Rethinking Annotator Simulation: Realistic Evaluation of Whole-Body PET Lesion Interactive Segmentation Methods

Zdravko Marinov[1,2], Moon Kim[3], Jens Kleesiek[3,4], and Rainer Stiefelhagen[1]

[1] Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[3] Institute for AI in Medicine, University Hospital Essen, Essen, Germany
[4] Cancer Research Center Cologne Essen (CCCE), University Medicine Essen, Essen,
[1]{firstname.lastname@kit.edu} , [3]{firstname.lastname@uk-essen.de}

Abstract. Interactive segmentation plays a crucial role in accelerating the annotation, particularly in domains requiring specialized expertise such as nuclear medicine. For example, annotating lesions in whole-body Positron Emission Tomography (PET) images can require over an hour per volume. While previous works evaluate interactive segmentation models through either real user studies or simulated annotators, both approaches present challenges. Real user studies are expensive and often limited in scale, while simulated annotators, also known as robot users, tend to overestimate model performance due to their idealized nature. To address these limitations, we introduce four evaluation metrics that quantify the user shift between real and simulated annotators. In an initial user study involving four annotators, we assess existing robot users using our proposed metrics and find that robot users significantly deviate in performance and annotation behavior compared to real annotators. Based on these findings, we propose a more realistic robot user that reduces the user shift by incorporating human factors such as click variation and inter-annotator disagreement. We validate our robot user in a second user study, involving four other annotators, and show it consistently reduces the simulated-to-real user shift compared to traditional robot users. By employing our robot user, we can conduct more large-scale and cost-efficient evaluations of interactive segmentation models, while preserving the fidelity of real user studies. Our implementation is based on MONAI Label and will be made publicly available.

Keywords: Interactive segmentation · Robot user · Realistic simulation

1 Introduction

Deep learning models have made significant progress in segmenting anatomical structures and lesions in medical images but often rely on manually labeled datasets [1–6]. This poses a challenge for volumetric medical data where annotating each voxel demands considerable time and expertise. Interactive segmentation mitigates this issue by leveraging less demanding annotations, such

as clicks, instead of dense voxelwise labels [7–15]. Clicks are combined with the image as a joint input for the interactive model and guide it spatially toward the segmentation target. Annotators can refine model outputs by placing clicks in missegmented areas, leading to an improved segmentation and high-quality predictions [9–15]. Once approved by medical experts, these predictions may serve as new labels [7]. Prior methods evaluate interactive models by simulating clicks on the test split (a "robot user") [15–18] or by involving real annotators in a user study [9–11]. However, real user studies are costly, with a limited sample size, and robot users often overestimate model performance due to their idealized nature. Similar to a domain shift encountered when assessing models with out-of-domain data (e.g., from a different scanner), a *user shift* arises when validating an interactive model via simulated robot users and deploying it in real clinical settings, where its performance often diverges [7]. We address these challenges for whole-body PET lesion segmentation with the following contributions:

1. We evaluate 4 robot users **(R1)**–**(R4)** on the AutoPET dataset [1] and conduct 2 user studies, each with 4 medical annotators, to show the disparity between simulated and real user performance of existing robot users.
2. We introduce 4 evaluation metrics **(M1)**–**(M4)** to quantify the simulated-to-real user shift in terms of segmentation accuracy, annotator behavior, and conformity to ground-truth labels.
3. We propose a novel robot user that mitigates the pitfalls identified in 1. by simulating clicks that disagree with the ground-truth labels. Our robot user reduces the user shift (defined in 2.) and the segmentation performance gap to real users compared to previous robot users in both our user studies.

Related Work. Previous research on robot users mainly explores classical non-deep learning methods and overlooks evaluating the disparity with real annotators. For example, Kohli et al. [18] compare four Graph Cut-based interactive models [19] and conclude that placing clicks at the center of the largest error consistently yields optimal results across all models. However, their comparison is limited to natural images, and they do not explore deep learning-based approaches. Moschidis and Graham [16] compare two robot users for 3D medical image segmentation: one targeting central regions and the other - boundary regions. However, their study also examines classical non-deep learning methods and lacks simulated clicks for iterative corrections. Benenson et al. [20] compare iterative boundary and central clicks, discovering that central clicks outperform boundary clicks, particularly when adding random noise perturbations, however, they also only explore the domain of natural images. The closest work to ours is Amrehn et al. [17], which compares robot users using an interactive U-Net [21] for liver lesion segmentation. Their results suggest that a U-Net trained with a robot user using more spatially distributed clicks generalizes well when evaluated with a different robot user. However, they do not explore the generalization to real annotator interactions. In contrast to previous work, our focus lies on evaluating deep learning-based methods incorporating iterative corrections, with an emphasis on reducing the disparity between simulated and real annotators. Interactive segmentation reviews [7, 8] have discussed the lack of user-centric metrics

