Temporal Gaze Dynamics as Zero-Shot Prompts for Volumetric Medical Segmentation

Tatyana Shmykova **Ilya Pershin**
Research Center of the Artificial Intelligence Institute, Innopolis University, Innopolis, Russia
{t.shmykova, i.pershin}@innopolis.ru

Abstract

Guiding foundation models like SAM-2 for volumetric medical segmentation typically relies on inefficient manual prompts. We introduce a more efficient, multimodal approach using eye gaze—a continuous physiological time series—to steer the model's focus in a zero-shot manner. By fusing a user's temporal gaze stream with spatial image data, we enable dynamic, interactive 3D segmentation. Evaluating with SAM-2 and its medical variant, MedSAM-2, our gaze-based method proves significantly more time-efficient (e.g., 62 vs. 88 seconds per volume) than manual bounding boxes, with a modest accuracy trade-off. This work establishes a practical framework for incorporating human physiological signals into sequential, human-in-the-loop clinical tasks, paving the way for more intuitive AI interfaces.

1 Introduction

Physiological time-series data are fundamental to healthcare [1, 2], and their integration with foundation models offers new opportunities for enhancing clinical workflows [3, 4]. A critical application is medical image segmentation, where guiding large vision models like the Segment Anything Model (SAM) [5] and its medical derivatives [6, 7] still relies on discrete, labor-intensive prompts. This interaction paradigm is a significant bottleneck for volumetric data [8], limiting the practical utility of these powerful models.

This paper proposes to bridge this gap by treating the human gaze—a rich physiological time-series signal reflecting an expert's cognitive focus—as a primary input modality. We frame this as a multimodal learning problem where a temporal gaze stream is fused with spatial image data to guide a foundation model in a zero-shot setting. This transforms segmentation into a more natural, sequential human-in-the-loop task where the AI adapts to the user's attention rather than waiting for discrete commands [9].

Our work addresses the limitations of other prompting modalities. For instance, text-based prompts showed poor performance in our experiments with MedCLIP-SAMv2 [10] on the WORD dataset [11]. Prior work on gaze-assisted segmentation has primarily focused on 2D images or required model fine-tuning [12, 13, 14], which is impractical for many narrow-domain medical tasks. Our primary contribution is a novel, zero-shot framework for 3D interactive segmentation that leverages the raw gaze time series to guide pre-trained foundation models. We show that this method is significantly faster than manual bounding boxes, paving the way for more efficient human-AI collaboration in medicine.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health (TS4H).

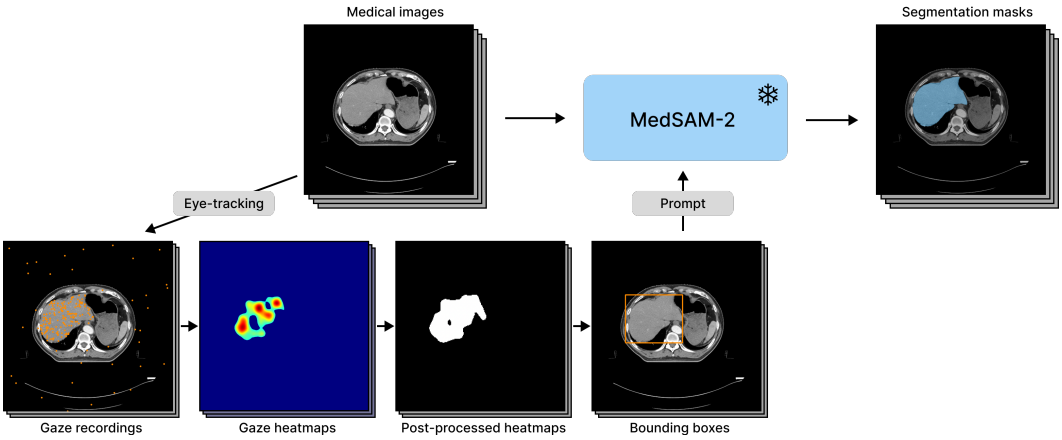

Figure 1: The process of generating segmentation masks based on the annotator's gaze stream involves the following steps: (1) tracking of eye gaze, (2) gaze heatmaps generation, (3) heatmaps post-processing, (4) bounding box-based prompt construction, and (5) segmentation using a pre-trained interactive segmentation model, e.g., MedSAM-2, with frozen parameters.

2 Related work

The zero-shot segmentation method in MedCLIP-SAMv2 integrates multimodal learning, starting with a specialized vision-language model (BiomedCLIP) to extract image and text embeddings. These are refined via a Multi-modal Information Bottleneck (M2IB) mechanism to enhance relevant visual information while suppressing noise, resulting in saliency maps. These maps are post-processed with K-Means clustering to generate a coarse segmentation mask, which then provides bounding box or point-based prompts to the Segment Anything Model (SAM) for final refinement.

