npj Digital Medicine

https://doi.org/10.1038/s41746-026-02544-2

Article in Press

User-preference alignment with uncertainty-aware interactive rectification for liver organ and tumor segmentation and analysis from CT images

Received: 8 November 2025

Accepted: 3 March 2026

Cite this article as: Zhao, G., Wang, Y., Gong, C. *et al.* User-preference alignment with uncertainty-aware interactive rectification for liver organ and tumor segmentation and analysis from CT images. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02544-2

Guangyuan Zhao, Yang Wang, Chen Gong, Zipei Wang, Guobin Huang, Xuechun Zhao, Tengfei Chao & Bo Yang

User-Preference Alignment with Uncertainty-Aware Interactive Rectification for Liver Organ and Tumor Segmentation and Analysis from CT Images

Guangyuan Zhao[1,2†], Yang Wang[3†], Chen Gong[4†], Zipei Wang[1,2], Guobin Huang[1,2], Xuechun Zhao[1,2], Tengfei Chao[4*], Bo Yang[1,2*]

[1*] Institute of Organ Transplantation, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology; Key Laboratory of Organ Transplantation, Ministry of Education;NHC Key Laboratory of Organ Transplantation, Wuhan, 430030, Hubei, China .

[2*] Key Laboratory of Organ Transplantation, Chinese Academy of Medical Sciences; Organ Transplantation Clinical Medical Research Center of Hubei Province, Wuhan, 430030, Hubei, China .

[3] Department of Hepatobiliary Surgery, Peking University People's Hospital, Beijing, 100044, Beijing, China .

[4*] Department of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430030, Hubei, China .

*Corresponding author(s). E-mail(s): turnface@126.com; yangbo@tjh.tjmu.edu.cn;
Contributing authors: zhaogy0923@live.cn; richardwang@pku.edu.cn; chengong@tjh.tjmu.edu.cn; 289213153@qq.com; 847477579@qq.com; SharonX_Zhao@outlook.com;
†These authors contributed equally to this work.

Abstract

Accurate liver and tumor segmentation from CT images is essential for cancer diagnosis, treatment planning, and response assessment. However, manual segmentation is labor-intensive and variable, while standard automated models lack the flexibility to adapt to diverse clinical needs or inherent image uncertainties.

To bridge this gap, we introduce User-Preference Alignment with Uncertainty-Aware Interactive Rectification (UAIR), a novel framework designed for efficient and adaptive segmentation. Instead of requiring laborious pixel-level corrections, UAIR presents the clinician with a small, curated set of diverse segmentation candidates generated by quantifying model uncertainty. The user simply selects the most suitable option, allowing the framework to iteratively refine its results and align with specific clinical preferences. This selection-based approach drastically reduces the human interaction cost. We validated UAIR on a large-scale, multi-center CT dataset, demonstrating superior accuracy (DSC 0.776) over existing manual positional prompting (DSC 0.685) and less prompting efforts. UAIR provides a clinically-viable solution that integrates seamless human guidance, enabling rapid and robust segmentation for downstream quantitative analysis.

Keywords: Liver Tumor Segmentation, Computed Tomography, Segment Anything Model, Semi-Supervised Learning, Label-Efficient Segmentation

1 Introduction

The liver plays a central role in maintaining metabolic balance and digestive function, making it one of the most vital organs in the human body. However, its importance is gravely threatened by hepatic cancers, which have become a leading cause of cancer-related morbidity and mortality worldwide. These malignancies pose an immense challenge to public health [1]. The liver's cancer landscape is strikingly complex, not only due to primary cancers such as hepatocellular carcinoma but also because of frequent metastatic involvement from distant tumors [2]. To navigate this complexity, accurate radiological evaluation is crucial. Among various imaging techniques, multi-phase contrast-enhanced Computed Tomography (CT) stands out as the clinical gold standard. Its exceptional spatial and contrast resolution enables a thorough examination of the liver parenchyma, its blood vessels, and any focal lesions [3, 4].

However, these detailed imaging results are only as valuable as their interpretation. The task of segmenting the liver and any intrahepatic tumors is key to transforming these images into clinically useful information [5, 6]. This segmentation forms the quantitative backbone of modern cancer care. The volumetric and morphometric data derived from such segmentation are indispensable for diagnosing, predicting, and stratifying the progression of the disease [7]. One example of how this plays out in clinical practice is the Response Evaluation Criteria in Solid Tumors (RECIST), which emphasize objective measurements over subjective radiological assessments when evaluating treatment effectiveness [8]. The importance of precise segmentation extends beyond diagnosis. It is integral in planning surgical resections, defining treatment areas for ablative therapies like stereotactic radiation and radiofrequency ablation, and assessing tumor burden to guide and track locoregional interventions [9].

Despite its critical role, manual segmentation remains a significant bottleneck in both clinical practice and scientific research. This traditional approach is not only labor-intensive but also prone to considerable inter- and intra-observer variability, which undermines the reproducibility and reliability of clinical data [10]. In clinical

settings, manual segmentation of liver tumors is a time-consuming, subjective process that demands extensive expertise from radiologists. This inefficiency disrupts the clinical workflow, leading to delays in diagnosis, treatment planning, and follow-up evaluations. Furthermore, the inherent variability in manually drawn contours presents substantial challenges for quantitative analysis and reproducibility, which are essential for robust scientific investigation. In response to these limitations, automated segmentation techniques based on deep learning have shown significant promise [11, 12]. However, the adoption of these models in clinical practice has been hindered by several persistent challenges unique to liver tumor segmentation [13].

Besides, current approaches often fall short when applied to real-world clinical scenarios. One of the most significant challenges remains addressing the inherent uncertainties in medical images [14]. In clinical practice, a single image can yield multiple valid segmentation results, depending on the specific labeling criteria required for a given medical context, as medical segmentation problems are often characterised by ambiguities and multiple hypotheses may be plausible [15]. For example, in liver tumor segmentation from CT scans, it is generally preferred to include a margin of surrounding liver tissue to ensure the entire tumor is captured and avoid missing small satellite lesions. On the other hand, in cases where surgical resection is planned, undersegmentation may be favored to preserve as much healthy liver parenchyma as possible [16]. These nuances highlight the need for adaptive segmentation methodologies that can account for such uncertainties and align with varying clinical requirements of different user preferences [17].

Building on the limitations of current automated segmentation methods, particularly their inability to effectively handle uncertainty, the emergence of promptable segmentation foundation models [18] represented by the Segment Anything Model (SAM) [19], MedSAM [20], and SAM-Med3D [21] offers a promising approach for this task. These methods introduce the potential to integrate interactive segmentation through user prompts, allowing for the refinement of segmentation results according to user preferences. By enabling the clinician to correct or adjust predictions via simple interactions such as clicks or bounding boxes, these models can better align with specific clinical requirements and uncertainties [22, 23].

