# SegMoTE: Token-Level Mixture of Experts for Medical Image Segmentation

## Supplementary Material

In this supplementary material, we provide a detailed description of the dataset construction process and its proportional allocation, along with a more comprehensive visual comparison of the results. The arrangement is as follows: Section 6 presents the construction of the MedSeg-HQ dataset and the filtering criteria, while Section 7 visually compares the segmentation results of different methods.

## 6. More Details of MedSeg-HQ

**Data composition of MedSeg-HQ.** Tab.5 presents detailed information about our newly constructed dataset, MedSeg-HQ. This dataset contains 154,569 high-precision image-mask annotations. MedSeg-HQ integrates 12 existing medical image datasets: ACDC, AMOS (CT), AMOS (MRI), BTCV, CHAOS (T1), CHAOS (T2), ISIC2016, ISIC2017, ISIC2018, SZ-CXR, Totalsegmentator (CT), and WORD. All these datasets provide extremely fine-grained segmentation masks, upon which we performed additional selection and refinement.

**Construction Process.** After collecting 12 public medical imaging datasets, five domain experts collaborated to design and implement a semi-automatic data construction pipeline for MedSeg-HQ. Their primary goal was to maximize both the quality and diversity of the images, ensuring that a broad range of data was included in the dataset. At the same time, they aimed to maintain consistent annotation reliability across various imaging modalities, such as CT, MRI, and X-ray, to ensure the robustness of the dataset. During their analysis, the experts observed significant variations in key image characteristics, such as brightness, contrast, and mask precision, across different modalities. If these images were simply merged without addressing these differences, it could result in uneven global quality, potentially affecting the performance of downstream models. To mitigate this risk and ensure high-quality data, they established a unified quantitative quality control (QC) protocol specifically designed to filter out low-quality samples. This protocol integrates clinical experience with standard image quality metrics, and the QC criteria were defined with four hard constraints to ensure that only the most reliable data was included:

$$\mu \in [5, 245] \tag{12}$$

$$\sigma \geq 6.0 \tag{13}$$

$$0.02 \leq R_{\mathrm{FG}} \leq 0.95 \tag{14}$$

$$C_{\mathrm{N}} \leq 300 \tag{15}$$

Table 5. Data composition of our constructed MedSeg-HQ.

| Dataset | Mask Num. |
|---|---|
| ACDC [1] | 3596 |
| AMOS (CT) [13] | 38598 |
| AMOS (MRI) [13] | 7110 |
| BTCV [22] | 4228 |
| CHAOS (T1) [15, 16] | 1942 |
| CHAOS (T2) [15, 16] | 950 |
| ISIS2016 [8] | 2526 |
| ISIC2017 [5] | 4274 |
| ISIC2018 [4] | 5302 |
| SZ-CXR [40] | 1132 |
| Totalsegmentator (CT) [45] | 49545 |
| WORD [25] | 35366 |
| **Total** | **154569** |

where $\mu$ represents the mean intensity of the image. This constraint ensures that the image intensity is within a reasonable range to avoid images that are too bright or too dark, which could result in poor contrast and reduced quality. $\sigma$ is the standard deviation, which measures the image's contrast. A minimum value of 6.0 ensures that the image has adequate contrast to make features distinguishable. Low contrast images might be rejected because they could make segmentation and analysis harder. $R_{\mathrm{FG}}$ is the foreground ratio of the mask, representing the proportion of the image that is occupied by the foreground object. The lower bound of 0.02 ensures that the mask captures at least some portion of the foreground, while the upper bound of 0.95 prevents the mask from including excessive background, which could result in incorrect segmentation. $C_{\mathrm{N}}$ represents the number of connected components in the mask. A value greater than 300 indicates that the mask has too many fragmented components (e.g., due to noise or segmentation errors), which would likely result in unreliable data. Any sample violating these thresholds was automatically rejected and logged.

