593	0.914
Kidney (R)	0.756	0.391	0.765
Stomach	0.715	0.729	0.751
Gallbladder	0.895	0.713	0.886
Esophagus	0.866	0.863	0.873
Pancreas	0.873	0.245	0.862
Duodenum	0.725	0.666	0.792
Colon	0.773	0.577	0.740
Intestine	0.657	0.638	0.671
Adrenal	0.667	0.423	0.491
Rectum	0.748	0.738	0.725
Bladder	0.765	0.306	0.725
Head of Femur (L)	0.809	0.712	0.824
Head of Femur (R)	0.764	0.771	0.764
Mean	0.745	0.601	0.745

points resulted in an average Dice coefficient of 0.82. Meanwhile, the method employing a model trained with a random number of points in the prompt, accumulating 2 random gaze points at each iteration, achieved an average Dice coefficient of 0.861. The original MedSAM with a bounding box prompt obtained an average Dice coefficient of 0.884. However, for certain organs, such as the right and left kidneys and the duodenum, the performance of MedSAM with bboxes was comparatively lower. The average time for the gaze-assisted segmentation approach with various gaze point selection strategies was 9.7 ± 4.9 seconds, compared to 5.7 ± 3.1 seconds required for MedSAM with bboxes. We provide visualization results of interactive prompts along with predictions made based on the expert's eye gaze in the Appendix (see Figure 4).

Additionally, we evaluated several gaze point selection strategies using a model trained on 20 points in the prompt along with the predicted segmentation mask from the previous iteration. Table 3 presents the segmentation quality metrics for each organ using gaze-assisted segmentation. The approach employing the strategy of selecting gaze data based on 20 randomly accumulated points achieved an average Dice coefficient of 0.745, while the method based on selecting 20 random fixation points with replacement yielded an average Dice coefficient of 0.601. The method using random fixation points with accumulation also resulted in an average Dice coefficient of 0.745.

5.3 Limitations

Our work has several limitations. 1) Despite the large number of human experiments, only one participant was involved. Future work should include more participants. 2) The participant segmented organs on individual slices. Future work should focus on using 3-dimensional medical data with the ability to manipulate slices.

6 Conclusion and future work

Medical image analysis has always been challenging, leaving room for improvement. The proposed gaze-assisted approach for organ segmentation in abdominal CT scans achieves state-of-the-art performance. Moreover, we conducted a series of synthetic experiments to identify the best approach, which was used in the eye-tracking experiments with a proxy radiologist. We observed that the performance of gaze-assisted MedSAM was superior to alternative state-of-the-art solutions.

Future work should expand the application of our gaze-assisted segmentation approach to a wider variety of medical datasets beyond abdominal imaging, including different anatomical regions and medical conditions. This will enhance the robustness and applicability of our method. We also intend to explore strategies for adapting our solution to 3D data, focusing on effective segmentation techniques for volumetric images.

7 Acknowledgement

This research has been financially supported by The Analytical Center for the Government of the Russian Federation (Agreement No. 70-2021-00143 01.11.2021, IGK 000000D730324P540002).

References

[1] Hands-free interactive image segmentation using eyegaze. In *Medical Imaging 2009: Computer-Aided Diagnosis*, volume 7260, pages 441–450. SPIE, 2009.

[2] Robert Alexander, Stephen Waite, Michael A. Bruno, Elizabeth A. Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. Mandating limits on workload, duty, and speed in radiology. *Radiology*, 304(2):274–282, August 2022.

[3] Pieter Blignaut and Tanya Beelders. The effect of fixational eye movements on fixation identification with a dispersion-based fixation detection algorithm. *Journal of eye movement research*, 2(5), 2008.

[4] Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M. Friedrich, and Felix Nensa. Explainable AI in medical imaging: An overview for clinical practitioners – beyond saliency-based XAI approaches. *European Journal of Radiology*, 162:110786, May 2023.

[5] Y.Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112 vol.1, 2001.

[6] R. J. M. Bruls and R. M. Kwee. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into Imaging*, 11(1), November 2020.

[7] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis*, 71:102062, 2021.

[8] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine*, 5(1), October 2022.

[9] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021.

[10] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[11] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1), February 2021.

[12] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[13] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018.

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

[15] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T. Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A. Krupinski, and Mehdi Moradi. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data*, 8(1), March 2021.

[16] Naji Khosravan, Haydar Celik, Baris Turkbey, Ruida Cheng, Evan McCreedy, Matthew McAuliffe, Sandra Bednarova, Elizabeth Jones, Xinjian Chen, Peter Choyke, et al. Gaze2segment: a pilot study for integrating eye-tracking technology into medical image segmentation. In *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging: MICCAI 2016 International Workshops, MCV and BAMBI, Athens, Greece, October 21, 2016, Revised Selected Papers 8*, pages 94–104. Springer, 2017.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[18] Christine Dan Lantsman, Yiftach Barash, Eyal Klang, Larisa Guranda, Eli Konen, and Noam Tau. Trend in radiologist workload compared to number of admissions in the emergency department. *European Journal of Radiology*, 149:110195, April 2022.

