YOU POINT, I LEARN: ONLINE ADAPTATION OF INTERACTIVE SEGMENTATION MODELS FOR HANDLING DISTRIBUTION SHIFTS IN MEDICAL IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive segmentation uses real-time user inputs, such as mouse clicks, to iteratively refine model predictions. Although not originally designed to address distribution shifts, this paradigm naturally lends itself to such challenges. In medical imaging, where distribution shifts are common, interactive methods can use user inputs to guide models towards improved predictions. Moreover, once a model is deployed, user corrections can be used to adapt the network parameters to the new data distribution, mitigating distribution shift. Based on these insights, we aim to develop a practical, effective method for improving the adaptive capabilities of interactive segmentation models to new data distributions in medical imaging. Firstly, we found that strengthening the model's responsiveness to clicks is important for the initial training process. Moreover, we show that by treating the post-interaction user-refined model output as pseudo-ground-truth, we can design a lean, practical online adaptation method that enables a model to learn effectively across sequential test images. The framework includes two components: (i) a Post-Interaction adaptation process, updating the model after the user has completed interactive refinement of an image, and (ii) a Mid-Interaction adaptation process, updating incrementally after each click. Both processes include a Click-Centered Gaussian loss that strengthens the model's reaction to clicks and enhances focus on user-guided, clinically relevant regions. Experiments on 5 fundus and 4 brain-MRI databases show that our approach consistently outperforms existing methods under diverse distribution shifts, including unseen imaging modalities and pathologies. Code and pretrained models will be released upon publication.

1 INTRODUCTION

Medical image segmentation facilitates disease analysis, diagnosis, and treatment. Deep-learning methods have driven notable advances in automated medical image segmentation (Azad et al., 2024). However, a major challenge is that the training-data distribution often differs from the test-data distribution—for example, images may be acquired on different scanners—severely hindering model performance. Although models lack knowledge about unseen test data distributions, human users (such as clinicians) are often still able to segment images in the target distribution with reasonable accuracy. Hence, their knowledge can be leveraged to guide models. Can we design an AI framework that enables models to be guided by human users in an easy, immediate, and continuous manner, so that they can effectively adapt to distribution shifts? Although not originally developed for solving data distribution shift problems, a class of deep learning models known as interactive segmentation models is well suited to this challenge.

Interactive segmentation models allow users to provide prompts, such as clicks, scribbles, or bounding boxes, which inform the model's prediction. A common strategy is to encode user prompts as additional input channels in convolutional networks, as seen in models like DeepIGeoS (Wang et al., 2018) and Interactive FCNN (Sakinis et al., 2019). More recent approaches, such as SAM (Kirillov et al., 2023), MedSAM (Ma et al., 2024), and Med-SA (Wu et al., 2023b), instead employ Transformers to encode user prompts. Both approaches have demonstrated strong performance on natural and medical images, highlighting the usefulness of incorporating user guidance. However, they do not include mechanisms for adapting model parameters from user corrections.

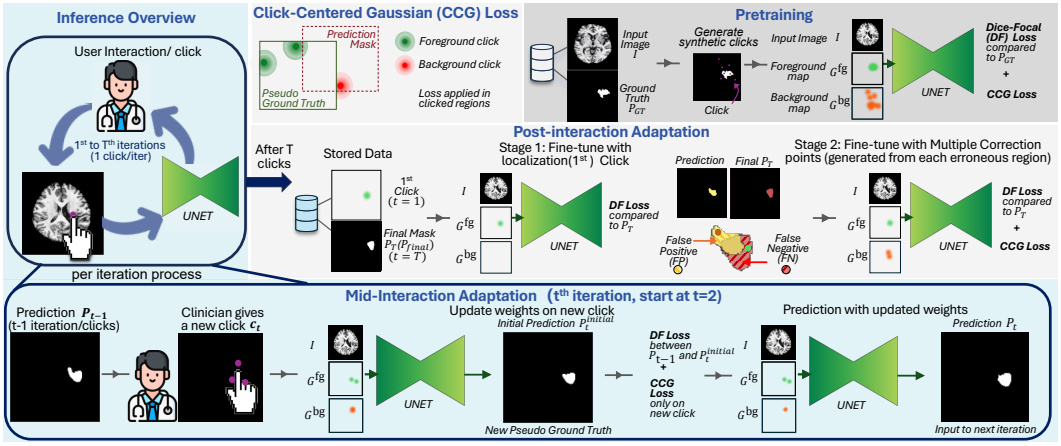

Figure 1: **Overview of the proposed method.** During **Pretraining**, the model is trained with simulated clicks provided as additional input channels to the image. During **Inference** and adaptation, images arrive sequentially. For each image, the user iteratively provides new clicks until the final prediction P_{final} (P_T) is obtained. After each new click c_t is provided, the model is updated using P_t^{initial} as pseudo ground truth (**Mid-Interaction Adaptation**). Once the final segmentation P_T is obtained, the model is first fine-tuned on the localization click (Stage 1) and then fine-tuned on multiple correction points generated from P_T and the prediction of Stage 1 (Stage 2) (**Post-Interaction Adaptation**). A Gaussian-weighted and class-limited **CCG loss** is applied in all three processes.

Our work explores how to leverage user guidance in interactive segmentation to improve performance under distribution shift most effectively. We identify that this requires complementary learning mechanisms for the pre-deployment training and post-deployment adaptation. For pre-deployment stage , we find that adding an optimization objective that enforces model predictions to align with user feedback in areas around given clicks improves performance under distribution shift.

We couple this with post-deployment learning mechanisms. **Post-deployment adaptation** methods for interactive segmentation **optimize a model for the specific data distribution encountered after deployment** using information from user prompts. They use **Online Learning** (Hoi et al., 2021) to update model parameters after each test image is processed sequentially.[1]. Prior work on online adaptation for interactive medical image segmentation is limited. An early related method is IA+SA (Kontogianni et al., 2020), originally developed for natural images, which combines independent image-level adaptation (IA) and image-sequence adaptation (SA). Another recent related method, TSCA (Atanyan et al., 2024), achieves further improvements in performance after online adaptation. These methods leverage user corrections through sparse cross-entropy or focal loss, applied only to the pixels clicked by the user. This focuses optimization narrowly on a small number of pixels while ignoring surrounding areas. Moreover, additional regularizers are often required to prevent overfitting to the few labeled pixels, increasing model complexity and the number of hyper-parameters that must be tuned.