Beyond the hard gates, a continuous scoring mechanism was proposed to further differentiate between valid samples of varying quality. This scoring system integrates four key image-domain metrics: sharpness (measured as the variance of the Laplacian), contrast (the global pixel standard deviation), entropy (the histogram information entropy), and boundary gradient (the mean Sobel gradient magnitude along the dilated–eroded mask boundary). These metrics together capture important aspects of image quality, such

as clarity, contrast, and boundary definition. Each metric is normalized using a smooth $\tanh(\cdot)$ transformation with empirically determined scales of 200 for sharpness, 40 for contrast, 6 for entropy, and 20 for boundary gradient. The final quality score is then computed as a weighted linear combination of these metrics, allowing the system to rank samples based on their overall quality:

$$\begin{aligned}\text{Score} = 0.30 \cdot \tanh\left(\frac{S}{200}\right) + 0.30 \cdot \tanh\left(\frac{C}{40}\right) + \\ 0.20 \cdot \tanh\left(\frac{H}{6}\right) + 0.20 \cdot \tanh\left(\frac{G}{20}\right)\end{aligned} \tag{16}$$

where $S$ represents sharpness, $C$ represents contrast, $H$ represents entropy, and $G$ represents the boundary gradient. Higher scores indicate sharper, more structured, and better-contrasted images with reliable boundary delineations.

To address the significant data imbalance across different datasets, organs, and scanning planes, the experts proposed a more refined sampling strategy. This approach combines proportional allocation, temperature-weighted sampling, and stratification based on quantiles. Initially, quotas are allocated to each class in proportion to the number of available samples, ensuring balance across classes. The quota $q_i$ for each class $i$ is calculated as follows:

$$q_i = \frac{N_i}{\sum_{j=1}^{K} N_j} \cdot T \tag{17}$$

where $N_i$ is the number of available samples in class $i$, $\sum_{j=1}^{K} N_j$ is the total number of samples across all classes, and $T$ is the target total sample size.

Then, within each class, samples are drawn based on the quality score using temperature-weighted probabilities. The probability $p_i$ of selecting the $i$-th sample is given by:

$$p_i \propto \text{Score}_i^{1/\tau} \tag{18}$$

where $\tau = 0.85$ is the temperature factor. A smaller $\tau$ biases the selection toward higher-quality samples, while a larger $\tau$ results in more uniform sampling.

The slices are divided into five quantile buckets according to slice index, and weighted sampling is performed independently within each bucket. The buckets are defined as:

$$\text{Bucket}_k = \{i_j \mid q_{k-1} \leq i_j < q_k\} \tag{19}$$

where $q_k$ represents the $k$-th quantile of the slice indices. After performing weighted sampling within each bucket, proportional allocation is carried out across buckets. Any quota deficit is filled by global re-sampling to reach the exact target count. The global re-sampling probability is computed as:

$$p_{\text{global}} \propto \text{Score}_i^{1/\tau} \tag{20}$$

with $\tau = 0.85$, which biases the selection toward high-quality samples while retaining diversity. Finally, slices are divided into five quantile buckets according to slice index, and weighted sampling is performed independently within each bucket, followed by cross-bucket proportional allocation. Any quota deficit is filled by global re-sampling to reach the exact target count.

# 7. More Visual Results Comparison

## 7.1. Visual Comparison of In-domain Dataset

We compare the visual segmentation results of different methods on in-domain datasets. For the segmentation task involving only foreground and background, we experiment by using points as prompts, and the results are shown in Fig.8. In contrast, for the multi-class segmentation task, we use bounding boxes as prompts, with the results presented in Fig.9. These comparisons effectively highlight the advantages of our method in terms of segmentation accuracy and scene consistency. They demonstrate that our approach exhibits outstanding robustness across a range of segmentation tasks, from simple foreground-background segmentation to more complex multi-class tasks.

## 7.2. Visual Comparison of Out-of-domain Dataset

In Fig.10, we further present the visual segmentation results of different methods on three challenging out-of-domain datasets: ISLES, SegThor, and Totalsegmentator (MRI). For the ISLES dataset, we use points as prompts, while for the other two datasets, we use bounding boxes as prompts. These results highlight the effectiveness and adaptability of our approach, demonstrating that even when facing complex out-of-domain scenarios, our method maintains stable performance across datasets and segmentation tasks.