for medical interactive segmentation. We address this by introducing 4 metrics that capture user behavior and quantify the simulated-to-real user shift.

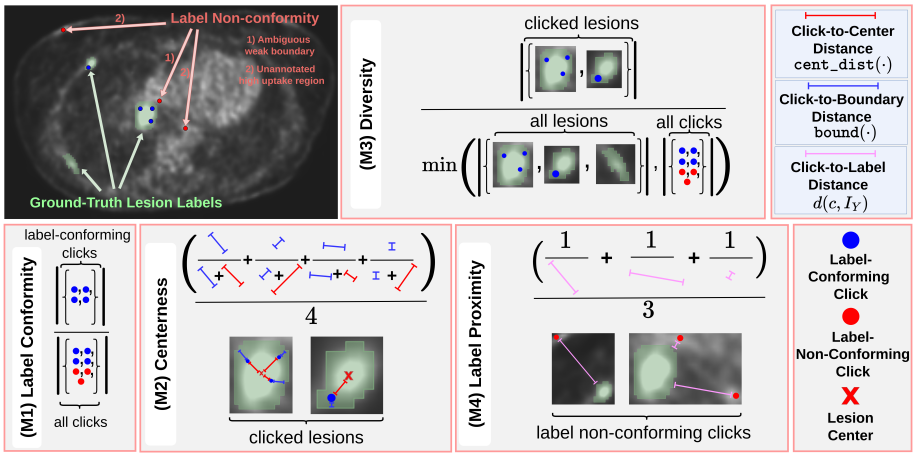

Fig. 1: Our proposed evaluation metrics and examples of label non-conformity.

2 Methods

We explore iterative interactive models that simulate clicks in a loop of 10 iterations. In each click iteration $i \in \{1, ..., 10\}$, a robot user R simulates a click, denoted as $\texttt{clicks}(R, I)[i] \in \mathbb{N}^3$, and combines it with the image $I \in \mathbb{R}^{W \times H \times D}$ as a joint input, where $W \times H \times D$ are the image dimensions. Using this joint input, the model predicts a segmentation mask $\texttt{pred}(I)[i] \in \{0, 1\}^{W \times H \times D}$. Then, the missegmented regions within this prediction, denoted as $\texttt{err}(I)[i] \in \{0, 1\}^{W \times H \times D}$, are employed to generate $\texttt{clicks}(R, I)[i + 1]$ for the next iteration. We provide a notation table for all our equation terms in the supplementary.

2.1 Robot Users

(R1) Center Click: A common approach is to simulate clicks in the center of the largest missegmented component [7, 22]. However, the first click is placed in the center of the largest component of the label I_Y. This is defined as:

$$\texttt{clicks}(R1, I)[i] = \begin{cases} \texttt{center}(\texttt{largest\_component}(I_Y)), & \text{if } i = 1 \\ \texttt{center}(\texttt{largest\_component}(\texttt{err}(I)[i-1])), & \text{if } i > 1 \end{cases} \quad (1)$$

where $I_Y \in \{0, 1\}^{W \times H \times D}$ is the ground-truth label for image I, $\texttt{center}(\cdot)$ computes the geometric center of a component as in [22], and $\texttt{largest\_component}(\cdot)$ computes the largest connected component.

(R2) Uncertainty: Zheng et al. [23] sample a click in each iteration using the epistemic uncertainty of the model as a sampling distribution, defined as:

$$\texttt{clicks}(R2, I)[i] \sim \begin{cases} \texttt{uniform}(I_Y), & \text{if } i = 1 \\ \texttt{epistemic}(\texttt{pred}(I)[i-1]), & \text{if } i > 1 \end{cases} \tag{2}$$

where $\texttt{epistemic}(\cdot)$ is the normalized epistemic uncertainty in $[0, 1]$ using Monte Carlo Dropout [24], and $\texttt{uniform}(X)$ defines a uniform distribution over the non-zero entries of X.