Our work is based on the insight that, in a real-world interactive segmentation workflow where the user provides clicks to correct a model output, the segmentation predicted at the end of interactions should have sufficient quality to serve as pseudo ground-truth. We propose a Post-Interaction adaptation method based on using that final prediction as optimization target and show it results in effective adaptation without requiring complex regularizers. Furthermore, we generate *artificial* correction clicks from the pseudo-ground-truth mask and use them with the Click-Centered Gaussian (CCG) loss that we introduce to *strengthen the model's response around clicks* under the new data distribution. This improves performance in all tested distribution-shift scenarios.

We further combine this with Mid-Interaction Adaptation, which adapts the model weights after each user click. We again rely on the segmentation mask predicted by the model and optimize the CCG loss, which emphasizes the region around the click. Unlike previous works that focus only

[1] Adaptation in this context targets only the distribution seen post-deployment, thus does not need to handle Catastrophic Forgetting of past distributions as Continual Learning (Wang et al., 2024)

on the clicked pixels for such adaptation, our method leverages the whole corrected segmentation mask together with the CCG loss, thereby optimizing over the greater surrounding region, improving adaptation performance.

We term the proposed method **OAIMS** (Online Adaptation for Interactive Medical-image Segmentation). Experiments under distribution shift across 5 fundus and 4 brain-MRI databases demonstrate that by using solely the proposed Post-Interaction method already results in adaptation performance that compares favorably to SOTA adaptation methods. When this is combined with Mid-Interaction adaptation, OAIMS consistently outperforms all previous methods, especially on brain MRI where Dice score improvements exceed 10%. Ablation studies show that the proposed CCG loss is consistently useful when employed in all 3 learning processes (pretraining, mid- and post- adaptation). Further analysis also shows strong robustness to settings that may cause overfitting to other methods.

2 METHODS

2.1 OVERVIEW: INTERACTIVE SEGMENTATION FRAMEWORK

We here provide an overview of the whole process, which is shown in Fig. 1. For simplicity, we describe it for binary segmentation, but it also applies for multiple classes, as shown in Experiments.

We define the interactive model as $f(I, C; \theta)$, where I is the input image, C is the set of user clicks, and θ are model parameters. A click is labeled either as foreground or background class. A foreground click indicates that the specific pixel belongs to the target object, background click indicates that it does not. We train the model on a source database with simulated clicks C. During inference, the model receives a sequence of images $\{I_1, I_2, \ldots, I_N\}$ from another database. For a single image I_n separately, the user (or simulated user) first provides a localization click c_1 to trigger the interactive process. The *localization click* used to start interaction is simply a foreground click placed anywhere inside the target foreground object. The model predicts initial segmentation $P_1^n = f(I_n, c_1; \theta)$. Afterwards, multiple iterations of interactions occur. At iteration t the user places a new click c_t in a region where prediction P_{t-1}^n is wrong. The click set is updated $C_t = C_{t-1} \cup \{c_t\}$, where $C_1 = \{c_1\}$. The model then predicts $P_t^n = f(I_n, C_t; \theta)$. Next interaction $t + 1$ then occurs, and so forth. After T interactions we obtain the final prediction P_T^n, which we call P_{n_final}. While here T is given a set value for simplicity, in a real-world setting T would be as much as user requires to be satisfied with segmentation output. The whole process is then repeated for the next image I_{n+1} in the sequence. For notational simplicity, we omit the image index n in most formulas below.

During inference, we perform two types of online adaptation. The **Post-Interaction adaptation** is a two-stage method that updates the model after the iterative, interactive corrections for a single image have finished and the model has produced final segmentation P_{final}. This improves performance for subsequent images. **Mid-Interaction adaptation** happens after each interaction. It takes place before the P_{final} is obtained. This strategy benefits both the current and subsequent images.

2.2 PRETRAINING THE INTERACTIVE MODEL

Our base interactive model is a U-Net (Ronneberger et al., 2015) modified to accept both the image and click prompts as input. We use the same strategy as ICNN (Sakinis et al., 2019), where we set 2 guidance maps that encode foreground and background clicks, respectively, each having the same spatial dimensions as I. The raw guidance maps are zero everywhere except at clicked pixels; we then apply a Gaussian smoothing kernel and normalize each map to $[0, 1]$. These maps are concatenated with the image along the channel dimension. The concatenated tensor (image + foreground map + background map) is input to the model. We train the base model using simulated clicks and a compound loss: **Dice–Focal** (Eq. equation 4) and **CCG Loss** (Eq. equation 3), which strengthens the model's response to user clicks. See Appendix A.2 for details regarding click simulation.

2.3 CLICK-CENTERED GAUSSIAN (CCG) LOSS

An interactive model should react to user clicks and update the surrounding region accordingly. We propose a Click-Centered Gaussian Loss to strengthen the model's reaction to clicks by penalizing wrong predictions near each click, weighted by a Gaussian kernel. The penalty is applied only to

pixels that *should* share the same class as the click (e.g. for a foreground click, the loss only applies to surrounding pixels that are foreground in the ground truth mask). This loss is employed in all three stages, pre-training, Post-Interaction adaptation, and Mid-Interaction adaptation.

Let c denote a user click at pixel (i', j') with class label $y_{i',j'} \in \{0, 1\}$. For any pixel (i, j) we define the Gaussian weight and an indicator

$$G_c(i,j) = \begin{cases} \exp\!\left(-\frac{(i-i')^2+(j-j')^2}{2\sigma^2}\right), & |i-i'| \le 3\sigma \text{ and } |j-j'| \le 3\sigma \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

$$I_c(i,j) = \begin{cases} 1, & P(i,j) = y_{i',j'}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

P denotes the ground-truth mask (used for pretraining) or pseudo ground-truth mask (used for adaptation), and $P(i, j)$ is its pixel value at coordinates (i, j).

Given the current prediction $\hat{P}$, the **CCG Loss** is

$$\mathcal{L}_{\text{CCG}} = \frac{\sum_{c \in C} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} G_c(i,j)\, I_c(i,j)\, \text{CE}\big(\hat{P}(i,j),\, P(i,j)\big)}{|C|HW} \tag{3}$$

where C is the set of clicks for the current sample, $H \times W$ is the image size, and $\text{CE}(\cdot, \cdot)$ denotes cross-entropy loss.

Why not apply the loss to all surrounding pixels? Each click only serves to change the surrounding pixels to a specific target class. The click does not provide information for clusters of pixels belonging to a different class. Applying extra penalties to pixels annotated as another class in the ground truth may cause the model to overfit to specific regions or images, ultimately degrading overall performance when facing distribution shifts or performing online learning.

2.4 Online Adaptation

Post-Interaction Adaptation: This process updates the model after the user has completed all T clicks for an image. The key assumption is that the final segmentation P_{final} after the user finishes their interactions is of "good enough" quality to serve as a pseudo ground-truth mask for updating the model.In real-world practice this is rather easy to ensure, if we only adapt based on segmentations for which the user has confirmed that the interactions resulted in satisfying outputs. Even if the mask is imperfect, it still provides new information from users to update the model's knowledge.