## 7.3. Visual of interactive segmentation

We visualize the interactive process of SAM and SAM2 in foreground and background segmentation tasks, as shown in Fig.11 and 12. Both SAM and SAM2 methods require five point interactions to achieve optimal results. In contrast, our approach employs a progressive prompt tokenization mechanism, which allows for single-step generation without the need for any interactions. This enables our method to not only achieve results that are comparable to, or even better than, those produced by SAM and SAM2, but also to do so with significantly higher efficiency. Our approach thus demonstrates a clear advantage in terms of both performance and computational efficiency.

# 8. More Quantitative Comparisons

## 8.1. Ablation Study of Hyperparameter

To investigate the impact of the hyperparameter $\lambda_{\text{balance}}$ on the overall segmentation performance, we conducted an ablation study on the TotalMRI dataset. As shown in Table 6, we varied the value of $\lambda_{\text{balance}}$ from 0.005 to 0.9. The results indicate that the model achieves its peak performance of 71.48 when $\lambda_{\text{balance}}$ is set to 0.01. Increasing the weight beyond this optimal point leads to a progressive performance degradation, with a significant drop to 60.55 observed at $\lambda_{\text{balance}} = 0.9$. This suggests that an excessively large $\lambda_{\text{balance}}$ may over-regularize the network and negatively interfere with the primary segmentation objective. Conversely, a value that is too small (e.g., 0.005) fails to provide sufficient constraint, leading to slightly sub-optimal results. Therefore, we empirically set $\lambda_{\text{balance}} = 0.01$ as the default configuration for our framework.

Table 6. Ablation study on the hyperparameter $\lambda_{\text{balance}}$ evaluated on the TotalMRI dataset.

| $\lambda_{\text{balance}}$ | **0.005** | **0.01** | **0.1** | **0.5** | **0.9** |
|---|---|---|---|---|---|
| TotalMRI | 71.12 | 71.48 | 69.87 | 68.64 | 60.55 |

## 8.2. Comparison of Fine-Tuning Methods

To further validate the effectiveness of our approach, we compare our method, denoted as Ours (MedSeg-HQ), against standard fine-tuning strategies including LoRA, Adapter, and Full Parameter (Full Param.) fine-tuning. We report the segmentation performance using both click and bounding box (bbox) as interactive prompts. The results are presented in Table 7 in the format of click (bbox).

As shown in the table, parameter-efficient methods like LoRA and Adapter yield relatively lower performance across the ISLES, SegThor, and TotalMRI datasets. While Full Parameter fine-tuning achieves strong results, it typically incurs significantly higher computational costs and risks compromising the generalized representations of the foundation model. Our proposed method, Ours (MedSeg-HQ), achieves highly competitive performance that closely matches the Full Parameter baseline across all three datasets under both prompt settings. For instance, on the ISLES dataset, our method achieves 67.57 (77.30), which is comparable to the 67.81 (77.54) of Full Parameter fine-tuning. This demonstrates that our approach can effectively and efficiently adapt the foundation model to medical image segmentation tasks while maintaining high accuracy and robustness.

## 8.3. Comparison with nnU-Net

To provide a comprehensive evaluation, we compare our proposed SegMoTE against nnU-Net. Since nnU-Net is designed for automatic inference and does not support interactive prompts, we report its standard automatic segmentation results. For SegMoTE, we report the performance using click and bounding box (bbox) prompts, denoted in the format of click (bbox).

Table 7. Comparison of different fine-tuning methods across three datasets. The results are reported using click and bounding box prompts in the format of Click (Bbox).

| **Dataset** | **LoRA** | **Adapter** | **Full Param.** | **Ours (MedSeg-HQ)** |
|---|---|---|---|---|
| ISLES | 61.44 (73.14) | 63.00 (74.52) | 67.81 (77.54) | 67.57 (77.30) |
| SegThor | 60.89 (77.95) | 62.34 (78.43) | 66.24 (83.67) | 65.92 (83.39) |
| TotalMRI | 52.42 (65.30) | 53.16 (67.58) | 58.19 (71.63) | 57.97 (71.48) |

As shown in Table 8, SegMoTE guided by bounding box prompts achieves superior performance on the BTCV (83.39) and ISIC2017 (88.20) datasets compared to the fully automatic nnU-Net. On the WORD dataset, nnU-Net obtains a strong score of 80.01, slightly outperforming our bbox-prompted result of 78.42. As expected, the single-click prompt results for SegMoTE are generally lower than nnU-Net's automatic results, because a single click provides minimal contextual guidance compared to a densely trained automatic network. However, with the spatial constraints provided by a bounding box, SegMoTE demonstrates highly competitive and often superior segmentation accuracy, highlighting its effectiveness and potential for precise, interactive medical image annotation.