(R3) Euclidean Distance Transform (EDT): Previous methods [9, 10] apply the EDT on missegmented regions as a sampling distribution for clicks:

$$\texttt{clicks}(R3, I)[i] \sim \begin{cases} \texttt{uniform}(I_Y), & \text{if } i = 1 \\ \texttt{EDT}(\texttt{err}(I)[i-1]), & \text{if } i > 1 \end{cases} \tag{3}$$

where $\texttt{EDT}(\texttt{err}(I)[i-1])$ is the normalized EDT of the non-zero entries in the missegmented regions $\texttt{err}(I)[i-1]$ from the previous iteration.

(R4) Uniform: The final robot user samples uniformly either from the previous error [17] or from the label for the first click as:

$$\texttt{clicks}(R4, I)[i] \sim \begin{cases} \texttt{uniform}(I_Y), & \text{if } i = 1 \\ \texttt{uniform}(\texttt{err}(I)[i-1]), & \text{if } i > 1 \end{cases} \tag{4}$$

Note: In each iteration we simulate two types of clicks: $\texttt{clicks}(R, I)[i]^{\text{lesion}}$ and $\texttt{clicks}(R, I)[i]^{\text{background}}$. We designate the under- and over-segmented regions as missegmented areas $\texttt{err}(I)[i]$ for the "lesion" and "background" classes respectively, and omit the class labels in Eq.(1)-(4), for clarity.

($\mathbf{R_{ours}}$): Our Robot User: In our first user study, we found that 25% of our annotators' clicks are outside the ground-truth labels. Label non-conforming clicks stem from two factors (see Fig. 1, top left): 1) ambiguous weak boundaries in the low-resolution PET scans, leading to clicks slightly outside the label boundaries; 2) and unannotated high uptake regions, spatially isolated from ground-truth labels. To address the first issue, we propose integrating click perturbations to spatially displace clicks with a probability p_{perturb}. For the second issue, we propose to systematically incorporate label non-conformity by sampling clicks in high uptake regions outside the ground-truth labels with a probability p_{system}. To achieve this, our robot user extends **(R1)** and is defined as:

$$\texttt{clicks}(R_{\text{ours}}, I)[i] = \begin{cases} \texttt{clicks}(R1, I)[i] & \text{if } p_{i,1} \geq p_{\text{perturb}} \text{ and } p_{i,2} \geq p_{\text{system}} \\ \texttt{clicks}(R1, I)[i] + \widetilde{z}, & \text{if } p_{i,1} < p_{\text{perturb}} \text{ and } p_{i,2} \geq p_{\text{system}} \\ \widetilde{s}, & \text{if } p_{i,1} \geq p_{\text{perturb}} \text{ and } p_{i,2} < p_{\text{system}} \\ \widetilde{s} + \widetilde{z}, & \text{if } p_{i,1} < p_{\text{perturb}} \text{ and } p_{i,2} < p_{\text{system}} \end{cases} \tag{5}$$

where $\widetilde{s} \sim \texttt{SUV}(I, I_Y)$ and $\widetilde{z} \sim \mathcal{U}_{[-a,a]^3}$. $\texttt{SUV}(I, I_Y)$ defines a distribution over the normalized Standardized Uptake Values in I which are outside the label I_Y, $\widetilde{z}$ is a random perturbation with a maximal amplitude $a \in \mathbb{N}$, and each iteration $p_{i,1}$, $p_{i,2}$ are independently sampled from $\mathcal{U}_{[0,1]}$ to decide which case is applied.

2.2 Model Architecture and Dataset

We use the pre-trained SW-FastEdit [9] interactive model based on MONAI Label [25] with a U-Net backbone [21] and conduct our user studies on the openly available AutoPET [1] dataset which consists of 1014 PET/CT volumes with annotated tumor lesions of melanoma, lung cancer, or lymphoma. We exclusively utilize PET data and use SW-FastEdit's official test split of 10% of the volumes. The PET volumes have a voxel size of $2.0 \times 2.0 \times 3.0\text{mm}^3$ and an average resolution of $400 \times 400 \times 352$ voxels. Both user studies were conducted using 3D Slicer [26] and its MONAI Label plugin. We implemented our robot user experiments with MONAI Label [25] and will release the code.