The user starts the interaction with a localization click. We therefore naturally split the post-interaction updates into: (i) Fine-tune with the initial localization click as input; (ii) Fine-tune with the correction clicks as input.

Stage 1 – Fine-tune with Localization Click as input: First, we fine-tune the model with one localization click, obtained from the previous inference step. With this click, we obtain $P_1 = f(I, c_1, \theta)$ and use $P_T = f(I, C_T, \theta)$ (the pseudo ground-truth mask) to update the model. We apply **Dice–Focal (DF) loss** (Milletari et al., 2016; Lin et al., 2017) between P_1 and the final mask P_{final}, where

$$\mathcal{L}_{DF} = (1-\alpha)\,\mathcal{L}_D + \alpha\,\mathcal{L}_F. \tag{4}$$

$\mathcal{L}_D$, $\mathcal{L}_F$ and α are the Dice loss, Focal loss, and a weighting hyper-parameter respectively. Only one Gradient Descent update is performed for each image using Eq. 4.

Stage 2 – Fine-tune with Multiple Correction Points: To improve ability of the model to leverage correction-clicks, we input artificial correction-clicks, obtain the output and update the model using P_{final} as target. We cannot reuse the user's original correction clicks, as they were already employed to produce P_{final} –they would result in identical output leading to trivial updates. Instead, we compare the Stage 1 output (P_1) with P_{final}, locate false-positive and false-negative regions, and generate one artificial click in each erroneous connected component (up to T clicks), without extra human input. The newly generated clicks are fed to the model, yielding new prediction $\hat{P}$. We then apply the proposed CCG loss, supplemented by the Dice–Focal loss for further guidance. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{DF} + \beta\,\mathcal{L}_{CCG}. \tag{5}$$

The CCG loss ensures the model **reacts** to each click in its surrounding region during adaptation. To our knowledge, no previous work addresses this point explicitly.

Mid-Interaction Adaptation: Besides Post-Interaction adaptation, we can also update the model *after each user click*. When the model is updated after every click, the new parameters directly influence the next prediction, so the update not only improves performance on the following images but also refines the current result. This in turn helps achieve a high quality P_{final} after all T interactions, assumption of Post-Interaction adaptation, thus complementing and enhancing it indirectly.

We keep the idea of using the model's own corrected output as pseudo ground truth. Let $P_{t-1} = f(I, C_{t-1}, \theta)$ be the prediction after $t-1$ clicks. When click c_t is given, we obtain $P_t^{\text{initial}} = f(I, C_t, \theta)$, which is used as the pseudo ground truth. We optimize Dice–Focal plus CCG loss (Eq. 5) between P_{t-1} and P_t^{initial}. Here the CCG loss only applies to the new click c_t. After the model is updated, it processes C_t again and produces P_t, which is shown to the user (or simulator) to get the next click. The process continues until the final prediction P_{final} is obtained. After this, we begin the Post-Interaction adaptation on that image.

The pseudo ground truth in Mid-Interaction adaptation is not perfect, so the CCG loss is very important. It helps the model to concentrate learning on regions close to the clicks, which are the most valuable and trustworthy areas. The CCG loss is not intended to strengthen the model's reaction here, because the c_t is not used for obtaining P_{t-1}.

3 Experiments

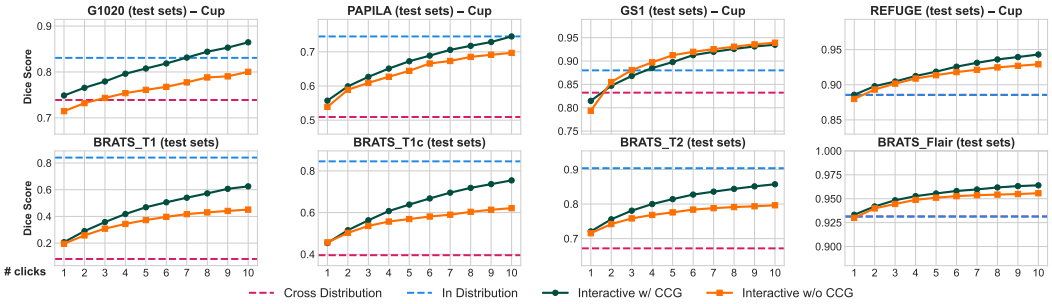

Figure 2: Dice-score performance for automatic and interactive models. All networks, except the in-distribution baseline, are trained on REFUGE (fundus) or BRATS-FLAIR (brain MRI). The x-axis represents the number of clicks. Horizontal lines mark automatic performance of automatic models in cross-distribution and in-distribution settings. Curves show the interactive segmentation model with and without the CCG loss; clicks significantly improve performance in all test cases, with the CCG loss providing additional gains, especially for large distribution gaps (e.g., BRATS-T1).

Databases: We evaluate our method on two types of data. **Fundus imaging:** We use 5 public databases: REFUGE2 (Orlando et al., 2020; Fang et al., 2022), G1020 (Bajwa et al., 2020), GS1-Drishti (Sivaswamy et al., 2014), GAMMA (Wu et al., 2023a), and PAPILA (Kovalyk et al., 2022). They are 2D RGB images acquired at different clinics with different scanners, hence each represents a different distribution. We perform multi-class segmentation {0: Background, 1: Outer-ring, 2: Cup}. We compute evaluation metrics on Cup and Disc, where Disc is the union of Outer-ring and Cup, as common in literature (Orlando et al., 2020). Unless stated otherwise, we treat $REFUGE2$ as the *source* database on which we pretrain the interactive model. We treat other databases as different *target* distributions for adaptation and evaluation. **MRIs of Brain Lesions:** We use 4 databases. Each contains a different type of pathology and some contain multiple MRI modalities: BRATS2023 - Glioma, with Flair, T1, T1c, T2 modalities (Baid et al., 2021); ATLAS v2.0 - Stroke, with T1 modality (Liew et al., 2022); WMH - white matter hyperintensities, with Flair and T1 (Kuijf et al., 2022); TBI - Traumatic Brain Injuries, with Flair and T1. TBI is the only non-public data we use. Each database is acquired from a different clinic, with different scanner. Each database and modality can therefore be regarded as a different data distribution. Although the MRI scans are 3D images, we test the models using 2D slices, by selecting the slice with the largest lesion area

Table 1: Performance on fundus imaging (average Dice, %). Each image receives 10 clicks; the base interactive model is trained on REFUGE. Each cell reports Disc/Cup Dice at 1, 5, and 10 clicks. While ICNN* and Med-SA are non-adapting models, the other methods perform online adaptation. PI is our Post-Interaction adaptation strategy. PI+MI combines the Post-Interaction with the Mid-Interaction strategy.