Table 8. Comparison with nnU-Net across three datasets. The results for SegMoTE are reported using click and bounding box prompts in the format of click (bbox), whereas nnU-Net is a fully automatic method without prompts. Red color indicates the best performance.

| **Method** | **BTCV** | **WORD** | **ISIC2017** |
|---|---|---|---|
| nnUNet | 81.92 | 80.01 | 86.32 |
| SegMoTE | 71.00 (83.39) | 60.37 (78.42) | 77.57 (88.20) |

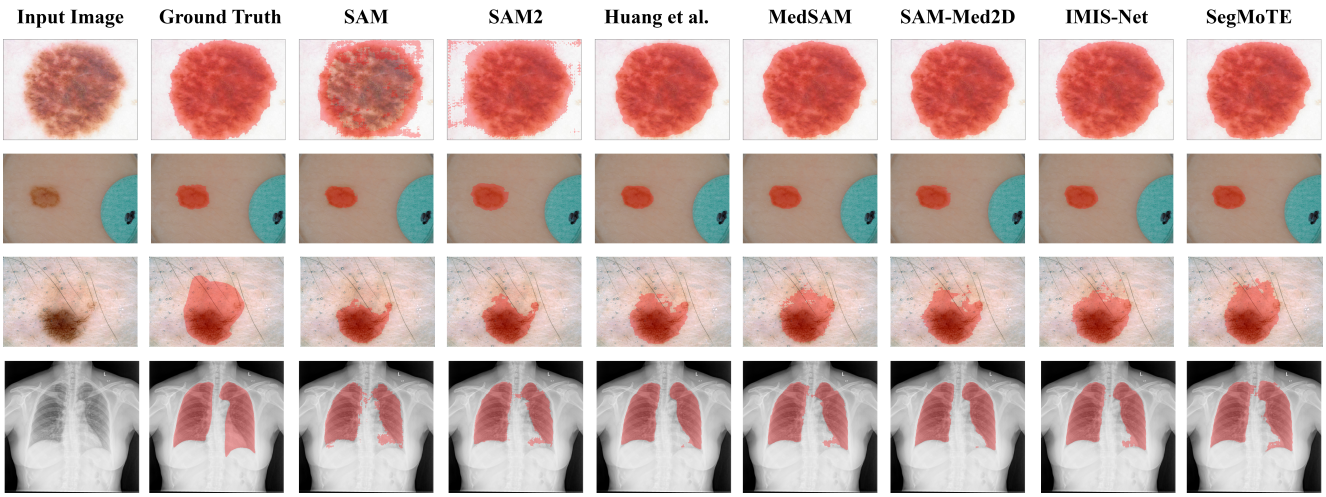


Figure 8. Visualization comparison of segmentation in binary in-domain datasets using different methods.

