

# autoPET IV challenge: Incorporating organ supervision and human guidance for lesion segmentation in PET/CT

Junwei Huang<sup>1</sup>, Yingqi Hao<sup>1</sup>, Yitong Luo<sup>1</sup>, Ziyu Wang<sup>1</sup>, Mingxuan Liu<sup>1</sup>, Yifei Chen<sup>1</sup>, Yuanhan Wang<sup>1</sup>, Lei Xiang<sup>2</sup>, and Qiyuan Tian<sup>1\*</sup>

<sup>1</sup> School of Biomedical Engineering, Tsinghua University, Beijing, China

<sup>2</sup> Subtle Medical Inc., Shanghai, China

**Abstract.** Lesion Segmentation in PET/CT scans is an essential part of modern oncological workflows. To address the challenges of time-intensive manual annotation and high inter-observer variability, the autoPET challenge series seeks to advance automated segmentation methods in complex multi-tracer and multi-center settings. Building on this foundation, autoPET IV introduces a human-in-the-loop scenario to efficiently utilize interactive human guidance in segmentation tasks. In this work, we incorporated tracer classification, organ supervision and simulated clicks guidance into the nnUNet Residual Encoder framework, forming an integrated pipeline that demonstrates robust performance in a fully automated (zero-guidance) context and efficiently leverages iterative interactions to progressively enhance segmentation accuracy. Our source code is available at <https://github.com/huang-jw22/autoPET-4-submission>.

**Keywords:** PET/CT lesion segmentation · Interactive Segmentation · autoPET 2025 · nnUNet

## 1 Introduction

Positron Emission Tomography combined with Computed Tomography imaging (PET/CT) supplies combined metabolic and anatomical information, acting as a powerful imaging modality for clinical diagnosis and treatment planning [2,11]. However, manual lesion segmentation in PET/CT scans remains a time-consuming process susceptible to inter-observer variability, which motivates the development of automated methods to improve both efficiency and reproducibility. Deep learning has emerged as a powerful paradigm for automated segmentation of medical images, and its feasibility for handling diverse multi-tracer and multi-center PET/CT data has been robustly demonstrated in preceding autoPET challenges [3].

The autoPET IV challenge extends this pursuit by investigating the role of interactive human guidance in the segmentation process. The challenge utilizes

a large-scale dataset inherited from previous iterations, comprising 1014 FDG-tracer and 597 PSMA-tracer scans [4,7]. The focus this year is a human-in-the-loop segmentation scenario, where incremental human guidance is provided in the form of foreground and background clicks to simulate a clinical workflow where an automated tool is progressively refined by an expert user.

Our method builds upon the well-established nnU-Net framework [5,6], which has consistently demonstrated high accuracy and strong generalization ability across numerous medical segmentation tasks, including prior autoPET competitions [3,9,10]. We took valuable experiences from previous top-performing teams, forming an integrated pipeline for this task. Various training strategies were investigated to teach the model progressively improve with incremental guidance while maintaining robustness in zero/few-guidance settings.

## 2 Methods

### 2.1 Data

**Training data.** The autoPET IV dataset was used for both training and validation, including 1,014 FDG cases and 597 PSMA cases [2,11]. Specifically, the FDG dataset comprises 501 patients diagnosed with histologically proven malignant melanoma, lymphoma, or lung cancer, along with 513 negative control patients. The PSMA dataset includes pre- and/or post-therapeutic PET/CT images of male individuals with prostate carcinoma, encompassing images with (537) and without PSMA-avid tumor lesions (60).

In autoPET IV, the role of human interaction is emphasized. For each case, 10 Foreground clicks and 10 Background clicks are simulated and modeled as 3D Gaussian maps using the provided script, then concatenated with the original PET/CT data to form a 4-channel input.

**Validation data.** 5-fold cross-validation was performed for model development and evaluation. An 80/20 train-val split was applied on the autoPET dataset to create the training data and validation data for 5 folds.

### 2.2 Data pre-processing

Experiment planning and data pre-processing were conducted using the default planner and preprocessor of nnUNetv2 [5]. CT images were normalized with the "CT Normalization" scheme that conducts percentile clipping before normalization, while ZScore Normalization was applied to PET images and two clicks channel. Default data augmentation of nnUNetv2 includes Gaussian noise, random rotation, cropping, Gaussian blur, down-sampling, and gamma correction.

### 2.3 Algorithm/model

Our segmentation pipeline is built upon the robust nnUNet Residual Encoder framework [6]. To tailor this powerful baseline for the specific challenges, we

integrated several key components, including an upstream tracer classification module, a dual-headed segmentation architecture with organ supervision, and a post-processing method based on PET SUV thresholding.