	G1020			PAPILA			GS1			GAMMA		
No. Cl	**1**	**5**	**10**	**1**	**5**	**10**	**1**	**5**	**10**	**1**	**5**	**10**
ICNN*	89.4/77.4	93.1/83.1	95.1/88.0	88.3/60.3	92.6/70.2	94.6/77.1	96.6/82.4	97.2/90.4	97.7/94.1	94.4/83.3	96.3/89.8	97.2/93.3
Med-SA	72.2/56.1	75.3/58.6	75.6/58.6	52.5/34.5	56.1/36.7	56.7/36.7	88.6/61.5	89.6/67.0	89.6/67.6	89.8/76.2	90.0/77.5	90.0/77.5
IA+SA	90.5/77.8	94.3/84.5	96.1/90.3	89.4/61.3	93.3/70.5	95.3/77.6	96.8/83.4	97.4/91.5	98.0/94.8	94.7/84.0	96.3/90.2	97.4/93.9
TSCA	89.9/77.6	94.2/85.0	96.1/90.7	89.6/61.7	94.0/72.2	96.1/79.0	96.9/85.4	97.5/92.9	98.0/95.5	94.6/84.2	96.4/90.7	97.5/**94.1**
PI	93.5/81.9	95.9/88.5	97.0/92.2	94.1/73.0	96.1/80.0	97.0/85.8	**97.3/89.5**	97.7/94.1	98.2/95.7	**95.3/85.9**	96.5/**91.4**	97.5/**94.1**
PI+MI	**93.6/82.2**	**96.4/90.0**	**97.5/92.7**	**94.7/73.0**	**96.5/81.0**	**97.5/86.2**	**97.3**/89.3	**97.8/94.5**	**98.4/96.5**	95.1/85.8	**96.7**/91.3	**97.9**/93.9

Table 2: Performance (Dice%) on different MRI modalities.

	BRATS T1			BRATS T1C			BRATS T2		
No. Clicks	1	5	10	1	5	10	1	5	10
ICNN*	20.7	46.8	62.5	45.4	63.8	75.4	72.1	81.4	85.7
Med-SA	23.4	33.4	35.7	38.7	47.0	49.2	75.7	78.8	79.2
IA+SA	28.6	57.0	70.6	48.3	67.1	78.2	76.5	84.1	87.9
TSCA	34.4	60.9	74.0	50.1	70.2	79.6	77.7	85.8	89.0
PI	61.1	72.4	78.9	64.3	74.7	80.4	82.5	88.1	90.9
PI+MI	**71.2**	**83.4**	**88.0**	**70.4**	**82.9**	**87.5**	**84.9**	**90.6**	**93.0**

per case. For multi-class databases (BRATS, TBI), we merge all lesion classes into one label, and perform binary classification (healthy tissue VS lesion). We split BRATS into 1002 training cases and 249 test cases. Unless stated otherwise, we pre-train on the train split of BRATS using the Flair modality. We then use the other modalities of BRATS's test split, and all other databases, as the target distributions for adaptation and evaluation. All interactive models are trained with up to 10 simulated clicks per image, unless stated otherwise.

3.1 Interactive Segmentation Under Distribution Shift

We start with experiments that test the performance of the base interactive model under data-distribution shift, without any adaptation. We also assess if the proposed CCG loss helps an interactive segmenter to handle distribution shifts. We consider 2 types of distribution shifts. The first is across fundus databases. We consider REFUGE as the *source* distribution, and other fundus databases as *target* distributions. The second is across BRATS modalities. We consider BRATS FLAIR as the *source* distribution, and other BRATS modalities as *target* distributions.

Four models are compared for each of the 2 settings: (1) Automatic (non-interactive) U-Net (same backbone as the interactive model) trained on the source database, and applied to each of the target databases (cross-distribution); (2) Automatic model trained on the target database and applied to the target database (in-distribution). This is to quantify performance on target, without influence of distribution shift; (3) Interactive segmentation model trained on the source database without CCG loss, applied to each target database; (4) Similar to (3) but pretraining also uses CCG loss. Results are shown in Fig.2. We see a large difference between *in-distribution* and *cross-distribution* performance of automatic methods, across all settings, due to distribution shifts between *source* and *target* data. Nonetheless, ten clicks with the interactive segmentation model largely close the gap, confirming that interactive segmentation remains effective under strong shifts. Adding the CCG loss during pretraining yields improvements in most settings. This is because the CCG loss enforces the model to depend more on user input when given, signal unrelated to distribution shift, and less on the image signal where the shift manifests.

3.2 Online Adaptation

We then evaluate our online adaptation methods, including *Mid-Interaction* and *Post-Interaction* adaptation. For all following experiments, during online adaptation, a sequence of images is input to the model. Each image is corrected through T iterations (one click per iteration, $T = 10$ by default). For every image, the Dice score is calculated with the prediction mask at each iteration t *after mid-interaction adaptation is performed in that iteration*. We measure average Dice score

Table 3: Performance (Dice%) on different brain pathologies.

	Trained on BRATS (Flair)						Trained on BRATS (Flair, T1, T1c)					
	TBI Flair			WMH Flair			TBI T1			ATLAS T1		
No. Clicks	1	5	10	1	5	10	1	5	10	1	5	10
ICNN*	49.9	64.1	69.6	47.9	61.2	67.6	42.0	49.3	55.3	40.6	46.8	52.1
Med-SA	43.5	47.9	48.5	52.6	60.0	61.5	34.0	41.3	43.1	35.7	42.4	43.9
IA+SA	50.6	66.4	73.9	49.4	64.2	72.0	44.5	52.7	59.8	43.4	53.9	62.6
TSCA	52.7	66.1	73.7	52.8	66.7	72.7	44.4	55.8	63.9	43.4	55.7	64.0
PI	53.8	68.8	73.6	53.7	66.7	72.3	**47.7**	61.1	68.0	62.7	77.0	81.8
PI+MI	**55.2**	**69.9**	**76.3**	**59.0**	**73.0**	**78.9**	**47.7**	**67.0**	**74.8**	**66.4**	**82.2**	**86.0**

Table 4: Adapting with maximum 5 or 3 clicks per image.