However, despite the potential of this approach, the need for manual interaction remains a significant drawback as shown in Figure. 1. While prompt-based segmentation provides a flexible and adaptive way to incorporate user feedback, it still requires the clinician to spend considerable time and effort providing input. This ongoing manual involvement limits the overall efficiency of the process, particularly when dealing with large datasets or in fast-paced clinical environments. Thus, while SAM-like models represent an innovative step toward improving segmentation accuracy through user engagement, the trade-off between accuracy and user effort remains a challenge. More research is needed to reduce the manual burden while maintaining the advantages of adaptive, uncertainty-aware segmentation.

To overcome these challenges, we introduce a novel user-preference alignment framework with uncertainty-aware interactive rectification (UAIR), designed to efficiently address diverse test-time preferences with minimal human intervention. As shown in Figure. 1, our framework streamlines the segmentation process by presenting

a small, curated set of distinct candidates that best capture inherent uncertainties in the data. This approach significantly reduces clinician workload, enabling rapid selection of segmentation results that align closely with specific user preferences—without the need to sift through numerous similar predictions. By incorporating real-time user feedback, UAIR refines its predictions iteratively, better adapting to the clinician's preferences and the unique demands of different clinical scenarios. This minimizes the need for time-consuming manual adjustments, allowing clinicians to guide the segmentation effectively while avoiding the substantial effort typically required by traditional methods. We validated the effectiveness of UAIR on large-scale multi-center liver tumor segmentation datasets, demonstrating superior accuracy and efficiency alongside a notable reduction in user interaction. The key contributions of our framework is organized as follows:

- **Uncertainty-Guided Multi-Candidate Selection:** Presenting a limited set of diverse segmentation candidates that capture model uncertainty, allowing users to quickly identify preferred results with minimal effort.
- **Iterative Preference Alignment:** Enabling multi-round interactive refinement whereby the framework adapts segmentation outputs based on user choices, improving accuracy through progressive feedback rather than relying solely on extensive manual corrections.
- **Reduced Human Interaction Cost:** Compared to traditional prompt-based or manual refinement methods, our selection-based approach drastically lowers the burden on clinicians by replacing pixel-level or click-based corrections with efficient candidate selection.
- **Real-Time Adaptability to Clinical Needs:** By learning from user preferences during test time, UAIR flexibly adjusts to diverse and variable clinical labeling criteria, making it broadly applicable across heterogeneous liver tumor presentations.

2 Results

2.1 Experimental Details

In this study, we use the AbdomenAtlas dataset [24] for the evaluation of our method, which a large-scale, detailed-annotated, and multi-center abdominal CT dataset sourced from 138 hospitals. This dataset provides high-quality annotations for anatomical structures including the liver, as well as preliminary annotated examples of liver tumors in a subset of its 9,262 CT volumes. These volumes were randomly split into training, validation, and testing sets with a ratio of 6:2:2, respectively. The CT volumes retain their original resolution to preserve fine-grained anatomical details, and we process the data with intensity normalization to ensure consistency for experimental evaluation. All the scans are resampled to the same resolution with intensity normalized to zero mean and unit variance. In our experiments, we meticulously designed a robust evaluation protocol to assess the performance of our models under varying conditions of data scarcity. To quantitatively evaluate the segmentation results, we use

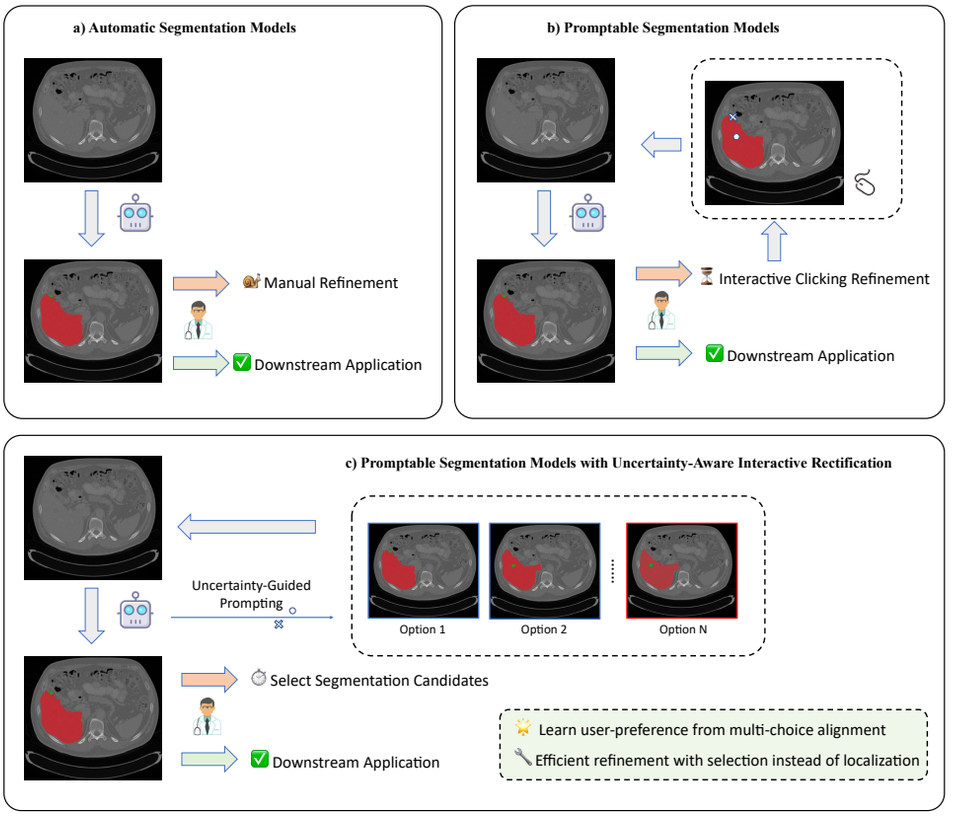

Fig. 1 Overview of our proposed framework. Comparison of different segmentation paradigms for clinical application. (a) Automatic segmentation models generate results directly, but when the outcomes are suboptimal, the model cannot correct them, requiring clinicians to perform manual refinements before downstream applications. (b) Promptable segmentation models enable interactive correction through visual prompts such as clicks, improving the flexibility of the segmentation system but still demanding considerable user effort. (c) The proposed framework presents users with a small set of diverse segmentation candidates based on uncertainty-guided prompting. Clinicians can efficiently refine results by selecting their preferred option, and if the outcome is unsatisfactory, the framework can further improve the segmentation through multiple iterative rounds. This process significantly reduces the amount of manual interaction required compared to existing promptable segmentation methods, enabling faster, more adaptive, and less labor-intensive segmentation.