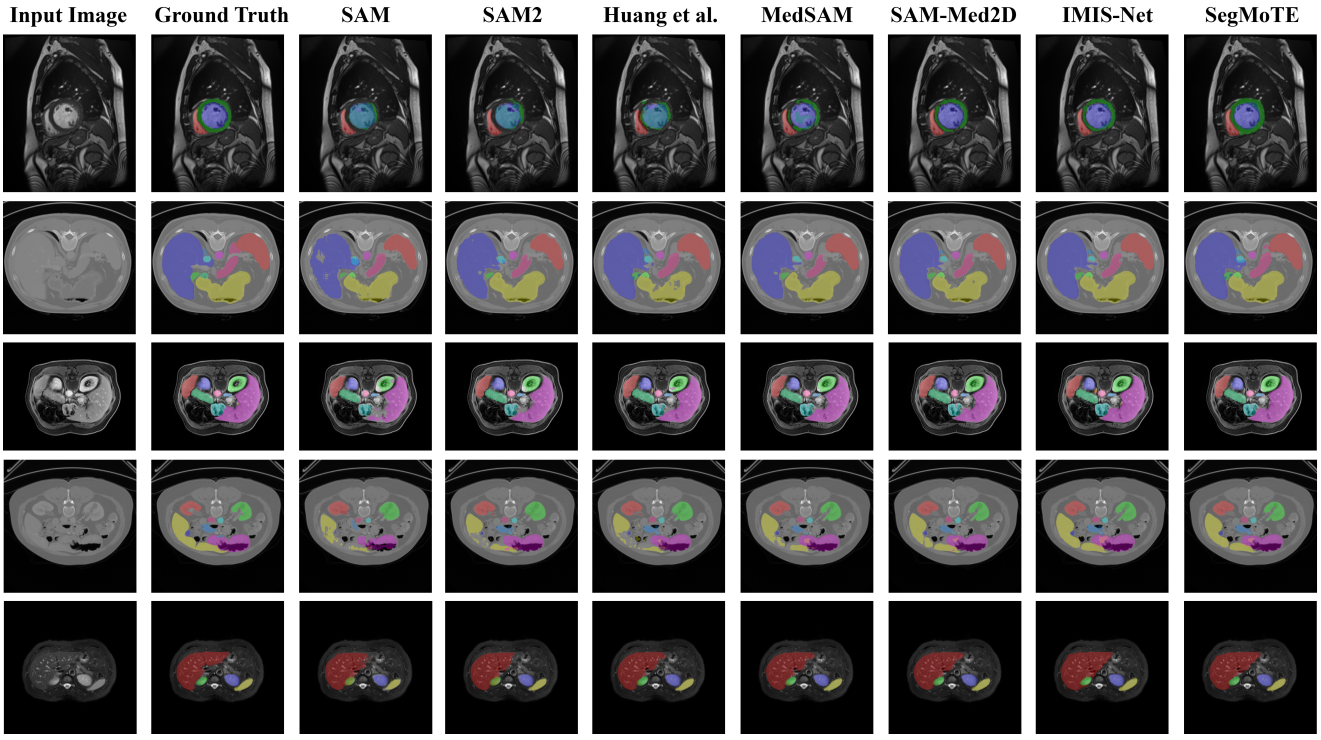


Figure 9. Visualization comparison of segmentation in multi-class in-domain datasets using different methods.