The authors of MedCLIP-SAMv2 demonstrated the importance of prompt engineering, comparing simple class-name prompts (P0) with more descriptive sentences (P2) and ensembles. On datasets like Lung CT, performance improved significantly with better prompts, with the Dice score rising from 69.89 (P0) to 80.38 (P2).

However, while MedCLIP-SAMv2 claims strong zero-shot capabilities, our own evaluation found its performance to be notably poor on the WORD dataset. We tested the approach on 600 slices using their prescribed P0 and P2 text prompts. While ground-truth bounding boxes achieved a high Dice score of 0.88, the text prompts performed drastically worse, yielding scores of only 0.16 (P0) and 0.2 (P2). This large performance drop indicates that the text-based approach is not universally effective and motivates our exploration of alternative prompting modalities.

3 Methods

We frame interactive volumetric segmentation as a sequential, human-in-the-loop process where a user's gaze stream—a continuous physiological time series—guides a foundation model. Our framework (Figure 1) has three stages: converting temporal gaze data into a spatial attention map, generating coarse prompts from it, and refining them with a pre-trained model.

Distilling Spatial Saliency from a Temporal Gaze Sequence First, we distill the raw gaze time series into a 2D spatial saliency map (a gaze heatmap) for each CT slice. A Gaussian filter aggregates gaze fixations over time, and K-Means clustering then binarizes the heatmap to produce a coarse segmentation mask, similar to the process in [10]. This method directly leverages the continuous signal of human visual focus, avoiding reliance on semantic text prompts.

Segmentation Refinement via Foundation Models Next, the coarse segmentation mask is refined by the pre-trained foundation models SAM-2 [8] and MedSAM-2 [7]. We automatically derive bounding box prompts from the coarse mask's contours and feed them to the models on a slice-by-slice basis. This two-step process allows for zero-shot segmentation refinement without any model fine-tuning, translating the gaze signal into an optimal prompt format for the models.

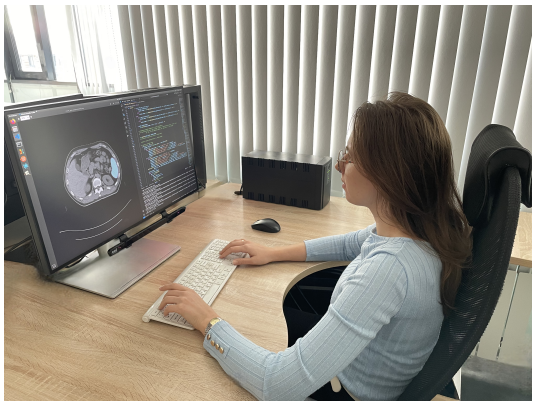

Figure 2: Gaze-based system that enables radiologists to segment abdominal organs on CT scans using an eye-tracker.

Efficient 3D Propagation via Sparse Prompts and Interpolation Finally, to minimize user effort, we employ an efficient strategy where prompts are provided on only a sparse subset of slices. The full 3D segmentation is then reconstructed via shape-based interpolation between the resulting boolean masks [15]. This method uses a distance transform and linear interpolation to generate smooth, anatomically plausible contours for the entire volume, significantly reducing the annotation burden.

4 Experiments and Results

We evaluate the performance and efficiency of SAM-2 and MedSAM-2 using both synthetic and real gaze data. Gaze-based prompts are compared against two bounding box (bbox) baselines: those synthetically generated from ground truth masks and real bboxes provided by a medical expert. For real-world validation, we employed a proxy radiologist who was trained in abdominal organ segmentation under expert supervision, ensuring an accurate and consistent evaluation of our method.

4.1 Experimental Setting

Gaze tracking. We develop a radiologist workstation with integrated eye-tracking functionality, designed for use in a dedicated, isolated room at our institution. The workstation is equipped with a lightweight, user-friendly, bar-shaped eye-tracking device positioned beneath the monitor for convenience. The hardware setup includes an LG diagnostic monitor featuring 10-bit color depth, a resolution of 3840×2160 pixels, and a pixel density of 7.21 px/mm. The eye-tracking functionality is facilitated by a Tobii Eye Tracker 4C, which operates at a frequency of 90 Hz.

Dataset. We utilize the WORD dataset [11], encompassing abdominal volumetric CT images, to explore using the annotator's gaze as prompt for SAM-2 and MedSAM-2. This dataset comprises 150 CT scans from 150 patients, covering 16 abdominal organs. Each CT scan contains between 159 and 330 slices, each with a resolution of 512 × 512 pixels. For experiments, we use 350 CT slices of 16 abdominal organs.