Dice similarity coefficient (Dice) and Jaccard Index (Jaccard) to measure the region mismatch.

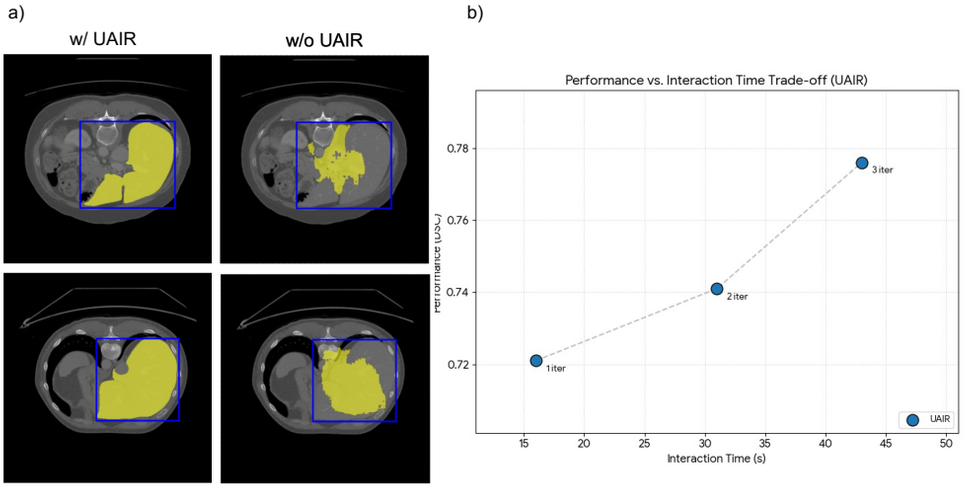

Fig. 2 Qianlilative and quantitative evaluation of the UAIR framework for interactive segmentation. (a) Qualitative comparison between segmentation results with (w/) and without (w/o) UAIR. (b) The trade-off b etween s egmentation p erformance (DSC) a nd c linician interaction time.

2.2 Automatic Segmentation with Uncertainty Estimation

In this section, we compare the performance of our proposed UAIR framework with existing segmentation methods, focusing on accuracy and the role of uncertainty estimation. We evaluate various approaches, including non-SAM training-based segmentation models [25, 26], and SAM-based manual prompting [19], auto-prompting [27], and ensemble methods [22]. The results are summarized in Table 1. Compared with SAM-based models, training-based segmentation methods require large amounts of annotated data for training and do not support interactive refinement, making manual correction still necessary when the segmentation results fail to meet application requirements. The Manual Prompting method with SAM, which relies on clinician input to guide the segmentation, yields a DSC of 0.685 and a Jaccard score of 0.521. While this method allows for human expertise, it is labor-intensive and time-consuming, limiting its practical application in large-scale or real-time clinical settings. The Auto-Prompting method improves upon manual prompting by automating some aspects of the process, achieving a DSC of 0.665 and a Jaccard index of 0.508. Despite being automated, it still falls short of the performance of more advanced models and does not incorporate uncertainty estimation to guide the clinician in selecting the most accurate segmentation result. The Ensemble method combines multiple generated prompts to improve segmentation results, yielding a DSC of 0.683 and a Jaccard of 0.518. While this method benefits f rom m odel d iversity, i t s till l acks t he adaptability offered b y i nteractive u ncertainty-aware t echniques. I n c ontrast, o ur UAIR framework, incorporating uncertainty-aware interactive rectification, demonstrates a

significant improvement in both metrics. With 1 iteration, UAIR achieves a DSC of 0.721 and a Jaccard index of 0.587, showcasing its effectiveness in providing better segmentation results with minimal clinician input. Furthermore, after 3 iterations, the performance improves further, with a DSC of 0.776 and a Jaccard of 0.616. These results highlight the power of iterative refinement and uncertainty-guided prompting, which enables the model to better align with user preferences while minimizing the need for extensive manual corrections. Overall, our framework outperforms existing methods, demonstrating the potential of uncertainty-aware frameworks in advancing automatic segmentation for clinical applications. In addition, compared with the manual positional prompting required by the original SAM, UAIR only requires the user to select from a set of generated candidates, which substantially reduces the time needed for human interaction.

The qualitative results in Fig. 2 visually underscore the robustness of our UAIR framework. Directly utilizing SAM without UAIR frequently results in fragmented masks and significant under-segmentation. In comparison, after several rounds of interaction, UAIR achieved relatively precise segmentation results. By transitioning from the precise, point-and-click manual prompting required by the original SAM to a selection-based paradigm, UAIR significantly reduce the time required for interaction refinement for segmentation. Our framework not only enhances the precision of 3D medical segmentation but also significantly reduces the cognitive and time burden on medical professionals, making it highly suitable for real-world clinical integration.

Method	DSC	Jaccard	Automatic	Interactable	Interactive Time
Conventional Segmentation Models					
U-Net	0.613	0.475	✓	-	-
TransUNet	0.654	0.493	✓	-	-
SAM-based Segmentation Models					
Manual Prompting	0.685	0.521	-	✓	64s
Auto-Prompting	0.665	0.508	✓	-	-
Ensemble	0.683	0.518	✓	-	-
UAIR (1 iter)	0.721	0.587	✓	✓	16s
UAIR (3 iter)	0.776	0.616	✓	✓	43s

Table 1 Comparison of segmentation performance metrics across different methods.

2.3 Efficient Refinement of Segmentation Results with Minimal Human Interaction Cost

One of the key advantages of the proposed UAIR framework is its ability to efficiently refine segmentation results with minimal clinician input, significantly reducing the time and effort required for manual adjustments. As shown in Figure 3, clinicians are presented with a small set of segmentation candidates, each generated with

uncertainty-aware prompting. This allows users to quickly compare and select the segmentation result that most closely aligns with their preferences. Unlike traditional manual segmentation methods, where clinicians must delineate tumor boundaries pixel by pixel, UAIR streamlines this process by offering multiple, pre-generated options that capture the inherent uncertainties of the model. In this way, clinicians are not required to perform intricate, time-consuming refinements. Instead, they can make an informed choice from a small set of candidates, thus reducing the decision-making process to just a few clicks. Furthermore, the iterative refinement process enables clinicians to provide feedback on the selected segmentation, which the model uses to adjust and generate improved candidates in subsequent rounds. As illustrated in the interface, users can see the results of each iteration and select their preferred candidate from a refined set. This multi-iteration approach allows for continuous improvement of segmentation accuracy, while minimizing the clinician's involvement after the initial feedback. This process not only enhances the efficiency of segmentation tasks but also ensures that segmentation results are adapted to the specific clinical context and user preferences. By reducing the overall interaction cost, UAIR provides a clinically viable solution for liver tumor segmentation.