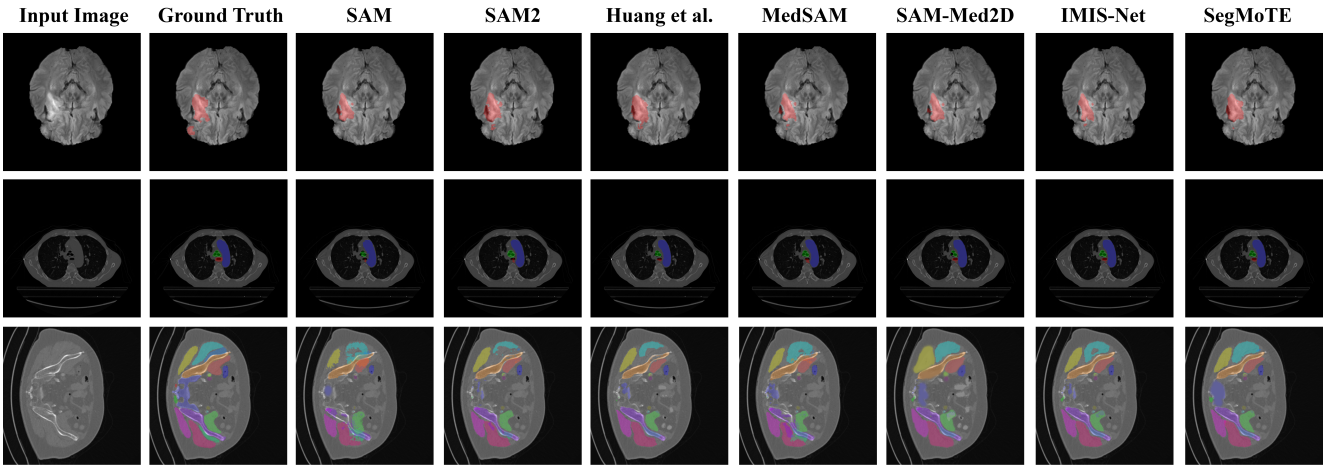


Figure 10. Visualization comparison of segmentation in out-of-domain datasets using different methods.

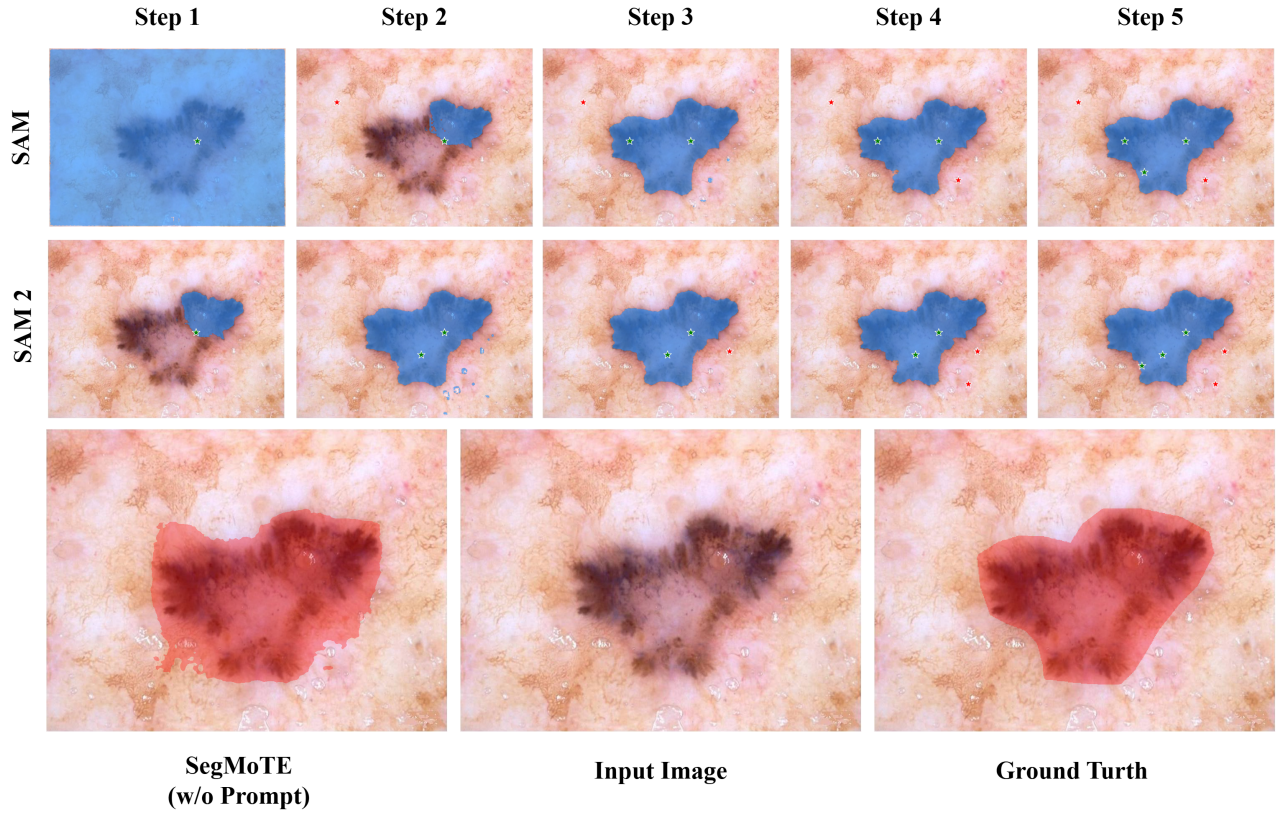


Figure 11. Segmentation results on the ISIC2016 dataset using point prompts for interaction.