Synthetic gaze data. For testing models on synthetic data, we generate gaze points by sampling coordinates from reference segmentation masks for abdominal organs in 2D CT slices from WORD [11]. Specifically, 80% of the points are randomly generated inside each organ area and 20% outside the organ, simulating natural gaze fluctuations and assuming potential inaccuracies in eye-tracking data [14].

4.2 Different strategies for prompt

We test different synthetic prompt strategies, changing the number of slices on which prompts are provided. According to Tables 1 and 2, when limited to the first slice, segmentation performance decreases substantially for both models (0.661 for SAM-2 and 0.670 for MedSAM-2). Limiting the number of prompts to 30 slices maintains similar accuracy compared to providing prompts

Table 1: Comparison of using different numbers of slices on which the prompt is provided on synthetic data: (1) only on the first slice, (2) on all slices, and (3) on 30 slices. Prompts generated based on **ground truth mask** boundaries.

Model	Method	Dice score	Time (sec)
SAM-2	First slice	0.661 ± 0.318	12 ± 1
	All slices	0.896 ± 0.073	133 ± 4
	30 slices	$\mathbf{0.896 \pm 0.073}$	$\mathbf{102 \pm 4}$
MedSAM-2	First slice	0.670 ± 0.353	11 ± 1
	All slices	0.904 ± 0.063	133 ± 4
	30 slices	$\mathbf{0.904 \pm 0.063}$	$\mathbf{102 \pm 4}$

Table 2: Comparison of using different numbers of slices on which the prompt is provided on synthetic data: (1) only on the first slice, (2) on all slices, and (3) on 30 slices. Prompts generated based on **synthetic gaze heatmaps**.

Model	Method	Dice score	Time (sec)
SAM-2	All slices	0.809 ± 0.171	108 ± 7
	30 slices	$\mathbf{0.812 \pm 0.172}$	$\mathbf{82 \pm 7}$
MedSAM-2	All slices	0.814 ± 0.163	107 ± 7
	30 slices	$\mathbf{0.817 \pm 0.162}$	$\mathbf{82 \pm 7}$

on all images (0.896 for SAM-2 and 0.904 for MedSAM-2 using bbox) and significantly reduces segmentation time (133 vs. 102 seconds using bbox). Synthetic gaze-based prompts, while slightly less accurate than bounding boxes, significantly reduce the time to get the masklet (82 vs. 102 secs).

4.3 Effectiveness of Gaze prompts

Table 3: Comparison of segmentation performance and mean time between SAM-2 and MedSAM-2 on real prompts provided by the proxy radiologist. Prompt methods: Bbox (bounding boxes), Gaze (prompts generated based on gaze heatmaps).

Model	Prompts	Dice	Time (sec)
SAM-2	Bbox	$\mathbf{0.834 \pm 0.124}$	88 ± 4
	Gaze	0.750 ± 0.204	$\mathbf{63 \pm 7}$
MedSAM-2	Bbox	$\mathbf{0.844 \pm 0.115}$	88 ± 4
	Gaze	0.759 ± 0.203	$\mathbf{62 \pm 7}$

According to Table 3 bounding boxes drawn by the radiologist results in better segmentation performance (0.834 for SAM-2, 0.844 for MedSAM-2) compared to gaze-based prompts (0.750 and 0.759, respectively). However, using gaze-based prompts remains efficient, requiring less time than manual bounding boxes (62 vs. 88 secs).

5 Conclusion and Future Work

In this work, we demonstrate a multimodal framework where a user's gaze, treated as a physiological time series, guides foundation models like SAM-2 for 3D medical segmentation. This zero-shot approach is significantly faster than manual prompting, with a minor accuracy trade-off. Future work will prioritize modeling the gaze stream directly as a time series (e.g., with Transformers), deriving digital biomarkers from gaze patterns (e.g., fatigue), and using multi-resolution analysis for more robust prompting. Clinical validation remains a crucial final step to ensure real-world impact.

6 Acknowledgment

The study was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

References

[1] Lei Clifton, David A Clifton, Marco AF Pimentel, Peter J Watkinson, and Lionel Tarassenko. Big data from physiological sensors. *IEEE Engineering in Medicine and Biology Magazine*, 33(3):21–28, 2014.

[2] Jaywardhan Shukla, Rotem Bar-Sela, and Shahar Kuten. Machine learning for the analysis of physiological time-series signals: A review. *Electronics*, 10(17):2057, 2021.