2.4 Application of Segmentation in Organ and Tumor Analysis

Once the UAIR framework generates segmentation results that meet the clinician's requirements in terms of accuracy and quality, these segmented images can be leveraged for a variety of downstream tasks. The primary benefit of accurate liver tumor segmentation is its ability to provide essential quantitative data that can guide diagnosis, treatment planning, and monitoring of therapeutic efficacy.

As exemplified by the analysis dashboard in Fig. 4, the system translates raw segmentation outputs into an integrated radiological analysis dashboard. This interface combines interactive 3D Visualization with an automatically generated quantitative metrics, demonstrating a key downstream application. Segmentation results provide a detailed delineation of liver tumors, which can be used to calculate organ and tumor volume, which a critical factor in diagnosing tumor stage, assessing growth, and planning treatment. By calculating the exact volume of a tumor from the segmented mask, clinicians can monitor the tumor's progression over time, helping to evaluate the effectiveness of treatments such as chemotherapy, radiation, or ablation therapies. Morphometric analysis, which includes measuring the shape, size, and boundary irregularity of tumors, can further help in assessing tumor heterogeneity, which is often associated with more aggressive cancer types.

3 Discussion

In modern medical imaging diagnosis and treatment planning, liver tumor segmentation holds an indispensable position. However, the inefficiency and inconsistent results of traditional manual segmentation methods have become a major barrier to clinical workflows. The findings of this study demonstrate that the proposed UAIR framework offers a substantial advancement in medical image segmentation, particularly in its capacity to balance high accuracy with the practical demands of a clinical workflow.

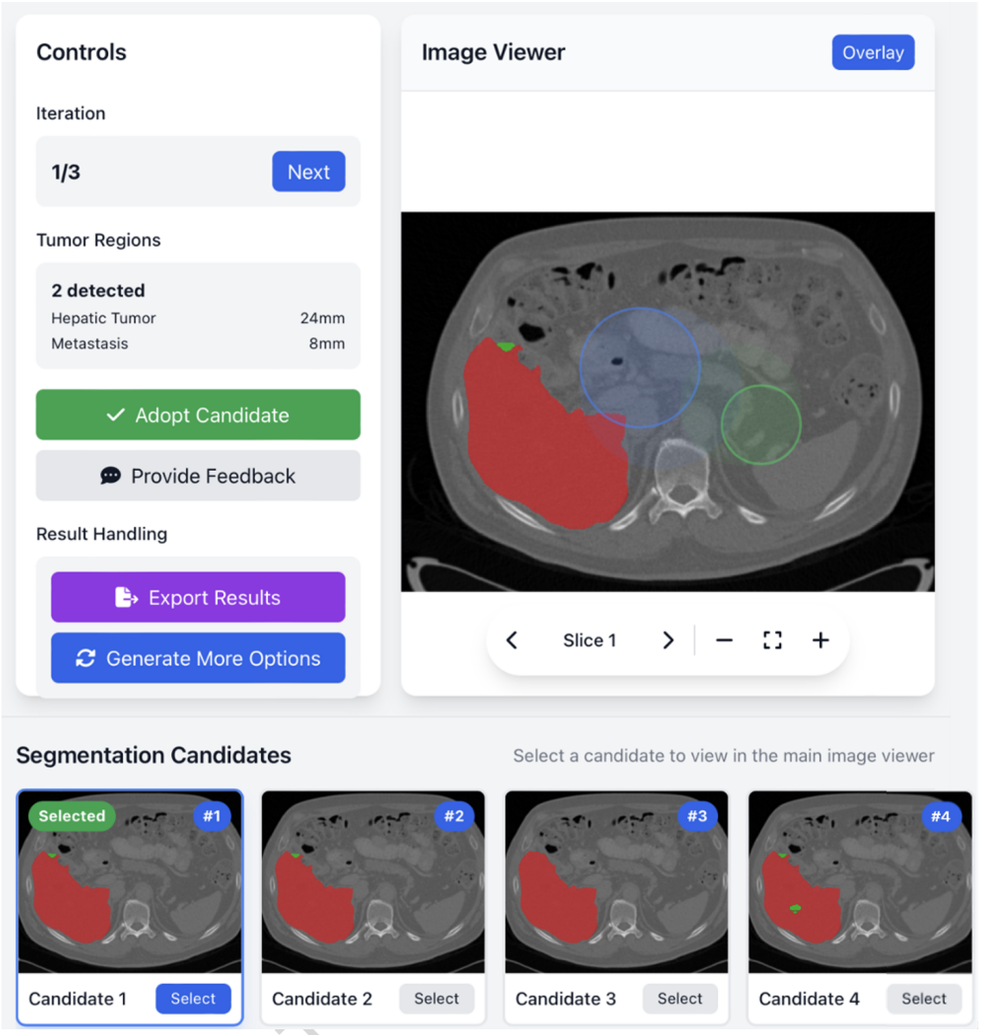

Fig. 3 User interface design of our proposed user-preference alignment segmentation framework. The left panel includes controls for navigating through segmentation iterations, tumor region summary, and buttons for exporting results or generating additional segmentation options. The central panel displays the selected CT slice with liver and tumor segmentation overlays. Below, multiple segmentation candidates are presented to facilitate efficient user preference alignment and iterative refinement.

As our results indicate, our framework achieves the highest segmentation performance compared with existing methods. However, the primary clinical significance lies not just in this quantitative accuracy, but in how it is achieved. The framework directly confronts the critical bottlenecks of both manual segmentation (labor-intensive, high variability [10]) and existing promptable models (high interaction cost [23]), which have thus far hindered the widespread adoption of automated tools in clinical practice.

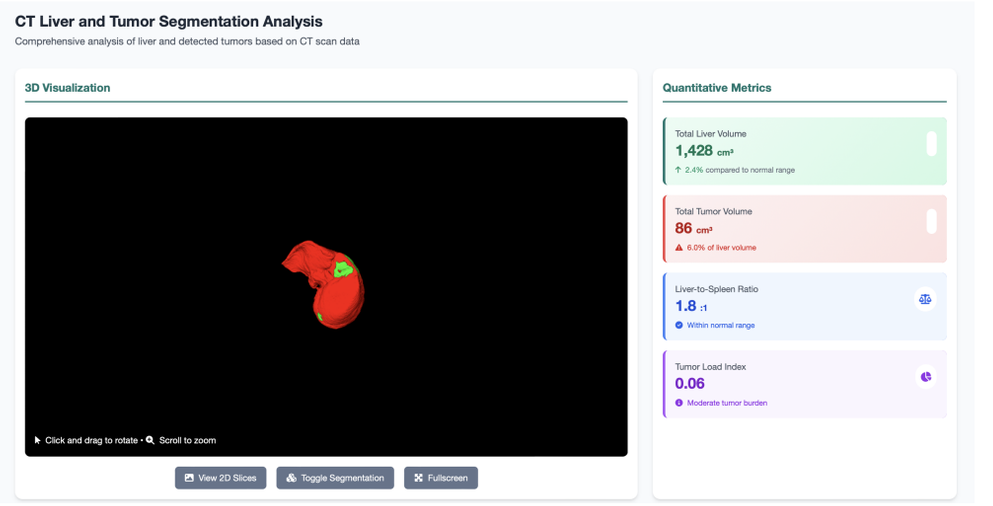

Fig. 4 Clinical analysis dashboard showcasing the downstream application of liver and tumor segmentation. The platform provides an interactive 3D visualization of the segmentation results, which enables precise calculation of organ and tumor volumes for tumor staging, assessing growth, and treatment planning, so as to facilitate morphometric analysis for evaluating tumor heterogeneity and monitoring therapeutic efficacy over time.