3 Experiments and Results

3.1 Evaluation Metrics

For all metrics, we denote $\mathcal{I}$ as the set of PET images labeled in a user study, $\mathcal{A}$ as the set of real annotators participating in the study, and fix the number of clicks per image to 10. We visualize examples for **(M1)-(M4)** in Fig. 1.

(M1) The Label Conformity for an annotator A is defined as:

$$\mathbf{M}_1(A) = \tfrac{1}{|\mathcal{I}|}\tfrac{1}{10}\textstyle\sum_{I\in\mathcal{I}}\sum_{i=1}^{10}\left[I_Y[\texttt{clicks}(A,I)[i]] = 1\right] \tag{6}$$

where $[\cdot]$ is the Iverson bracket. **(M1)** measures to what extent an annotator's clicks agree with the ground-truth labels of the PET images.

(M2) The Centerness for annotator A is defined as:

$$\mathbf{M}_2(A) = \tfrac{1}{|\mathcal{I}|}\tfrac{1}{|\bar{C}(A,I)|}\textstyle\sum_{I\in\mathcal{I}}\sum_{c\in\bar{C}(A,I)}\frac{\texttt{bound}(c,I_Y)}{\texttt{bound}(c,I_Y)+\texttt{cent\_dist}(c,I_Y)} \tag{7}$$

where $\bar{C}(A,I) = \{c \mid c \in \texttt{clicks}(A,I) \text{ and } I_Y[c] = 1\}$ is the set of label conforming clicks of annotator A for image I, $\texttt{bound}(c, I_Y)$ is the minimum distance of click c to the boundary of the label I_Y, and $\texttt{cent\_dist}(c, I_Y)$ is the minimum distance of click c to the center of the label I_Y. Small **(M2)** values indicate that label-conforming clicks are placed near the boundary, whereas large values show that clicks are placed near the central regions of the label.

(M3) The Click Diversity for annotator A is defined as:

$$\mathbf{M}_3(A) = \tfrac{1}{|\mathcal{I}|}\textstyle\sum_{I\in\mathcal{I}}\frac{|\{\widetilde{Y} \mid \widetilde{Y}\in\texttt{components}(I_Y) \text{ and } \exists c\in\texttt{clicks}(A,I)\text{: } c\in\widetilde{Y}\}|}{\min(|\texttt{components}(I_Y)|,\ |\texttt{clicks}(A,I)|)} \tag{8}$$

where $\texttt{components}(\cdot)$ is the set of all connected components. **(M3)** measures to what extent clicks are spread out in different connected components in the label.

(M4) The Label Proximity for an annotator A is defined as:

$$\mathbf{M}_4(A) = \tfrac{1}{|\mathcal{I}|}\tfrac{1}{|\hat{C}(A,I)|}\textstyle\sum_{I\in\mathcal{I}}\sum_{c\in\hat{C}(A,I)}\frac{1}{d(c,I_Y)} \tag{9}$$

where $\hat{C}(A,I) = \{c \mid c \in \texttt{clicks}(A,I) \text{ and } I_Y[c] = 0\}$ is the set of label non-conforming clicks of annotator A for image I, and $d(c, I_Y) = \min(\{||c-y|| \mid y \in$

$\mathbb{N}^{W\times H\times D}$ and $I_Y[y]=1\}$). **(M4)** computes the average inverse distance of the annotator clicks outside the ground-truth label to the label I_Y. Higher **(M4)** values suggest non-conforming clicks are close to the label boundary, while lower values indicate clicks are far from any component of the label I_Y, suggesting systematic non-conformity.

(M5) The Consistent Improvement is defined in [15] as:

$$\mathbf{M}_5(A) = \tfrac{1}{|\mathcal{I}|}\tfrac{1}{10}\sum_{I\in\mathcal{I}}\sum_{i=1}^{10}\left[\texttt{dice}(A,I)[i] > \texttt{dice}(A,I)[i-1]\right] \tag{10}$$

where $\texttt{dice}(A,I)[i]$ is the Dice score after annotator A's i^{th} click on image I.