[3] Michael Moor, Oana Oala, Kerstin N Vokinger, Larissa Retschnig, Anja Musan, Thomas C Kwee, Frank Kading, Anja Bucher, Amir Amini, Verena Girardi, et al. Foundation models for generalist medical artificial intelligence. *Nature*, 624(7992):522–531, 2023.

[4] Arun James Thirunavukarasu, Daniel SW Ting, Kavya Elangovan, Laura Gutierrez, Tock Han Tan, and Chee-Kiong Tan. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.

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

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

[7] Jiayuan Zhu Abdullah Hamdi Yunli Qi Yueming Jin Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

[8] Nikhila Ravi Valentin Gabeur Yuan-Ting Hu Ronghang Hu Chaitanya Ryali Tengyu Ma Haitham Khedr Roman Rädle Chloe Rolland Laura Gustafson Eric Mintun Junting Pan Kalyan Vasudev Alwala Nicolas Carion Chao-Yuan Wu Ross Girshick Piotr Dollár Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[9] Saleema Amershi, Maya Cakmak, W Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. In *AI Magazine*, volume 35, 4, pages 105–120, 2014.

[10] Taha Koleilat Hojat Asgariandehkordi Hassan Rivaz Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *arXiv preprint arXiv:2409.19483v3*, 2024.

[11] Luo Xiangde, Liao Wenjun, Xiao Jianghong, Chen Jieneng, Song Tao, Zhang Xiaofan, Li Kang, Metaxas Dimitris N, Wang Guotai, and Zhang Shaoting. Word: A large scale dataset benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022.

[12] Bin Wang Armstrong Aboah Zheyuan Zhang Ulas Bagci. Gazesam: What you see is what you segment. *arXiv preprint arXiv:2304.13844*, 2023.

[13] Leila Khaertdinova Ilya Pershin Tatiana Shmykova Bulat Ibragimov. Gaze-assisted medical image segmentation. *arXiv preprint arXiv:2410.17920*, 2024.

[14] Khaertdinova Leila, Shmykova Tatyana, Pershin Ilya, Laryukov Andrey, Khanov Albert, Zidikhanov Damir, and Ibragimov Bulat. Gaze assistance for efficient segmentation correction of medical images. *IEEE Access*, 2025.

[15] Schenk Andrea, Prause Guido, and Peitgen Heinz-Otto. Efficient semiautomatic segmentation of 3d objects in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2000.

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 and introduction clearly state the main contributions: a novel framework for zero-shot 3D interactive segmentation using eye gaze as a time-series input to guide foundation models. The claims on efficiency and performance trade-offs are directly supported by the results in Section 4.3.

   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: The "Conclusion and Future Work" section explicitly discusses limitations, including the aggregation of gaze into a static heatmap (instead of direct time-series modeling) and the need for validation with practicing clinicians on more diverse datasets.

   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: This paper is empirical and does not present theoretical results that would require mathematical proofs.

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 4.1 describes the hardware (Tobii Eye Tracker 4C, LG monitor), public dataset (WORD), and methodology for creating synthetic data. Section 3 details the methods for heatmap generation and prompting, providing a clear path for reproduction.

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: [No]

   Justification: Due to patient data privacy constraints associated with the real gaze data, and institutional policies, the code and collected data are not publicly released at this time. However, the methods are described in detail to allow for reimplementation.

   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: The paper does not involve model training (it is a zero-shot approach). The experimental setting, including the dataset, number of slices used for evaluation, and hardware setup, is detailed in Section 4.1.

   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: Tables 1, 2, and 3 report mean results along with standard deviations for Dice scores and segmentation times, indicating the variability of the results.

   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: [No]

   Justification: While segmentation times are provided, specific details on the GPU/CPU models and memory are not listed. However, the methods are not computationally intensive and can be reproduced on standard modern workstations.

   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: The research was conducted in line with ethical guidelines. The work involving a human participant was performed under institutional review with their informed consent.

   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 paper focuses on the positive impact of making clinical annotation workflows more efficient. The "Conclusion and Future Work" section implicitly addresses potential negative impacts by highlighting the need for extensive clinical validation before real-world deployment.

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: The paper does not release any new models or datasets that pose a high risk for misuse.

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 external assets used (e.g., SAM-2, MedSAM-2, WORD dataset) are properly cited in the references. The WORD dataset is publicly available, and its use conforms to its terms.

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: [NA]

    Justification: The paper does not introduce or release any new public assets.

    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: [Yes]

    Justification: The study involved a single, trained expert participant (a proxy radiologist), not crowdsourcing. Details about the human experiments are provided in Section 4.

    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: [Yes]

Justification: The study was conducted following institutional ethical guidelines, which include review and approval for research with human subjects. The participant was informed of the study's nature, and there were no significant risks involved.

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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used as a core component of our research methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.