A key innovation is its explicit and efficient handling of inherent uncertainties and user-preference, a challenge highlighted in our introduction. Standard automated models provide a single, fixed output, which fails to account for the fact that different clinical scenarios require different segmentation criteria (e.g., inclusive boundaries for detection vs. conservative boundaries for radiotherapy planning [16, 17]). UAIR is designed for this exact ambiguity. By presenting a small, curated set of diverse candidates guided by model uncertainty, it empowers the clinician to select the segmentation that aligns with their specific diagnostic or therapeutic goal. This selection-based paradigm is a significant departure from the laborious correction-based interaction of SAM-like models [19, 20], which still require manual clicks or box drawing.

The efficiency gained from the UAIR framework translates directly into significant clinical advantages. Traditional objective criteria like RECIST [8] rely on one-dimensional linear measurements, which are prone to high inter-observer variability and can compromise the reliability of therapeutic response assessment. UAIR addresses this limitation by providing reproducible 3D volumetric segmentation, enabling volumetric RECIST (vRECIST), a more accurate and comprehensive standard for evaluating treatment response. It enhances reproducibility by enabling a clinician to consistently apply a specific segmentation preference by selecting the most appropriate candidate. This consistency across longitudinal follow-up scans is critical for accurately tracking tumor burden and making confident decisions about treatment continuation or modification. Besides, in high-stakes procedures such as surgical resection or ablative therapies [9], segmentation accuracy is non-negotiable. UAIR's

real-time adaptability is a major asset here. It allows surgeons and radiation oncologists to interactively refine segmentations to their exacting standards. This accelerates the planning workflow without sacrificing the personalized precision required for safe and effective intervention.

While UAIR demonstrates significant promise, its effectiveness relies on the quality and diversity of the initial uncertainty-guided candidates. Generating prompts for large organs such as the liver is relatively straightforward, and even a small set of candidates usually provides at least one satisfactory option. However, for small or irregular tumors, initial candidates may occasionally fail to capture the full lesion, particularly in regions with low contrast or ambiguous boundaries [28]. In such cases, the iterative refinement can still improve results if the user selects the best available candidate, as subsequent iterations update the uncertainty map and regenerate candidates based on the selected reference. However, if all initial candidates are poor, the iterative process may provide limited improvement, highlighting the importance of ensuring sufficient candidate diversity in the first iteration. Future work could explore optimizing this candidate generation to ensure a relevant option is presented in the first iteration for most cases. Furthermore, the framework could be extended to learn a specific user's or department's preferences over time, further minimizing the need for interaction [29, 30]. Although the current study validates UAIR only on CT imaging, the framework is modality-agnostic and can in principle be applied to liver tumor segmentation from other imaging modalities [31–33], as it relies on highlighting high-uncertainty regions and allowing user-guided refinement rather than modality-specific features. In addition, incorporating a longitudinal comparison of segmentation results across multiple time points could provide valuable insights into disease progression and treatment response. Such an extension would allow the dashboard to track temporal trends, assess the stability of the segmentation framework over successive scans, and further evaluate the effectiveness of the UAIR interaction strategy in longitudinal monitoring.

In conclusion, the UAIR framework represents a pivotal step toward the practical adoption of AI segmentation in oncology. By re-framing the human-computer interaction from laborious correction to efficient selection, it respects the clinician's valuable time while harnessing their essential expertise. This uncertainty-aware, preference-aligned approach provides the accuracy, adaptability, and efficiency needed to transform quantitative imaging from a research-level tool into a robust, everyday component of personalized cancer care.

4 Methods

4.1 Foundation Model for Medical Image Segmentation

A plethora of foundation models have been released for various computer vision tasks [34, 35], but most of these models rely on interactions between images and text. The Segment Anything Model (SAM) [19], a foundation model for image segmentation, addresses this limitation by performing general segmentation tasks without the need for additional text guidance. SAM consists of an Image Encoder that generates cached embeddings and a Prompt Encoder that processes user inputs into sparse and dense