[19] Guangqi Li, Xin Wu, and Xuelei Ma. Artificial intelligence in radiotherapy. *Seminars in Cancer Biology*, 86:160–171, November 2022.

[20] Matthew D. Li and Brent P. Little. Appropriate reliance on artificial intelligence in radiology education. *Journal of the American College of Radiology*, June 2023.

[21] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *arXiv preprint arXiv:2111.02403*, 2021.

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

[23] Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 2023.

[24] Anika L. McGrath, Katerina Dodelzon, Omer A. Awan, Nicholas Said, and Puneet Bhargava. Optimizing radiologist productivity and efficiency: Work smarter, not harder. *European Journal of Radiology*, 155:110131, October 2022.

[25] Ilya Pershin, Maksim Kholiavchenko, Bulat Maksudov, Tamerlan Mustafaev, Dilyara Ibragimova, and Bulat Ibragimov. Artificial intelligence for the analysis of workload-related changes in radiologists' gaze patterns. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4541–4550, September 2022.

[26] Ilya Pershin, Tamerlan Mustafaev, Dilyara Ibragimova, and Bulat Ibragimov. Changes in radiologists' gaze patterns against lung x-rays with different abnormalities: a randomized experiment. *Journal of Digital Imaging*, 36(3):767–775, January 2023.

[27] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, May 2020.

[28] Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics*, 35(1):3–14, 2010.

[29] Mohammed Tahri Sqalli, Begali Aslonov, Mukhammadjon Gafurov, Nurmukhammad Mukhammadiev, and Yahya Sqalli Houssaini. Eye tracking technology in medical practice: a perspective on its diverse applications. *Frontiers in Medical Technology*, 5:1253001, 2023.

[30] Sian Taylor-Phillips and Chris Stinton. Fatigue in radiology: a fertile area for future research. *The British Journal of Radiology*, 92(1099):20190043, July 2019.

[31] Bin Wang, Armstrong Aboah, Zheyuan Zhang, Hongyi Pan, and Ulas Bagci. Gazesam: Interactive image segmentation with eye gaze and segment anything model. In *NeuRIPS 2023 Workshop on Gaze Meets ML*, 2023.

[32] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 41(7):1688–1698, July 2022.

[33] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond, 2023.

[34] Lian Zhang, Zhengliang Liu, Lu Zhang, Zihao Wu, Xiaowei Yu, Jason Holmes, Hongying Feng, Haixing Dai, Xiang Li, Quanzheng Li, et al. Segment anything model (sam) for radiation oncology. *arXiv preprint arXiv:2306.11730*, 2023.

[35] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.