	5 clicks			3 clicks			
	BRATS			TBI	WMH	TBI	ATLAS
Method	**T1**	**T1c**	**T2**	**Flair**	**Flair**	**T1**	**T1**
ICNN*	46.8	63.8	81.4	58.9	55.6	45.8	44.6
TSCA	62.9	69.0	86.0	61.9	59.5	52.9	51.4
PI	71.2	72.8	87.4	62.5	61.2	52.6	68.4
PI+MI	**80.4**	**80.5**	**90.3**	**65.3**	**64.3**	**54.6**	**73.3**

achieved for each image in the image sequence after 1, 5, and 10 interactions. The adaptation methods update the model after getting the segmentation result after each click and each image.

We set $\alpha = 0.7$ and $\beta = 200$, with $\sigma = 3$ in CCG loss, found adequate in preliminary experiments. The effect of different hyperparameter settings can be seen in Appendix A.3. We implement a base interactive model we denote it as ICNN*, using a U-NET with the interactive method proposed by ICNN (Sakinis et al., 2019), and trained with our CCG loss. We implement IA+SA (Kontogianni et al., 2020) and TSCA (Atanyan et al., 2024) using the same pretrained base interactive model (with CCG loss) as our method for fair comparison. In addition, we include a SAM-based interactive medical image segmentation model, the Medical SAM Adapter (Med-SA) (Wu et al., 2023b), which is fine-tuned on the source data and frozen during testing (target data).

Evaluation on Fundus data. We pretrain the models on REFUGE as the source distribution, for multi-class segmentation. We then adapt and evaluate using each of the 4 other fundus databases separately as *target* distributions. Tab. 1 shows the average Dice for both disc and cup. On G1020 and PAPILA, where the data-distribution shift is large, our fast *Post-Interaction* method outperforms previous methods—especially on cups (disc segmentation is nearly perfect for most models and thus hard to improve)—and all adaptation approaches surpass the frozen base model. On GS1 and GAMMA, where the shift is small, our *Post-Interaction* method remains comparable or better. Using only the *Post-Interaction* adaptation, which requires two back-propagations, already surpasses previous methods that need more than ten back-propagations. Adding the Mid-Interaction adaptation gives slightly better results in most cases. The improvement becomes much more significant when facing large data-distribution shifts in the brain-MRI databases.

Evaluation on MRI Modalities: We here adapt our model to scenarios with larger distribution shifts – between different MRI modalities. The model is initially trained using the FLAIR scans of the training split. It is then adapted and evaluated on T1, T1c, and T2 scans of the test split (separate experiment per modality). As shown in Tab. 2, all online-adaptation methods outperform the base interactive model, ICNN*. Largest improvements shown in T1. Among online adaptation methods, our approach surpasses TSCA and IA+SA even with only Post-Interaction adaptation, especially when few clicks are given. Including Mid-Interaction adaptation yields even greater gains.

Adapting to Different Brain Pathologies: In addition, we test our model across different brain pathologies. We pretrain 2 models, one on BRATS-Flair, and one on a combination of Flair/T1/T1c. We then adapt and evaluate the first on TBI-Flair and WMH-Flair, and the second on TBI-T1 and ATLAS-T1. Results are shown in Tab. 3. Even in these challenging settings, online-adaptation methods significantly boost performance, with our approach outperforming previous methods on all tasks. For TBI and WMH on FLAIR, our Post-Interaction method achieves results comparable to TSCA after 10 clicks but attains higher dice scores with fewer clicks. After adding Mid-Interaction adaptation, our method achieves significantly better results.For TBI-T1 and ATLAS, Post-Interaction alone significantly outperforms previous methods, and Mid-Interaction further improves performance. Although the pseudo ground truth in early iterations is suboptimal, as shown in the table (low Dice score for 1 click), PI+MI can still learn from it and achieve higher scores.

We also observe that in all three tables, TSCA performs better than IA+SA, consistent with the previous studies (Atanyan et al., 2024). Thus, we compare only with TSCA in subsequent experiments for simplicity. Furthermore, we observe that in most cases, Med-SA performs significantly worse than the base interactive model, ICNN, across all three tables, especially after 10 points. Therefore, we do not employ the computationally expensive SAM-based model further.

Adapting with fewer allowed corrections: All previous experiments used $T = 10$ maximum clicks for correction of each image. However, a method should ideally also perform well with fewer maximum performed corrective interactions. Here, we test our online-adaptation methods on brain MRI using maximum $T = 3$ or 5 clicks for interactive correction of each image. This also assesses the capability of our method to adapt using a less optimal pseudo ground-truth. Tab. 4 shows results using 5 or 3 clicks max per image, under the same experiment settings as Tab. 2 and Tab. 3. Even with fewer clicks, online-adaptation methods perform significantly better than the frozen model ICNN*. Our Post-Interaction (PI) adaptation continues to outperform previous methods in most cases. The addition of Mid-Interaction (PI+MI) further improves results. Although the model output after 3/5 clicks may be suboptimal, learning from it as pseudo ground-truth remains effective.

Table 5: Ablation study by including different terms in PI and MI, after 1, 5, 10 clicks. CCGL_{MI} and DFL_{MI} represent the Mid-Interaction adaptation. CCGL_{PI} and DFL_{PI} represent the stage 2 of the Post-Interaction adaptation. S1_{PI} represent the stage 1 of the Post-Interaction adaptation.

Loss terms					G1020 (Cup)			ATLAS			BRATS T1			BRATS T2		
DFL_{MI}	CCGL_{MI}	DFL_{PI}	CCGL_{PI}	S1_{PI}	1	5	10	1	5	10	1	5	10	1	5	10
✓	✓	✓	✓	✓	82.2	90.0	92.7	66.4	82.2	86.0	71.2	83.4	88.0	84.9	90.6	93.0
–	✓	✓	✓	✓	82.1	88.9	93.0	65.9	82.0	85.8	69.4	81.9	86.8	84.3	90.4	92.8
✓	–	✓	✓	✓	81.5	87.0	90.0	65.2	80.1	83.8	42.3	46.6	48.9	84.6	90.4	92.6
–	–	✓	✓	✓	81.9	88.5	92.2	62.7	77.0	81.8	61.1	72.4	78.9	82.5	88.1	90.9
–	–	✓	–	✓	82.2	87.6	90.6	58.4	66.2	69.2	60.6	71.7	76.8	82.6	88.1	90.5
–	–	–	✓	✓	81.7	86.8	89.6	60.7	74.7	79.8	55.7	66.3	72.1	81.7	87.3	89.9
–	–	–	–	✓	81.6	87.7	91.4	59.0	73.0	78.7	48.6	61.3	67.9	80.9	86.4	89.4

Ablation Study: To evaluate the benefit of each component of our method, we conduct an ablation study on each term of our online adaptation method. The results are shown in Tab. 5. Dice scores are reported on four target databases: G1020 (cup), ATLAS, BRATS-T1, and BRATS-T2. The source databases are as follows: REFUGE2 for G1020 (cup), a combination of BRATS Flair/T1/T1c for ATLAS, and BRATS Flair for both BRATS-T1 and BRATS-T2. The source-target pairs are consistent with previous experiments. The ablation terms are divided into two groups: PI (Post-Interaction adaptation) and MI (Mid-Interaction adaptation). $\mathbf{S1_{PI}}$ is the first stage of the Post-Interaction adaptation approach. $\mathbf{DFL_{PI}}$ and $\mathbf{CCGL_{PI}}$ are the second stage of the Post-Interaction adaptation processes with the Dice-Focal loss or Click-Centered Gaussian loss. $\mathbf{DFL_{MI}}$ and $\mathbf{CCGL_{MI}}$ are the Mid-Interaction adaptation processes with the Dice-Focal loss or Click-Centered Gaussian loss. Overall, the ablation study confirms the contribution of each component and stage in both the Mid-Interaction and Post-Interaction approaches. The two loss terms should be used together in each process.