**Tracer Classification.** FDG and PSMA tracers exhibit fundamentally different biodistribution patterns, which presents a significant challenge for a unified segmentation model. Training models on separate tracer datasets showed prospects in enhancing model’s performance on corresponding tracer images. To achieve efficient and accurate tracer classification, We adopted the public weights from the autoPET 2024 runner-up team [8]. Their model incorporates two separate ResNet trained on coronal and sagittal Maximum Intensity Projections, followed by a multilayer perceptron (MLP) that receives the concatenated features from the frozen backbones of both models, and outputs a binary prediction of the tracer type. We verified that this pre-trained model achieved 100% classification accuracy across all training cases.

**Incorporating Human Guidance into Training Data.** In autoPET IV challenge, each methods are evaluated in two aspects: The ability to efficiently utilize incremental human guidance information, and the ability to perform optimally in densely-guided scenarios. The way of incorporating guidance information into the training data largely affects the model’s robustness and generalization ability across different amounts of guiding clicks. We mainly explored two ways of incorporating interactive information into the training process:

- **Full-Guidance Training.** This initial strategy aims to train a model that quickly adapts to human guidance and achieve maximum segmentation accuracy when provided with dense user interactions. In this approach, we concatenate all 10 foreground/background clicks (modelled as 3D Gaussian heatmaps) together with the PET/CT images to form a 4-channel input for each case (Channel 0: CT; Channel 1: PET; Channel 2: FG Clicks; Channel 3: BG Clicks).
- **Stochastic Click Sampling.** A model trained exclusively on dense guidance often fails to establish a strong zero-guidance baseline and struggles with sparse inputs, which is detrimental to the AUC-based evaluation metrics. To address this, we implemented a stochastic click sampling strategy. For each training sample loaded, we randomly sampled an integer  $k$  (from 0 to 10) according to a predefined probability distribution. The network was then provided with guidance maps generated from only the first  $k$  foreground and  $k$  background clicks. This approach exposes the model to the full spectrum of interactive scenarios, forcing it to learn a robust and flexible response to varying levels of user input.

**PSMA-specific Model Development.** Our lesion segmentation models are primarily built upon the nnUNetv2 framework, utilizing the most recent Resid-

ual Encoder UNet architecture (ResEnc-M/L) [6]. [8] reported significant improvement in PSMA segmentation accuracy with models trained separately on the PSMA dataset. We tested various training strategies and selected 3 most promising models for final 5-fold cross validation.

- **Model V0: Weighted loss function with Dense Guidance.** The specificity and high sensitivity of PSMA in prostate carcinoma cases contributes to the appearance of numerous small and sparsely-distributed metastatic lesions [1], which are challenging for standard segmentation model. To guide the model place more emphasis on these small foreground lesions, we adjusted the weight of DiceCE loss from equal weights to Dice:CE = 2:1. Besides, the smoothing term is omitted in the Dice loss calculation to make training more stable, as suggested by [10]. The model adopted the ResEnc-M architecture and was trained on the full-guidance dataset (10 clicks) for 1000 epochs with a patch size of [192, 192, 192] and a batch size of 3.
- **Model V1: Second-Stage Fine-Tuning for Interactive Performance.** The dense-guidance training would potentially result in a specialist model that is highly dependent on user input and underperform in zero- or few-click scenarios. To enhance robustness with sparse guidance, a second stage fine-tuning of model V0 was conducted on the same training data but using stochastic click sampling. The distribution was heavily skewed towards scenarios with minimal guidance (e.g., 40% probability for 0 clicks, 20% for 1 click) to specifically improve the model’s baseline and sparse-guidance performance. This stage was run for 250 epochs with a reduced initial learning rate of  $2e-4$  to ensure stable adaptation without catastrophic forgetting.
- **Model V2: Fine-tuning pre-trained model with Balanced Stochastic Sampling.** A strong anatomical background has been proved helpful for enhancing lesion segmentation accuracy and mitigating false positive segmentation on organs with high physiological uptake [10,8]. To provide anatomical prior knowledge, we utilized the publicly available weights from [10], which had been trained on a diverse, multi-modal medical imaging dataset. The single-channel pre-trained weights were expanded to our 4-channel input by duplicating them for the CT and PET channels and initializing the two click-guidance channels to zero. We then fine-tuned this model on the PSMA dataset for 1000 epochs using stochastic click sampling. The click sampling distribution for this model was designed to be more balanced, with significant weight on both zero-click (10%) and dense-click (30%) scenarios. This strategy was chosen to leverage the strong baseline prior from the pre-trained weights while progressively training the model to respond to dense interactive guidance.