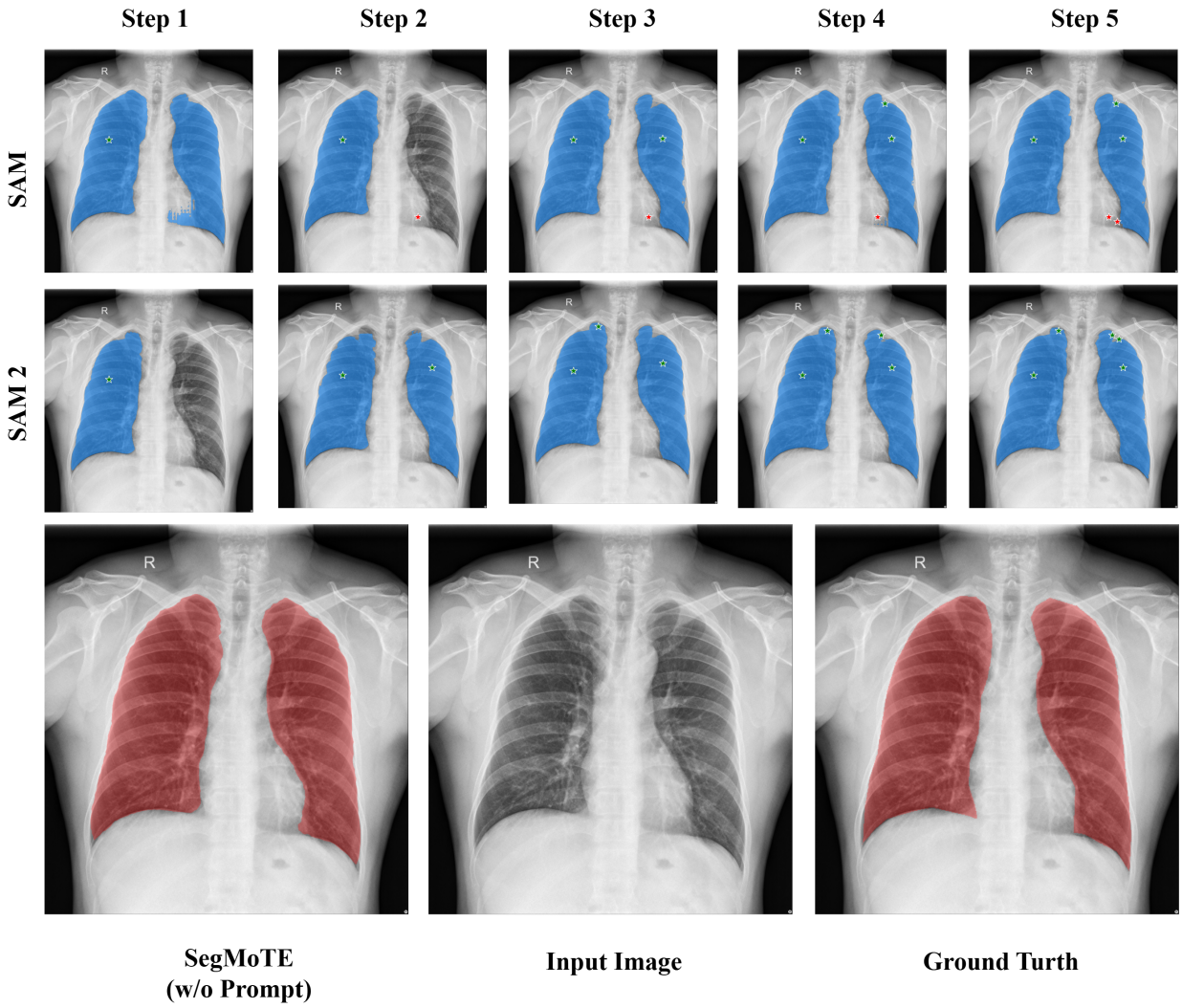


Figure 12. Segmentation results on the SZ-CXR dataset using point prompts for interaction.