(M6) The User Shift determines the mean absolute difference in all metrics **(M1)-(M5)** between a simulated robot user R and all real annotators $\mathcal{A}$:

$$\mathbf{M}_6(R,\mathcal{A}) = \tfrac{1}{|\mathcal{A}|}\tfrac{1}{5}\sum_{A\in\mathcal{A}}\sum_{i=1}^{5}|\mathbf{M}_i(R)-\mathbf{M}_i(A)| \tag{11}$$

(M7) The Dice Difference for a robot user R is defined as:

$$\mathbf{M}_7(R,\mathcal{A}) = \tfrac{1}{|\mathcal{I}|}\tfrac{1}{|\mathcal{A}|}\tfrac{1}{10}\sum_{I\in\mathcal{I}}\sum_{A\in\mathcal{A}}\sum_{i=1}^{10}|\texttt{dice}(A,I)[i]-\texttt{dice}(R,I)[i]| \tag{12}$$

(M6) quantifies the fidelity of the robot user in emulating annotator behavior, while **(M7)** evaluates its ability to reproduce the segmentation performance of the interactive model as used by real annotators.

3.2 User Studies and Results

Setup. We conduct two user studies, each with four annotators from a medical background. In both studies, annotators were instructed to place 10 "lesion" and 10 "background" clicks, updating the model prediction after each pair of clicks to replicate the workflow of simulated robot users. In our first user study, four annotators labeled the same 10 PET volumes from the test split. We used this user study to determine the optimal values of p_{perturb} and p_{system} for our robot user. In our second user study, four different annotators labeled 6 PET volumes. We conducted this as a "validation" user study to confirm that our results from the first user study generalize to other volumes and annotators. For both studies, we applied each robot user to the same PET images annotated by the real users.

Table 1: User Shift and Dice Difference of all robot users on both user studies.

		Previous Work				Ours ($a=35$)				
		(R1)	**(R2)**	**(R3)**	**(R4)**	p_{perturb} 25%	19.6%	13.4%	6.7%	0%
						p_{system} 0%	6.7%	13.4%	19.6%	25%
User Study 1	**(M6)** User Shift	27.4	35.0	28.5	29.5	9.4	8.4	**6.8**	9.0	11.6
	(M7) Dice Difference	8.7	10.0	9.2	11.6	6.0	5.3	**3.6**	5.8	6.9
User Study 2	**(M6)** User Shift	30.0	31.7	33.8	30.0	8.4	7.6	**6.7**	8.6	9.2
	(M7) Dice Difference	8.5	9.0	7.0	7.5	5.3	4.8	**3.7**	6.2	6.7

Results: Our Robot User. In the first user study, we assessed our robot user by varying p_{perturb}, p_{system} and the perturbation amplitude a and plotted

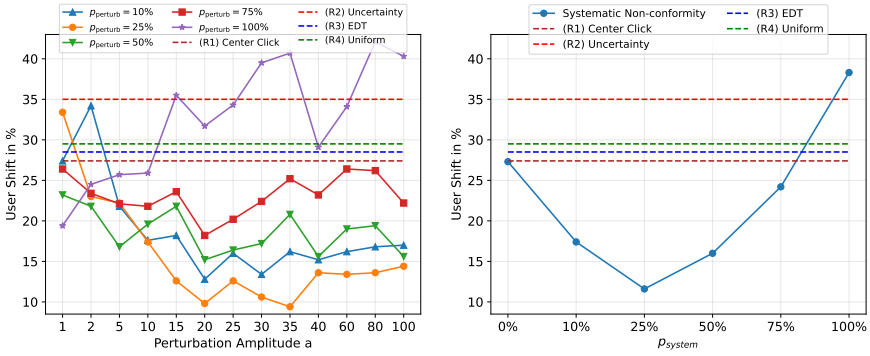

Fig. 2: Analysis of p_{perturb} (left) and p_{system} (right) in our first user study.

the results in Fig. 2. Spatial perturbations with $p_{\text{perturb}} \leq 75\%$ consistently outperform existing robot users in terms of user shift. The optimal user shift is achieved with $p_{\text{perturb}} \leq 75\%$ and $a \in [20, 35]$, in particular with $p_{\text{perturb}} = 25\%$ and $a = 35$, deteriorating with $a > 35$ or $p_{\text{perturb}} = 100\%$ due to the excessive spatial noise. Incorporating systematic non-conformity also consistently reduces the user shift, with $p_{\text{system}} = 25\%$ as the optimal value, similar to p_{perturb}. Since 25% is the optimal value for both p_{perturb} and p_{system}, we explore mixing them with a joint probability of 25%. The results in Table 1 show that mixing further reduces the user shift as well as the Dice difference, leading to optimal results when $p_{\text{system}} = p_{\text{perturb}}$.