**Unified Multi-Tracer Model with Organ Supervision.** For FDG dataset, however, the improvement of separate training is not significant. This conclusion is drawn both from [8] and our empirical experiments. Besides, given that all

FDG training data were acquired from UKT while all PSMA training data were from LMU, we hypothesized that training a unified model on the combined dataset would force it to learn more robust, center-invariant features, thereby improving its generalization potential.

To better cope with multi-tracer and multi-center data, Organ Supervision was implemented to guide the model form a strong anatomical understanding which is invariant of tracer types and center differences. We utilized the "autoPET3" Trainer class proposed by [10], which introduced an auxiliary organ segmentation head focusing on a set of 10 key organs (including spleen, kidneys, liver, urinary bladder, lung, brain, heart, stomach, prostate, and glands in the head region). Pseudo organ labels were created using TotalSegmentator [12]. Equal loss weighting was applied on both segmentation head. For final unified training, we adopted the same methods (Organ Supervision + Pre-trained model), and investigated the performance with different amount of guidance incorporated in the training data:

- **Model V3: Full-Guidance Training.** The integration of anatomical prior and organ supervision improved our confidence in the model’s ability to handle few-clicks situations, so we first evaluated the model’s performance when trained with maximal human guidance (all 10 clicks). To ensure consistency with pre-trained weights, ResEnc-L architecture was adopted with patch size [192, 192, 192] and batch size 2. Training was conducted for 1000 epochs with initial learning rate  $1e-3$ .
- **Model V4: Stochastic Click Sampling.** The second experiment aimed to create a more balanced model optimized for progressive human interaction. It followed the identical architecture and hyperparameters, but was trained using our stochastic click sampling strategy. The weights distribution for generating different numbers of clicks was [0.10, 0.10, 0.10, 0.08, 0.04, 0.04, 0.04, 0.04, 0.08, 0.08, 0.30] (corresponding to 0-10 clicks), emphasizing dense-guidance scenarios while maintaining the prior ability of the pre-trained model under 0/few-click circumstances.

## 2.4 Data post-processing

For data post-processing, we adopted the thresholding methods from [8] to reduce false positive volumes, by applying a tracer-specific SUV threshold to remove segmentation masks with PET values below the threshold. The thresholds were set to 1.5 for FDG cases and 1 for PSMA cases.

## 2.5 Training and test parameters

All models were trained using the nnUNetv2 framework. 3 PSMA models were trained on NVIDIA A100 GPUs, while the 2 unified models were trained on NVIDIA A800 GPUs. Specific training parameters such as patch size, batch size, epoch counts and initial learning rate for each experiment are detailed in

their respective sections above. Any parameters not explicitly mentioned were kept to the nnUNetv2 default settings.

For inference, we enable test-time augmentation (TTA) to enhance prediction robustness. To comply with the challenge’s time constraints, we estimate the time required for predicting the original image, and compute the number of mirrored axes allowed within a 40 seconds limit for each fold. The final segmentation is produced by averaging the softmax probabilities from multiple augmented predictions.

## 2.6 Github repository

Our source code and trained weights are available at <https://github.com/huang-jw22/autoPET-4-submission/tree/master>.

# 3 Results

## 3.1 PSMA lesion segmentation model

5-folds training has been performed for all 3 methods described in Section 2.3. Average cross validation results are presented in Tab 1. It is important to note that these metrics are not directly comparable, as each model was trained on a dataset with a different click-guidance distribution. Models trained with stochastic sampling (V1, V2) were evaluated on validation sets that also contained varied click counts, which naturally results in lower average performance compared to the dense-guidance model (V0) evaluated on a dense-guidance validation set.

**Table 1.** Averaged 5-fold cross-validation results for the PSMA-specific models.

Version	Training Data	DICE	FPV	FNV
<b>Model V0</b>	Full Guidance	0.75	5.86	10.87
<b>Model V1</b>	Stochastic Sampling	0.66	11.30	13.12
<b>Model V2</b>	Stochastic Sampling	0.72	8.11	12.76

To have a fair comparison of the 3 models and evaluate interactive abilities, we randomly chose 100 PSMA cases and generated 11 different inputs (with 0-10 clicks) for each of them. Interactive evaluation was then conducted by giving 11 predictions for all cases using the 3 models.