We have seen that MI adds benefits on top of PI. But does the opposite also hold? We evaluate whether adapting with PI offers benefits when MI is already performed. With a budget of five clicks per image, Post-Interaction adaptation improves performance in nearly every scenario as show in Tab. 6. When the budget rises to ten clicks, Post-Interaction adaptation continues to provide substantial gains in the early iterations, but by the final click, its advantage narrows: WMH still benefits, while others do not. Exact numbers are given in the Appendix A.4. With more clicks, the model leans more heavily on MI, which may partially cover the updates supplied by Post-Interaction adaptation. Even though the influence of PI may diminish with extensive interaction provided, it helps users reach satisfactory results with *fewer* clicks, which is important for interactive workflows. We therefore recommend deploying both mechanisms in most situations.

Finally, we investigated variants of the CCG loss, as shown in Tab. 7. Removing either the Gaussian kernel or the class-limited mechanism lowers performance for both PI and PI+MI in most cases.

3.3 Robustness and overfitting

In this section, we conduct additional experiments to assess the robustness of our method and potential overfitting. When plenty of clicks are provided for an image, the model may overfit to that image. To explore this, we consider an extreme case where each image receives 50 clicks (TSCA

Table 6: Ablation study for PI under a 5-click budget. (Average Dice% over 3 runs (different seeds))

	BRATS			WMH	TBI
PI+MI	**80.4**	**80.5**	**90.3**	**70.7**	68.9
MI	78.4	79.8	89.8	68.0	**69.0**

Table 7: Ablation study on the design of CCG loss. Performance shown as Dice% .

	all		no_class		no_gaussian	
	PI+MI	PI	PI+MI	PI	PI+MI	PI
BRATS (T2)	**93.0**	**90.9**	92.1	87.7	90.6	90.1
WMH	**78.9**	72.3	77.6	**73.1**	72.2	69.8
ATLAS	**86.0**	**81.8**	80.7	75.7	79.8	80.5

(50), OAIMS (50)). The result is shown in Tab. 8. In this scenario, TSCA (50) exhibits lower performance at the early clicks (e.g., click 1, 3) compared to TSCA(10), indicating potential overfitting to previously seen images. In contrast, our method performs significantly better with 50 clicks compared to 10, and does not exhibit signs of overfitting.

We also examine a challenging scenario, where images from different databases—BraTS T1, BraTS T2, and WMH FLAIR—are randomly shuffled together, with each database contributing 25 images. The result is shown in Tab. 9. Despite substantial domain differences, our method continues to outperform both the ICNN* and TSCA. Notably, while TSCA's performance approaches that of ICNN*, our method maintains a clear advantage. Removing Post-Interaction adaptation leads to performance drops. Hence our adaptation approach allows clinicians to use a single model that adapts to multiple diseases simultaneously, eliminating the need to manage multiple models.

Table 8: Performance on BRATS T1 with (10) or (50) clicks in total per image (in brackets). Dice shown at 1, 3, 10, 20, 50 clicks. Overfitting past images lowers Dice on next image with few clicks (1-3) using TSCA but not our method.

No. Clicks	1	3	10	20	50
ICNN*	22.1	35.8	62.9	78.7	87.1
TSCA(50)	28.9	49.6	75.3	88.4	92.9
TSCA(10)	34.4	51.4	74.0	N/A	N/A
OAIMS (50)	73.9	81.8	89.8	94.3	95.9
OAIMS (10)	61.1	68.9	78.9	N/A	N/A

Table 9: Adapting to a database composed of images from BRATS T1,T2, and WMH FLAIR.

No. Clicks	1	5	10
ISFCNN	50.1	65.6	74.3
TSCA	52.0	67.1	75.3
OAIMS(MI)	52.1	72.1	80.6
OAIMS(PI+MI)	55.5	73.7	81.6

4 Conclusion

This study investigates how to train and adapt an interactive segmentation model for medical imaging to better handle data distribution shifts. We proposed an online adaptation framework that integrates both Post-Interaction and Mid-Interaction approaches, enabling the model to continuously adapt to new data distributions. A Click-Centered Gaussian loss is proposed, which enhances the model's responsiveness to user inputs. We demonstrate the effectiveness of our method through extensive experiments with diverse distribution shifts. The promising performance underscores the transformative potential of adaptive interactive segmentation in advancing both clinical practice and research applications.

REFERENCES

Barsegh Atanyan, Levon Khachatryan, Shant Navasardyan, Yunchao Wei, and Humphrey Shi. Continuous adaptation for interactive segmentation using teacher-student architecture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 789–799, 2024.

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, et al. Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. *arXiv preprint arXiv:2202.08994*, 2022.

Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

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*, pp. 4015–4026, 2023.

Theodora Kontogianni, Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Continuous adaptation for interactive object segmentation by learning from corrections. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 579–596. Springer, 2020.

Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 9(1):291, 2022.

Hugo Kuijf, Matthijs Biesbroek, Jeroen de Bresser, Rutger Heinen, Christopher Chen, Wiesje van der Flier, Barkhof, Max Viergever, and Geert Jan Biessels. Data of the White Matter Hyperintensity (WMH) Segmentation Challenge, 2022. URL `https://doi.org/10.34894/AECRSD`.

Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1):320, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

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

Fausto Milletari et al. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020.

Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

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

Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pp. 53–56. IEEE, 2014.

Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, Yue Huang, Qinji Yu, Sifan Song, Xinxing Xu, et al. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90:102938, 2023a.

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. *arXiv preprint arXiv:2304.12620*, 2023b.

A Appendix

A.1 Visualization Results

Fig. 3 presents the visualization results demonstrating the adaptation performance on the BRATS dataset. The segmentation map is overlaid in red on the original image. Our OAIMS (PI+MI) method produces segmentations that are closest to the ground truth (GT), with more accurate boundaries compared to other methods.