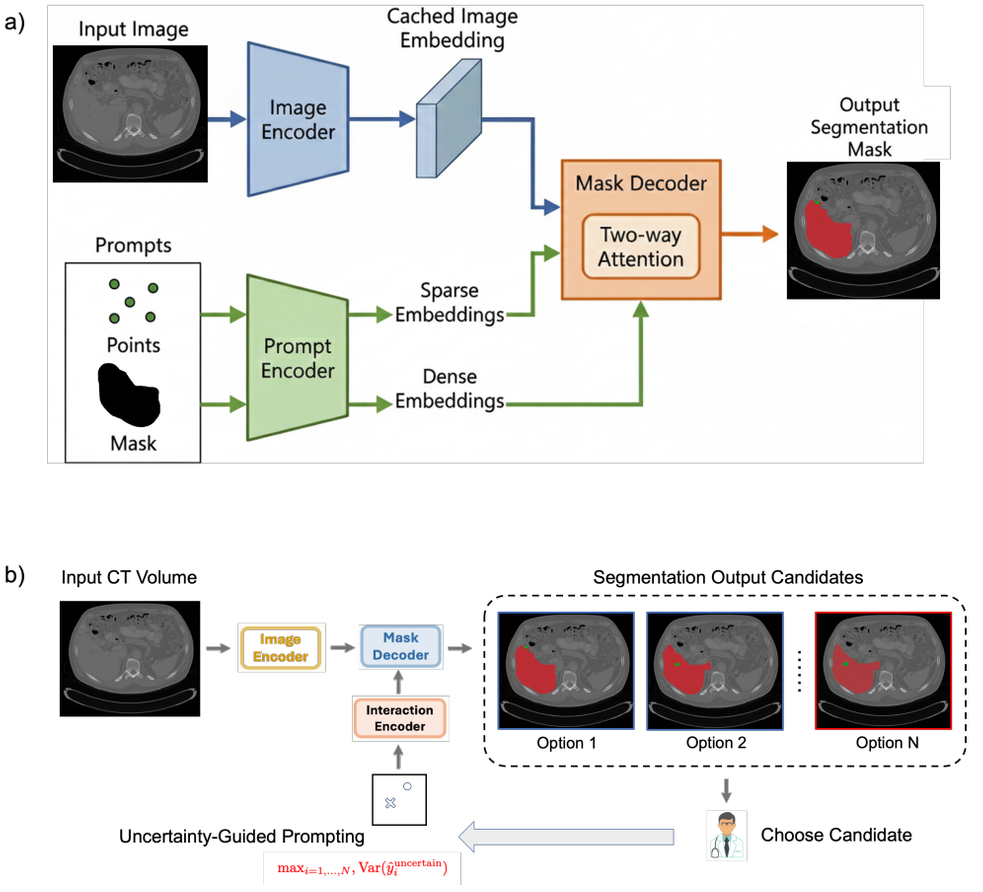

Fig. 5 Methodological details of our proposed framework. (a) The segmentation builds upon the architecture of the Segment Anything Model (SAM), which consists of an Image Encoder that generates cached embeddings and a Prompt Encoder that processes user inputs into sparse and dense embeddings. These are integrated by a Mask Decoder using two-way attention to produce the final segmentation. (b) Building upon the framework of SAM, our framework generates multiple segmentation output candidates based on different prompts. A clinician selects the optimal candidate, then the uncertainty-guided prompting is adopted via the maximum variance of uncertain predictions to generate new series of segmentation output candidates through iterative feedback for refinement of segmentation.

embeddings. These are integrated by a Mask Decoder using two-way attention to produce the final segmentation, as shown in Fig. 5. SAM is pretrained on a large and diverse dataset, enabling it to generalize across a wide range of domains, including medical images. SAM is capable of segmenting any object in an image with minimal input from the user, thanks to its novel approach of combining pre-trained visual representations with interactive prompts. It supports multiple input modalities, including

points, boxes, and free-form masks, making it highly flexible for various segmentation applications.

In the context of medical image segmentation, the MedSAM [20] extension is a specialized adaptation of SAM, fine-tuned specifically for tasks in the medical domain. By leveraging the generalizable capabilities of SAM and incorporating domain-specific knowledge, MedSAM is tailored to handle complex structures in medical images, such as organs, lesions, and tumors. The model enhances segmentation accuracy and robustness in clinical settings, offering great potential for applications in radiology, pathology, and other medical imaging fields [18, 36, 37]. This adaptation brings powerful AI-driven segmentation tools into the hands of clinicians and researchers, providing them with more efficient and accurate tools for diagnosing and analyzing medical conditions.

4.2 Automatic Prompting and Segmentation with Limited Templates

Despite the impressive generalization capabilities of foundation models for medical image segmentation, these models heavily rely on precise user-provided prompts like points and boxes to guide the segmentation process. In medical scenarios such as 3D CT imaging, manual annotations are not only costly and time-consuming but are also subject to human bias, making them impractical for large-scale clinical applications [38].

To address this challenge, automatic prompting technology is designed to address this challenge by enabling the model to autonomously generate effective prompts, thereby eliminating the need for manual intervention. The core idea is to leverage limited templates to learn generalized anatomical localization capabilities. MedLSAM [27] achieves automatic prompting by utilizing the anatomical similarity between templates and target images to automatically discover key localization cues and utilizes self-supervised learning to extract generalized anatomical features from large-scale medical datasets, which can then support the generation of relevant prompts. These techniques do not require dense annotations across the entire dataset. Instead, they only need a small number of template scans to effectively localize and generate prompts for target structures, significantly r educing a nnotation c osts w hile m aintaining high segmentation accuracy.

4.3 Uncertainty Estimation of Segmentation Targets

In medical image segmentation, uncertainty estimation plays a crucial role in assessing the reliability of segmentation results, which is especially important in high-stakes domains such as healthcare [39–41]. In particular, the ability to assess segmentation uncertainty enables the identification of regions with less confidence, allowing clinicians to focus on potentially ambiguous areas.

This section presents a framework for estimating uncertainty in segmentation targets, which is particularly useful in tasks where precise and reliable predictions are critical. The framework distinguishes between two main types of uncertainty: aleatoric and epistemic uncertainty. Aleatoric uncertainty arises from inherent noise or variability in the data itself such as image artifacts or unclear boundaries, while epistemic

uncertainty reflects the model's lack of knowledge, which can be mitigated with more training data. To represent and quantify uncertainty, the model generates a probabilistic prediction over possible segmentations $\hat{y}$ given an input image x. The uncertainty is captured by the variance of the predicted segmentations, which can be expressed as $\sigma_{\hat{y}}^2$, where $\hat{y}$ represents the model's output. Following [22], we utilize a series of different generated prompts $b_1, b_2, \ldots, b_N$ to generate different predictions $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N$, which are used to compute the mean and variance of the predicted segmentation as follows:

$$\mu_{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i, \quad \sigma_{\hat{y}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{y}_i - \mu_{\hat{y}})^2. \tag{1}$$

These statistics capture the central tendency and the variability of the segmentation masks generated by the ensemble. Intuitively, higher variance in the segmentation predictions indicates higher uncertainty in the model's output, particularly in areas where the boundaries of the target are ambiguous. In such regions, the model's predictions may differ significantly across different prompts, leading to a lack of confidence in the segmentation result. To quantify this uncertainty, we define the uncertainty measure $u(\hat{y})$, which reflects the variability in the model's output for each pixel or region in the image. This measure can be derived directly from the variance $\sigma_{\hat{y}}^2$, which serves as a proxy for the uncertainty in the segmentation process. Specifically, for each pixel, the uncertainty is proportional to the standard deviation $\sigma_{\hat{y}}$, with higher values indicating greater uncertainty about the correct label.

When segmenting targets with ambiguous boundaries, SAM tends to mis-segment regions with higher uncertainty, leading to poor performance. To address this, our framework uses a threshold u_{th} to identify high-uncertainty areas where $S_{\hat{y}}$ and S_b represent the areas of the ensembled segmentation mask and bounding box, respectively. The high-uncertainty areas are then selected using:

$$M_{\text{unc}} = u(\hat{y}) > T_{\text{unc}} \tag{2}$$

where M_{unc} represents the binary mask indicating high-uncertainty regions, and T_{unc} is a threshold value that determines which areas are considered uncertain. By focusing on these high-uncertainty regions, the prompts effectively guide the candidate generation process toward plausible alternative segmentations, enabling the user to efficiently refine and correct the segmentation results.