Results of the evaluation is presented in Tab 2. It must be noted that the predictions were made by the 5-folds ensemble and all cases had already been seen by the models during training, so no guarantee of final test performance can be made. However, some valuable conclusions can still be yielded:

- Model V0 showed terrible performance on 0/few-clicks circumstances. This confirms that training exclusively on dense guidance makes the model too dependent on user interaction and fails in sparsely-guided context.
- Despite high Dice score, Model V1 showed significantly higher False Positive Volume across all click counts. This observation suggests that a sudden absence of interactive information might cause the model to develop an overly aggressive prediction strategy to compromise the lack of external guidance.
- Model V2 demonstrated most stable performance and a significant correlation between number of clicks and segmentation quality. An increasing Dice Score and decreasing FPV/FNV values indicates that the model efficiently utilized human guidance. The baseline performance with no guidance provided is also acceptable.

**Table 2.** Average evaluation results of PSMA models across 0-10 clicks interaction.

Clicks	DICE			FPV			FNV		
	V0	V1	V2	V0	V1	V2	V0	V1	V2
0	0.000	0.708	0.619	0.000	13.927	0.991	190.238	4.033	5.641
1	0.599	0.808	0.775	0.637	13.814	0.378	27.328	4.057	4.000
2	0.723	0.832	0.816	0.687	13.382	0.377	11.061	3.533	2.801
3	0.781	0.845	0.837	0.703	13.347	0.367	5.503	3.263	2.290
4	0.810	0.847	0.849	0.707	13.261	0.365	4.935	3.139	2.190
5	0.823	0.851	0.855	0.668	13.146	0.365	3.752	2.876	2.119
6	0.839	0.855	0.860	0.703	13.159	0.364	3.205	2.729	1.987
7	0.848	0.857	0.865	0.677	13.138	0.361	2.724	2.571	1.956
8	0.854	0.859	0.866	0.667	12.948	0.352	2.399	2.390	1.831
9	0.860	0.861	0.869	0.640	12.886	<b>0.330</b>	2.038	2.361	1.782
10	0.864	0.862	<b>0.871</b>	0.635	12.868	0.332	1.823	2.202	<b>1.613</b>

### 3.2 Unified lesion segmentation model

Due to computational constraints, the 5-fold cross-validation for these models was partially completed at the time of this analysis, with results available from three folds for Model V3 (full guidance) and two folds for Model V4 (stochastic sampling). Similar interactive evaluation was conducted on 250 FDG cases. Results of the evaluation is presented in Tab 3.

Both methods exhibited a stable 0-click performance and progressive performance gains with incremental user guidance. The results confirmed our assumptions that anatomical prior and organ supervision guarantees the model’s

generalization ability across different interactive levels, even if all training data was provided with full guidance (model V3).

A closer analysis reveals a performance trade-off between the two models: Model V3 presents better performance under dense guidance scenarios, while Model V4 outperforms with sparse clicks inputs. To better leverage their individual specialty, we decided to adopt a hybrid strategy during inference, by flexibly selecting model version based on the number of input guiding clicks. Model V4 would be utilized in early interactive phase (0-4 clicks), while model V3 would be chosen in densely guided steps (5-10 clicks). This approach ensures that the optimal model is deployed to achieve better precision in different interactive scenarios.

**Table 3.** Average evaluation results of unified models across 0-10 clicks interaction.

Clicks	DICE		FPV		FNV	
	V3	V4	V3	V4	V3	V4
0	0.739	0.788	0.694	0.651	8.869	5.958
1	0.811	0.837	0.647	0.500	7.040	5.631
2	0.844	0.853	0.660	0.469	5.728	4.420
3	0.861	0.862	0.700	0.459	4.459	3.874
4	0.869	0.865	0.678	0.424	3.805	3.310
5	0.875	0.870	0.684	0.422	2.798	2.537
6	0.880	0.873	0.688	<b>0.418</b>	2.630	2.249
7	0.883	0.873	0.703	0.435	2.411	2.090
8	0.886	0.875	0.714	0.446	2.065	1.934
9	0.888	0.876	0.721	0.450	1.805	1.848
10	<b>0.889</b>	0.877	0.725	0.434	<b>1.589</b>	1.658

## 4 Final Submission

For final submission, we followed the described pipeline, integrating a tracer classifier, separate models for different tracer types, with SUV thresholding as final post-processing. For PSMA cases, we submitted Model V2 (fine-tuning pre-trained model with balanced stochastic sampling), which exhibited best interactive performance during evaluation; for FDG cases, we adopted the aforementioned hybrid strategy, flexibly utilizing two different models based on different scenarios. Training parameters and algorithm details have been described above and also summarized in Tab. 4.