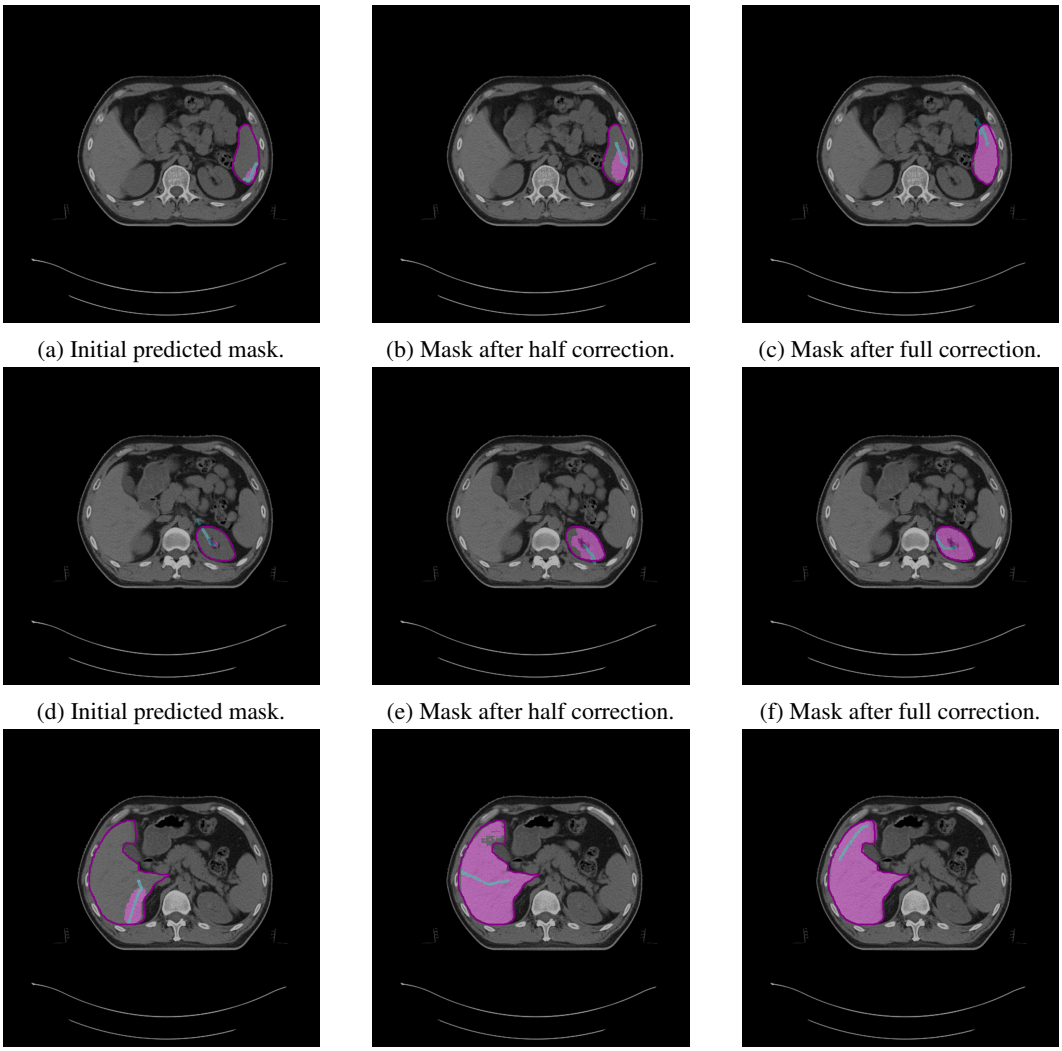
A Appendix / supplemental material

Table 4: Comparison of segmentation performance (DSC) in synthetic experiments with accumulation-based gaze data selection strategy with the distribution of points within the ground truth mask (prop gt), in the region of differences between the predicted and ground truth masks (prop mask diff) and outside both regions (prop out).

prop gt	prop out	prop mask diff	DSC
0%	100%	0%	0.07115
10%	90%		0.20640
20%	80%		0.34013
30%	70%		0.52645
40%	60%		0.74141
50%	50%		0.82396
60%	40%		0.85157
70%	30%		0.87200
80%	20%		0.88210
90%	10%		0.89290
95%	5%		0.89940
100%	**0%**		**0.90010**
20%	10%	70%	0.82960
5%	5%	90%	0.80380
40%	10%	50%	0.85080
30%	10%	60%	0.84320
60%	10%	30%	0.86662
70%	10%	20%	0.87379
80%	5%	15%	0.87848
90%	5%	5%	0.87930

Table 5: Comparison of segmentation performance (DSC) in synthetic experiments with gaze data selection strategy based on sequential accumulation of gaze data in the prompt at each iteration with the distribution of points in the gt mask (prop gt), in the region of differences between the predicted and reference masks (prop mask diff) and outside both regions (prop out). For each case, a fixed number of points (num points) was selected at each iteration.

prop gt	prop out	prop mask diff	num points	dice
20%	10%	70%	1	0.84142
5%	5%	90%		0.83411
40%	10%	50%		0.82446
30%	10%	60%		0.80465
60%	10%	30%		0.75430
70%	10%	20%		0.77780
80%	5%	15%		0.84267
90%	5%	5%		0.82022
20%	**10%**	**70%**	**2**	**0.88769**
5%	5%	90%		0.77403
40%	10%	50%		0.75456
30%	10%	60%		0.82651
60%	10%	30%		0.86370
70%	10%	20%		0.81802
80%	5%	15%		0.83697
90%	5%	5%		0.79542
20%	10%	70%	4	0.67183
5%	5%	90%		0.70947
40%	10%	50%		0.80588
30%	10%	60%		0.80906
60%	10%	30%		0.88311
70%	10%	20%		0.86640
80%	5%	15%		0.85710
90%	5%	5%		0.87252
20%	10%	70%	5	0.85304
5%	5%	90%		0.87967
40%	10%	50%		0.87501
30%	10%	60%		0.86181
60%	10%	30%		0.84695
70%	10%	20%		0.83883
80%	5%	15%		0.88035
90%	5%	5%		0.86934



(a) Initial predicted mask. (b) Mask after half correction. (c) Mask after full correction.

(d) Initial predicted mask. (e) Mask after half correction. (f) Mask after full correction.

(g) Initial predicted mask. (h) Mask after half correction. (i) Mask after full correction.

Figure 4: Steps to correct segmentation masks for various abdominal organs, such as the spleen, left kidney, and liver, on different CT slices. Each subfigure shows the outline of reference segmentation contours, the predicted segmentation mask, and gaze points (blue) used for predictions based on gaze.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract describes the results and experiments, which are presented in Sections 4 and 5. Description of methods in Section 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5.3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: -

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 details the methodology and Sections 4 and 5 describes the experiments. Experiments were conducted with human, so results may vary due to natural factors.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Provides code for reproducing experiments locally with instructions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental setting/details are stated in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have provided metric values with standard deviation in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.3 describes experiments compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research is conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The proposed approach will optimize the process of planning radiation therapy. Consequently, this will improve the quality of healthcare. In addition, the proposed approach can be used to optimize the process of annotating datasets. This will save money on the development of solutions based on artificial intelligence.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All resources used for experiments are open source and/or have proper attribution in their links.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Attached code includes implementation of the proposed approach and has comprehensive instructions on running experiments.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: -

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: -

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.