4.4 Uncertainty-Guided Prompting

In medical image segmentation, one of the primary challenges is addressing uncertainty caused by inherent image quality defects and varying annotation preferences, often due to differences in clinical scenarios [17]. These sources of uncertainty must be effectively captured and leveraged to generate a minimal yet diverse set of segmentation candidates that cover a wide range of potential user preferences. The candidate generation process is automatic and does not rely on additional user interaction during the initial step. Instead, it is driven by the model's ability to detect and quantify

uncertainty. We accomplish this by focusing on the identification and selection of the most distinct segmentation candidates based on high-uncertainty regions. These high-uncertainty zones are used as prompts to guide the candidate generation process. Specifically, the model first selects four segmentation outcomes that exhibit the largest variation in the uncertain regions. This is formalized as:

$$\max_{i=1,\ldots,N}, \mathrm{Var}(\hat{y}_i^{\mathrm{uncertain}}) \tag{3}$$

where $\hat{y}_i^{\mathrm{uncertain}}$ represents the predicted masks in the uncertain regions, and $\mathrm{Var}(\hat{y}_i^{\mathrm{uncertain}})$ measures the variation across all predictions in those regions. The candidates selected exhibit the highest uncertainty, ensuring the set covers a broad spectrum of potential segmentation variations. Upon the user's selection of the most appropriate candidate from the initial set, the model updates its candidate pool in an iterative manner. The user's feedback provides a reference for refining subsequent candidate generation. This is achieved by recalculating the uncertainty regions based on the selected candidate and adjusting the uncertainty map accordingly. The updated candidates are then generated by:

$$\hat{y}_i^{\mathrm{new}} = f(\hat{y}_i^{\mathrm{previous}}, \mathrm{Feedback}_i) \tag{4}$$

where $\hat{y}_i^{\mathrm{new}}$ denotes the new candidate after incorporating the feedback from the user, and $f(\cdot)$ represents the function that refines the segmentation output based on the feedback. After the user selects the most preferred candidate from the initial set, the framework enters an iterative refinement loop. The selected segmentation is fed back as the updated initial reference for the next iteration, replacing the original automatic prediction. Based on this updated reference, uncertainty regions are re-estimated, and a new uncertainty map is constructed.

These high-uncertainty zones are used as prompts to guide the candidate generation process. Specifically, we focus on selecting the four most distinct segmentation outcomes, those that exhibit the largest variation in the uncertain regions. This selection process ensures that the candidate set provided to the user captures a broad spectrum of potential segmentation variations, reflecting diverse interpretations that may arise from differences in clinical contexts or annotation preferences.

The selected candidates are automatically presented to the user for further interaction, with the aim of refining and improving the segmentation results. By presenting multiple, distinct candidates, we reduce the risk of missing potential alternatives and empower the user to make more informed decisions. This strategy minimizes manual intervention, as users are provided with a manageable yet diverse set of outputs, covering a range of possible segmentations within the uncertain areas. In doing so, our approach not only addresses the inherent uncertainty in medical image segmentation but also facilitates a more flexible and adaptive framework that can accommodate the varying preferences and judgment of different clinicians. Furthermore, our framework effectively leverages uncertainty as a tool to enhance the segmentation process. Rather than ignoring or smoothing over uncertain regions, we use these areas as an opportunity to present the user with meaningful options. This leads to a more interactive and user-guided refinement process, ensuring that the final segmentation result is

both accurate and reflective of the user's preferences, while maintaining the system's efficiency and robustness.

4.5 Training and Implementation Details

In our study, all of our experiments are implemented in Python with PyTorch, using an NVIDIA A100 GPU. For auto-prompting, we follow the approach outlined in [27] by randomly selecting five scans from each dataset, using the real bounding box prompts for each target organ as support images for the localization model. For each organ in these selected scans, we calculate the maximum and minimum coordinates and compute the average of both the coordinates and features across all the support images. This process results in an averaged representation of the latent coordinates and feature extremities, which are then used for auto-prompting. To standardize the input images, we rescale the voxel spacing to $3\times3\times3$ mm^3 and crop the patches to a size of $64 \times 64 \times 64$ pixels. The 3D bounding box generated during the localization step is then extended by [2, 10, 10] pixels in the z, x, and y directions, respectively, ensuring that the entire target organ is fully enclosed within the bounding box. For the segmentation task, we apply the same preprocessing procedure as described in [20]. This includes adjusting the slice resolution to 1024×1024 pixels and normalizing the pixel values to a range of [-500, 1000], as this range effectively encompasses most tissue types, ensuring consistency in the input data across different scans.

To quantitatively evaluate segmentation results, we employ two different metrics to measure the overlap between the predicted segmentation and the ground truth. The first evaluation metric is the Dice Similarity Coefficient (DSC), which measures the similarity between two regions. It is calculated as twice the area of intersection divided by the sum of the areas of both regions. The Dice score ranges from 0 to 1, with 1 indicating perfect overlap. The second metric is the Jaccard Index, which also measures region overlap. It calculates the ratio of the intersection to the union of the two regions. The Jaccard Index ranges from 0 to 1, with higher values indicating better segmentation. Given predicted segmentation (A) and ground truth (B), the formulas are as follows:

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{5}$$

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

Ethics approval and consent to participate

This study exclusively uses publicly available CT datasets and does not involve any new experiments with human participants or animals performed by any of the authors. Therefore, additional ethical approval and consent were not required.

Declarations

Data availability

All imaging datasets analyzed in this study are publicly accessible. The AbdomenAtlas

dataset is available at (https://github.com/MrGiovanni/AbdomenAtlas). Processed or derived data supporting the findings of this study are available from the corresponding author upon reasonable request.

Materials availability

No new biological or chemical materials were generated or analyzed in this study.

Acknowledgement

This work was supported by the Natural Science Foundation of Hubei Province (Grant No. 2025AFD774).

Author contributions

G.Z., C.G. and Y.W. contributed equally to this work, having full access to all study data and assuming responsibility for the integrity and accuracy of the analyses (Validation, Formal analysis). G.Z., C.G. and G.H. conceptualized the study, designed the methodology, and participated in securing research funding (Conceptualization, Methodology, Funding acquisition). Y.W. and X.Z. carried out data acquisition, curation, and investigation (Investigation, Data curation) and provided key resources, instruments, and technical support (Resources, Software). Z.W. drafted the initial manuscript and generated visualizations (Writing – Original Draft, Visualization). T.C. and B.Y. supervised the project, coordinated collaborations, and ensured administrative support (Supervision, Project administration). All authors contributed to reviewing and revising the manuscript critically for important intellectual content (Writing – Review & Editing) and approved the final version for submission.

Conflict of interest/Competing interests

The authors declare no competing financial or non-financial interests.