## 5 Conclusion

In this work, we developed and validated a comprehensive pipeline for interactive lesion segmentation in multi-tracer and multi-center PET/CT data. The proposed pipeline integrates several key methodologies, including tracer classification, unified & tracer-specific model training, organ supervision and PET SUV thresholding. We also explored various ways of incorporating human guidance into training data. Our experiments yielded two critical insights. First, we showed that a carefully designed stochastic sampling curriculum can effectively enhance the model’s generalization ability across different guidance level. Second, we demonstrated the role of introducing anatomical prior and organ supervision in securing a strong non-guiding baseline precision while reaching a high-performance ceiling with dense interactions. Our pipeline exhibits robust performance with sparse human guidance and a proficient ability to leverage incremental human feedback to refine segmentation accuracy.

**Table 4.** Algorithm details

<b>Team name</b>	<b>algorithm name</b>	<b>data pre-processing</b>	<b>data post-processing</b>	<b>training data augmentation</b>
BIRTH	BIRTH submission	nnUNet Preprocessor	PET SUV Thresholding	nnUNet DA scheme (Gaussian noise, random rotation, cropping, Gaussian blur, down-sampling, and gamma correction)
<b>test time augmentation</b>	<b>ensembling</b>	<b>standardized framework</b>	<b>network architecture</b>	<b>loss</b>
Mirroring	multi-folds cross-validation	nnUNetv2 (3D)	ResEnc UNet (3D)	DSC + CE
<b>training data</b>	<b>data/model dimensionality and size</b>	<b>use of pre-trained models</b>	<b>GPU hardware for training</b>	
1014 FDG + 597 PSMA PET-CT of autoPET	3D: 192x192x192	Public available pre-trained models	5x Nvidia A100 1x Nvidia A800	

**Acknowledgments.** This work was supported by Tsinghua University Initiative Scientific Research Program (Student Academic Research Advancement Program).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bubendorf, L., Schöpfer, A., Wagner, U., Sauter, G., Moch, H., Willi, N., Gasser, T.C., Mihatsch, M.J.: Metastatic patterns of prostate cancer: an autopsy study of 1,589 patients. *Human pathology* **31**(5), 578–583 (2000)
2. Farwell, M.D., Pryma, D.A., Mankoff, D.A.: Pet/ct imaging in cancer: Current applications and future directions. *Cancer* **120**(22), 3433–3445 (November 2014). <https://doi.org/10.1002/cncr.28860>, epub 2014 Jun 19
3. Gatidis, S., Früh, M., Fabritius, M.P., Gu, S., Nikolaou, K., Fougère, C.L., Ye, J., He, J., Peng, Y., Bi, L., et al.: Results from the autpet challenge on fully automated lesion segmentation in oncologic pet/ct imaging. *Nature Machine Intelligence* **6**(11), 1396–1405 (2024)
4. Gatidis, S., Kuestner, T.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-ct-lesions) [dataset]. <https://doi.org/10.7937/gkr0-xv29> (2022). <https://doi.org/10.7937/gkr0-xv29>, the Cancer Imaging Archive
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
6. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *arXiv preprint arXiv:2404.09556* (2024)
7. Jeblick, K., et al.: A whole-body psma-pet/ct dataset with manually annotated tumor lesions (psma-pet-ct-lesions) (version 1) [dataset]. <https://doi.org/10.7937/r7ep-3x37> (2024). <https://doi.org/10.7937/r7ep-3x37>, the Cancer Imaging Archive
8. Kalisch, H., Hörst, F., Herrmann, K., Kleesiek, J., Seibold, C.: Autopet iii challenge: Incorporating anatomical knowledge into nnunet for lesion segmentation in pet/ct (2024), <https://arxiv.org/abs/2409.12155>
9. Murugesan, G.K., McCrumb, D., Brunner, E., Kumar, J., Soni, R., Grigorash, V., Moore, S., Oss, J.V.: Improving lesion segmentation in fdg-18 whole-body pet/ct scans using multilabel approach: Autopet ii challenge (2023), <https://arxiv.org/abs/2311.01574>
10. Rokuss, M., Kovacs, B., Kirchhoff, Y., Xiao, S., Ulrich, C., Maier-Hein, K.H., Isensee, F.: From fdg to psma: A hitchhiker’s guide to multitracer, multicenter lesion segmentation in pet/ct imaging (2024), <https://arxiv.org/abs/2409.09478>
11. Schwenck, J., Sonanini, D., Cotton, J.M., Rammensee, H.G., la Fougère, C., Zender, L., Pichler, B.J.: Advances in pet imaging of cancer. *Nature Reviews Cancer* **23**(7), 474–490 (July 2023). <https://doi.org/10.1038/s41568-023-00576-4>, epub 2023 May 31
12. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence* **5**(5) (2023)