Fig. 4 illustrates the databases used in our experiments. This visualization helps to better understand the distribution shifts across different databases and modalities.

Fig. 5 illustrates how the predicted segmentation evolves across interaction clicks.

A.2 Details of the Simulation Process

To train and evaluate the interactive model, we simulate user interactions with an automatic point-generation procedure that places clicks in incorrectly segmented regions.

During training, we first generate a random click inside the target foreground object as a localization click. Based on the resulting segmentation mask and the ground truth, we identify incorrectly segmented regions with connected components. Each erroneous component is ranked by size, and a random point is generated within each. We then select the first M points from this queue, where M is the desired number of clicks, and feed them into the model simultaneously. In our training, M is randomly sampled for each iteration from a uniform distribution in the range $[1, 10]$.

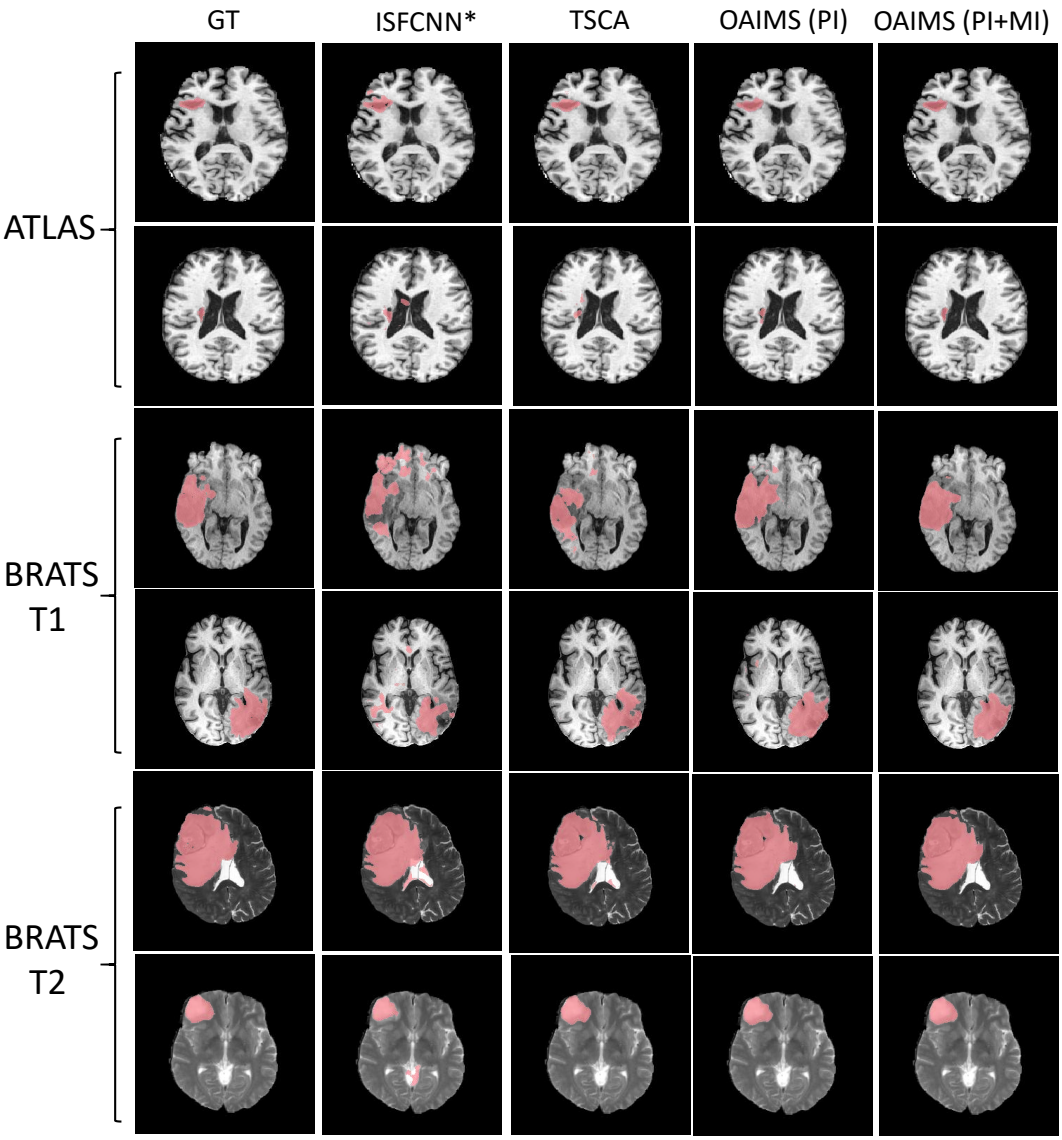

Figure 3: Visualizations on BRATS Databases demonstrating adaptation to different modalities and pathologies (Trained on BRATS Flair for BRATS T1 and BRATS T2, Trained on BRATS Flair/T1/T1c for ATLAS

At inference time, user clicks are simulated iteratively. First, one random click is placed inside the target foreground object. Then, based on the predicted segmentation and the ground truth, a correction click is placed in the largest erroneous component (including both false positives and false negatives). A new segmentation is generated using all previous clicks, and the process repeats until a total of T clicks is reached.

For the simulation process in the Post-Interaction stage, the procedure is similar to training. However, instead of using the real ground truth, this step relies on a pseudo ground-truth mask.

A.3 Effect of Different Hyperparameter Settings

In this section, we evaluate the effect of three hyperparameters, α, β, and σ, on the performance of our method. The results are presented in Table 10.

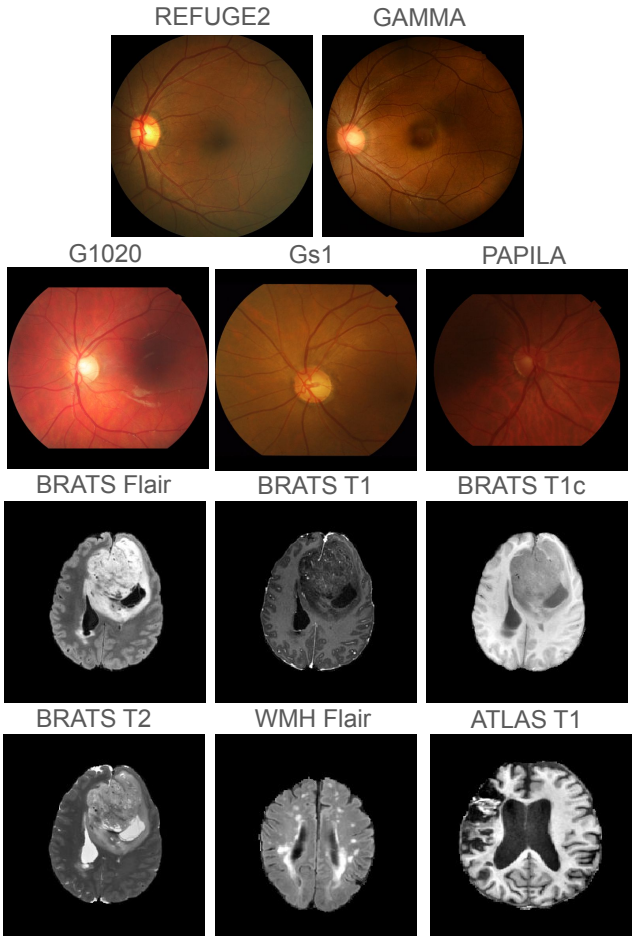

Figure 4: Illustration of the databases used. The distribution across different datasets and modalities can be visually observed.