Results: Previous Work. The results, plotted in Fig. 3 and Table 1 reveal a large discrepancy between existing robot users and the average annotator in all metrics. This contrast is especially notable in **(M1)** and **(M4)** since robot users always produce label-conforming clicks, while real annotators click outside the label in 25% of their interactions. Building on this insight, our robot user introduces label non-conformity in 25% of its simulated clicks by spatially per-

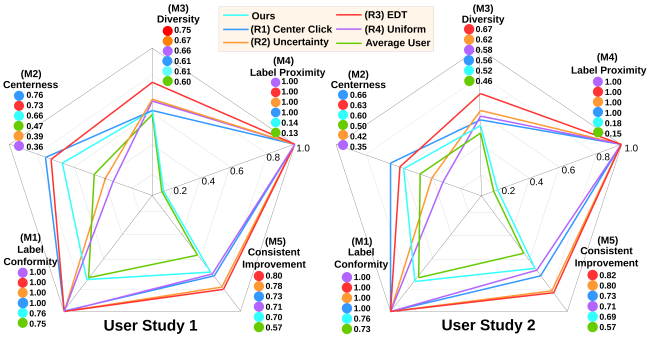

Fig. 3: Metric values **(M1)-(M5)** of all robot users on both user studies.

turbing clicks and systematically sampling from high-uptake regions outside the label. This non-conformity achieves the optimal user shift and Dice difference in both user studies. Our robot user reduces the Dice difference from 8.7% to 3.6% and from 7.0% to 3.7% on the first and second user study respectively, which confirms that the Dice score reported when evaluating with our robot user is much more realistic. The Dice curves are visualized in Fig. 4.

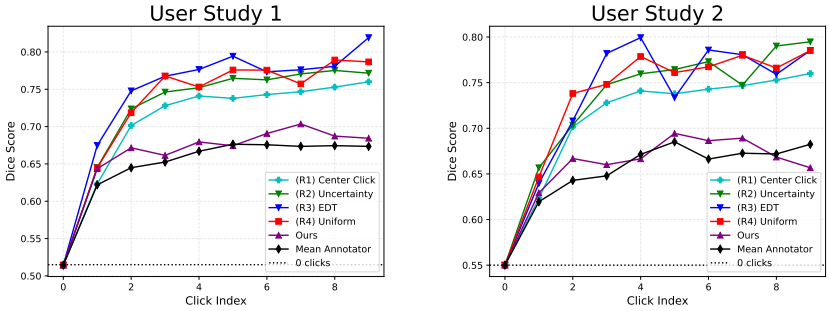

Fig. 4: Mean Dice curves of all robot users for both user studies.

User Shift vs. Dice Difference. As the user shift only quantifies the behavioral shift, we examine its correlation with the Dice difference for all our robot user configurations in the first user study. Fig. 5 reveals a Pearson correlation of $\rho = 0.89$ between the user shift and the Dice difference. Importantly, omitting any of our metrics **(M1)-(M5)** from **(M6)** decreases the correlation to $\rho < 0.8$. This confirms that our proposed metrics not only quantify the annotation style but also quantify how this style influences the segmentation performance.

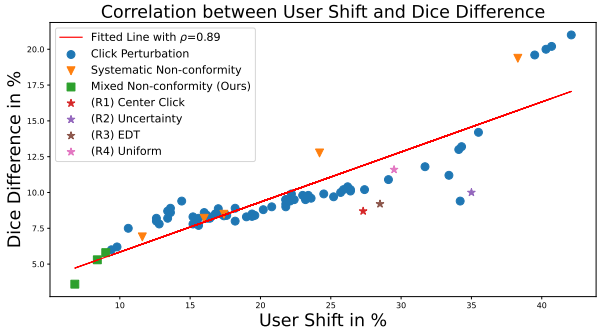

Fig. 5: Correlation between **(M6)** and **(M7)** on the first user study results.