References

[1] Wild, C.P., Weiderpass, E., Stewart, B.W.: World cancer report (2020)

[2] Sedano, R., Cabrera, D., Jiménez, A., Ma, C., Jairath, V., Arrese, M., Arab, J.P.: Immunotherapy for cancer: common gastrointestinal, liver, and pancreatic side effects and their management. Official journal of the American College of Gastroenterology— ACG **117**(12), 1917–1932 (2022)

[3] Midya, A., Chakraborty, J., Srouji, R., Narayan, R.R., Boerner, T., Zheng, J., Pak, L.M., Creasy, J.M., Escobar, L.A., Harrington, K.A., *et al.*: Computerized diagnosis of liver tumors from ct scans using a deep neural network approach. IEEE journal of biomedical and health informatics **27**(5), 2456–2464 (2023)

[4] Shiina, S., Sato, K., Tateishi, R., Shimizu, M., Ohama, H., Hatanaka, T., Takawa, M., Nagamatsu, H., Imai, Y.: Percutaneous ablation for hepatocellular carcinoma: comparison of various ablation techniques and surgery. Canadian Journal of Gastroenterology and Hepatology **2018**(1), 4756147 (2018)

[5] Gul, S., Khan, M.S., Bibi, A., Khandakar, A., Ayari, M.A., Chowdhury, M.E.: Deep learning techniques for liver and liver tumor segmentation: A review. Computers in Biology and Medicine **147**, 105620 (2022)

[6] Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., *et al.*: The liver tumor segmentation benchmark (lits). Medical image analysis **84**, 102680 (2023)

[7] Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2022)

[8] Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., *et al.*: New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). European journal of cancer **45**(2), 228–247 (2009)

[9] Virdis, F., Reccia, I., Di Saverio, S., Tugnoli, G., Kwan, S., Kumar, J., Atzeni, J., Podda, M.: Clinical outcomes of primary arterial embolization in severe hepatic trauma: a systematic review. Diagnostic and interventional imaging **100**(2), 65–75 (2019)

[10] Todorov, M.I., Paetzold, J.C., Schoppe, O., Tetteh, G., Shit, S., Efremov, V., Todorov-Völgyi, K., Düring, M., Dichgans, M., Piraud, M., *et al.*: Machine learning analysis of whole mouse brain vasculature. Nature methods **17**(4), 442–449 (2020)

[11] Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[12] Alirr, O.I., Rahni, A.A.A.: Survey on liver tumour resection planning system: steps, techniques, and parameters. Journal of Digital Imaging **33**(2), 304–323 (2020)

[13] Moghbel, M., Mashohor, S., Mahmud, R., Saripan, M.I.B.: Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. Artificial Intelligence Review **50**(4), 497–537 (2018)

[14] Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging **32**(4), 582–596 (2019)

[15] Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 119–127 (2019). Springer

[16] Prasanna, P.G., Stone, H.B., Wong, R.S., Capala, J., Bernhard, E.J., Vikram, B., Coleman, C.: Normal tissue protection for improving radiotherapy: Where are the gaps? Translational cancer research **1**(1), 35 (2012)

[17] Zhu, J., Wu, J., Ouyang, C., Kamnitsas, K., Noble, J.A.: Spa: Efficient user-preference alignment against uncertainty in medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23731–23740 (2025)

[18] Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in Biology and Medicine **171**, 108238 (2024)

[19] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., *et al.*: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)

[20] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)

[21] Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., *et al.*: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In: European Conference on Computer Vision, pp. 51–67 (2024). Springer

[22] Zhang, Y., Hu, S., Xue, L., Ren, S., Hu, Z., Cheng, Y., Qi, Y.: Enhancing the reliability of auto-prompting sam for medical image segmentation with uncertainty estimation and rectification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1282–1291 (2025)

[23] Deng, G., Zou, K., Ren, K., Wang, M., Yuan, X., Ying, S., Fu, H.: Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 368–377 (2023). Springer

[24] Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., *et al.*: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis **97**, 103285 (2024)

[25] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[26] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., *et al.*: Transunet: Rethinking the u-net architecture design for medical

image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)

[27] Lei, W., Xu, W., Li, K., Zhang, X., Zhang, S.: Medlsam: Localize and segment anything model for 3d ct images. Medical Image Analysis **99**, 103370 (2025)

[28] Zhang, D., Chen, B., Chong, J., Li, S.: Weakly-supervised teacher-student network for liver tumor segmentation from non-enhanced images. Medical Image Analysis **70**, 102005 (2021)

[29] Ma, Y., Wang, J., Yang, J., Wang, L.: Model-heterogeneous semi-supervised federated learning for medical image segmentation. IEEE Transactions on Medical Imaging **43**(5), 1804–1815 (2024)

[30] Jiang, M., Yang, H., Cheng, C., Dou, Q.: Iop-fl: Inside-outside personalization for federated medical image segmentation. IEEE Transactions on Medical Imaging **42**(7), 2106–2117 (2023)

[31] Zhao, J., Li, D., Xiao, X., Accorsi, F., Marshall, H., Cossetto, T., Kim, D., McCarthy, D., Dawson, C., Knezevic, S., *et al.*: United adversarial learning for liver tumor segmentation and detection of multi-modality non-contrast mri. Medical image analysis **73**, 102154 (2021)

[32] Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., *et al.*: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022)

[33] Zhang, Y., Xue, L., Zhang, W., Li, L., Liu, Y., Jiang, C., Cheng, Y., Qi, Y.: Seganypet: Universal promptable segmentation from positron emission tomography images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21107–21116 (2025)

[34] Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2022, p. 3876 (2022)

[35] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H), pp. 353–367 (2023). PMLR

[36] Zhang, Y., Lv, B., Xue, L., Zhang, W., Liu, Y., Fu, Y., Cheng, Y., Qi, Y.: Semisam+: Rethinking semi-supervised medical image segmentation in the era of foundation models. Medical Image Analysis (2025)

[37] Ali, M., Wu, T., Hu, H., Luo, Q., Xu, D., Zheng, W., Jin, N., Yang, C., Yao,

J.: A review of the segment anything model (sam) for medical image analysis: Accomplishments and perspectives. Computerized Medical Imaging and Graphics **119**, 102473 (2025)

[38] Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., Jin, C.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Computers in Biology and Medicine **169**, 107840 (2024)

[39] Zhang, Y., Jiao, R., Liao, Q., Li, D., Zhang, J.: Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artificial Intelligence in Medicine **138**, 102476 (2023)

[40] Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. Meta-Radiology **1**(1), 100003 (2023)

[41] Zhou, N., Zou, K., Ren, K., Luo, M., He, L., Wang, M., Chen, Y., Zhang, Y., Chen, H., Fu, H.: Medsam-u: Uncertainty-guided auto multi-prompt adaptation for reliable medsam. IEEE Transactions on Circuits and Systems for Video Technology (2025)