We observe that relatively small values of β (e.g., 100) and σ (e.g., 1) lead to noticeable performance drops on the BRATS T1 dataset. This suggests that overly small values can hinder the model's ability to effectively utilize the CCG loss. Therefore, we recommend selecting values greater than 100 for β and greater than 1 for σ. The choice of α also influences performance on BRATS T1, while its impact on WMH is minimal. We do not specifically tune α to achieve the best results on BRATS T1.

Overall, performance remains stable in most cases. Note that in our experiments, when comparing with other methods, we did not perform hyperparameter tuning to obtain the best-performing configuration for any specific database. This decision was made due to the relatively stable performance of our method across tasks and to better demonstrate its robustness.

A.4 Ablation Study on PI (10 clicks in total)

As supplementary information to the main paper, we provide the numerical results of the ablation study of Post-Interaction (PI) adaptation under a budget of 10 clicks. The results are shown in Tab. 11. PI continues to offer substantial performance gains in the early iterations; however, by the final click, its advantage diminishes—WMH still benefits, while other methods do not. We continue to recommend deploying both mechanisms in most situations, with further explanation provided in the main paper.

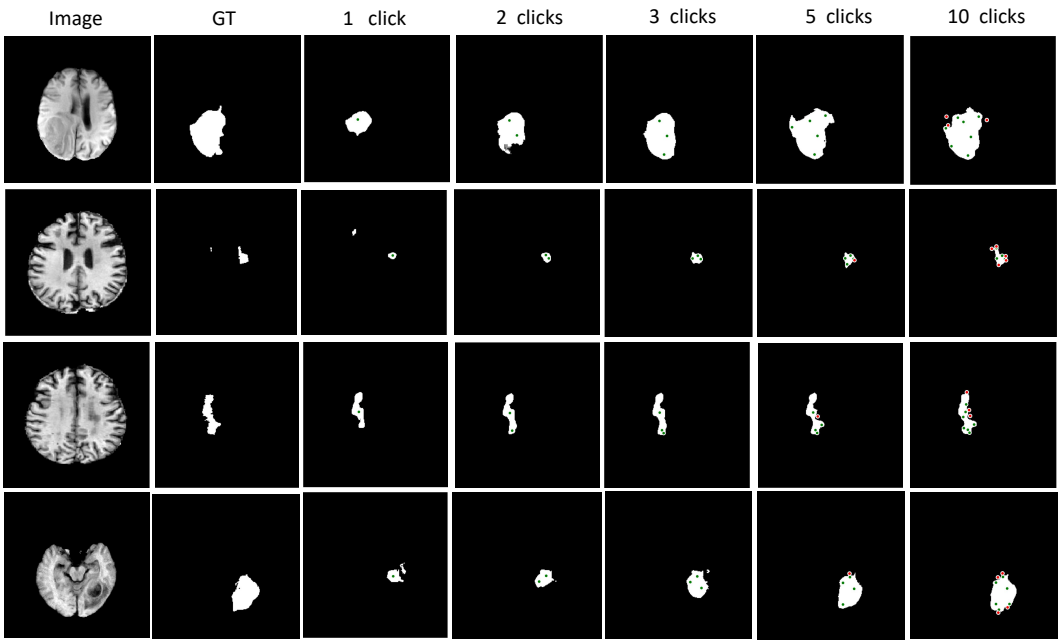

Figure 5: Illustration of how the predicted segmentation from our OAIMS (PI+MI) method evolves as more interaction clicks are provided. The examples are brain MRI images (ATLAS and BRATS T1). From left to right: input image, ground truth, and predictions with 1, 2, 3, 5, and 10 clicks. The results show progressive refinement of the segmentation as more clicks are provided.

Table 10: Dice scores (%) on WMH and BRATS T1 under a 10-click interaction budget. Each value of α, β, and σ is tested in combination with all values of the other two hyperparameters (i.e., 3 α × 4 β × 3 σ total combinations). The reported score for each value is the average over all combinations that include it.

Parameter	Value	BRATS T1 (%)	WMH (%)
α	0.3	85.6	77.7
	0.5	87.0	77.6
	0.7	84.3	78.3
β	100	81.3	77.5
	200	86.9	77.9
	300	87.2	78.1
	400	87.4	78.1
σ	1	81.0	77.2
	3	87.9	78.2
	5	88.1	78.3

A.5 Dice Score

To evaluate segmentation performance, we use the Dice score. It measures the overlap between the predicted segmentation mask P and the ground truth G, and is defined as:

$$\text{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN} \tag{6}$$

Here, TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. A higher Dice score indicates a greater overlap between the prediction and the ground truth.

Table 11: Ablation study for PI with a 10-click budget. Dice shown at 1, 3, 10 clicks. PI and MI are the proposed Post-Interaction process and the Mid-Interaction process of our method.

Dataset	**PI+MI**			**MI**		
	1	3	10	1	3	10
BRATS T1	71.2	79.4	88.0	69.1	77.7	87.6
BRATS T1c	70.4	78.9	87.5	68.3	77.9	87.7
BRATS T2	84.9	88.6	93.0	84.6	88.6	93.0
WMH Flair	58.9	68.6	78.9	58.4	68.1	78.0
TBI Flair	55.2	65.6	76.3	53.3	64.4	76.4

The Dice score is particularly well-suited for medical image segmentation, as it emphasizes accurate delineation of foreground regions—such as lesions—which are often small and sparse. As a result, we adopt Dice as our evaluation metric.

A.6 USE OF LLMS

We used a large language model (LLM) only to improve the writing. Specifically, the LLM was employed to revise some sentences, focusing on grammar and style. The LLM was not used for generating ideas, searching related work, or contributing to the scientific content of the paper.