4 Conclusion

Our user studies reveal the challenges in evaluating interactive models through simulated interactions. Despite its simplicity and dependence on the careful choice of hyperparameters, our robot user exposes fundamental flaws in traditional robot users that heavily rely on ground-truth labels. This is particularly problematic in domains where experts disagree on the ground truth in 25% of

their interactions, as observed in our user studies for whole-body PET lesion annotation. Traditional robot users exhibit significant user shift and Dice difference compared to real annotators, resulting in overly optimistic Dice scores and unrealistic annotation behavior. By incorporating click perturbations and systematic label non-conformity, we substantially reduce the user shift and Dice difference compared to previous robot users. This facilitates a more realistic evaluation of interactive model performance without the need for extensive user studies involving the entire test split.

Acknowledgements. The user studies were done in collaboration with the Annotation Lab Essen (https://annotationlab.ikim.nrw/). The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health. This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

References

1. Gatidis, Sergios, et al. "The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging." (2023).
2. Menze, Bjoern H., et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." IEEE transactions on medical imaging 34.10 (2014): 1993-2024.
3. Antonelli, Michela, et al. "The medical segmentation decathlon." Nature communications 13.1 (2022): 4128.
4. Ji, Yuanfeng, et al. "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation." Advances in Neural Information Processing Systems 35 (2022): 36722-36732.
5. Wasserthal, Jakob, et al. "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images." Radiology: Artificial Intelligence 5.5 (2023).
6. Hernandez Petzsche, Moritz R., et al. "ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset." Scientific data 9.1 (2022): 762.
7. Marinov, Zdravko, et al. "Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy." arXiv preprint arXiv:2311.13964 (2023).
8. Zhao, Feng, and Xianghua Xie. "An overview of interactive medical image segmentation." Annals of the BMVA 2013.7 (2013): 1-22.
9. Hadlich, Matthias, et al. "Sliding Window FastEdit: A Framework for Lesion Annotation in Whole-body PET Images." arXiv preprint arXiv:2311.14482 (2023).
10. Hallitschke, V.J., et al. "Multimodal Interactive Lung Lesion Segmentation: A Framework for Annotating PET/CT Images Based on Physiological and Anatomical Cues," 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 2023, pp. 1-5.
11. Asad, Muhammad, et al. "Adaptive Multi-scale Online Likelihood Network for AI-Assisted Interactive Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.
12. Wang, Guotai, et al. "DeepIGeoS: a deep interactive geodesic framework for medical image segmentation." IEEE transactions on pattern analysis and machine intelligence 41.7 (2018): 1559-1572.

13. Luo, Xiangde, et al. "MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning." Medical image analysis 72 (2021): 102102.
14. Wang, Guotai, et al. "Interactive medical image segmentation using deep learning with image-specific fine tuning." IEEE transactions on medical imaging 37.7 (2018): 1562-1573.
15. Marinov, Z., Stiefelhagen R., Kleesiek J. "Guiding the Guidance: A Comparative Analysis of User Guidance Signals for Interactive Segmentation of Volumetric Images." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.
16. Moschidis, Emmanouil, and Jim Graham. "A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction." 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, 2010.
17. Amrehn, Mario, et al. "Interactive neural network robot user investigation for medical image segmentation." Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck. Springer Fachmedien Wiesbaden, 2019.
18. Kohli, Pushmeet, et al. "User-centric learning and evaluation of interactive segmentation systems." International journal of computer vision 100 (2012): 261-274.
19. Boykov, Yuri, and Gareth Funka-Lea. "Graph cuts and efficient ND image segmentation." International journal of computer vision 70.2 (2006): 109-131.
20. Benenson, Rodrigo, Stefan Popov, and Vittorio Ferrari. "Large-scale interactive object segmentation with human annotators." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
21. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.
22. Liu, Qin, et al. "iSegFormer: interactive segmentation via transformers with application to 3D knee MR images." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022.
23. Zheng, Ervine, et al. "A continual learning framework for uncertainty-aware interactive image segmentation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 7. 2021.
24. Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. PMLR, 2016.
25. Diaz-Pinto, Andres, et al. "Monai label: A framework for ai-assisted interactive labeling of 3d medical images." arXiv preprint arXiv:2203.12362 (2022).
26. Fedorov, Andriy, et al. "3D Slicer as an image computing platform for the Quantitative Imaging Network." Magnetic resonance imaging 30.9 (2012): 1323-1341.