New paradigms for radiological segmentation with AI foundation models: automatic prompting and interactive VR agent design

Pascal Spiegler

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

August 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Pascal Spiegler**

Entitled: **New paradigms for radiological segmentation with AI foundation models: automatic prompting and interactive VR agent design**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Charalambos Poullis

_____ Examiner
Dr. Charalambos Poullis

_____ Examiner
Dr. Marta Kersten-Oertel

_____ Supervisor
Dr. Yiming Xiao

Approved by _____
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

New paradigms for radiological segmentation with AI foundation models: automatic prompting and interactive VR agent design

Pascal Spiegler

This thesis investigates data-efficient and interactive methods for medical image segmentation, addressing two major related hurdles: the scarcity of high-quality annotations for fully-supervised lesion segmentation and the labor/expertise-intensive nature of manual segmentation workflows. Our first contribution, YOLO-URSAM, is a weakly supervised intracranial hemorrhage (ICH) segmentation model that leverages the Segment-Anything Model (SAM), where we fine-tune YOLOv8 on bounding-box annotated CT scans, introduce automatic point prompting for SAM, and apply perturbation-based majority-voting for uncertainty rectification. YOLO-URSAM achieves 0.933 detection accuracy, 0.796 AUC, and a mean Dice score of 0.629, surpassing existing weakly supervised methods and popular supervised models (U-Net and Swin-UNETR) on available public data. Our second contribution, SAMIRA, is a virtual reality system with a conversational AI agent for semi-automated 3D radiological segmentation. Users issue simple voice commands to initialize masks, correct them using a human-in-the-loop approach, then view the 3D segmentations as a true-to-scale 3D mesh. In a user study, SAMIRA achieved a System Usability Scale score of 90.0±9.0, demonstrated a low cognitive load, and was praised for its intuitive guidance, educational benefits, and immersive visualization. Together, these methods combine AI foundation models, prompting, and limited human interaction to deliver accurate and efficient segmentation for clinical imaging.

Acknowledgments

Above all, I want to thank my supervisor, Dr. Yiming Xiao, for his support, always taking the time to answer my many questions, and for giving me the freedom and resources to follow my curiosity. I'm especially grateful for his trust in me during my transition from life sciences to computer science, and for believing in me, both within and beyond research. I feel incredibly lucky to have been mentored by someone so caring, hard working, and insightful.

I would like to thank my colleagues from the Health-X Lab, who I feel privileged to have collaborated with, and who fostered a supportive environment. I am also grateful to Dr. Marta Kersten-Oertel, from whom I learned a great deal about human-computer interaction, and to our clinical collaborators, Dr. Alexander G. Weil, Dr. Aris Hadjinicolaou, and Dr. Corey S. Miller, who trusted me with medical projects beyond the scope of this thesis, which I thoroughly enjoyed working on.

To my family: thank you for the constant and unconditional support, not just throughout this degree, but throughout the course of my life. Your presence made all the difference, especially following my injury this year. To my girlfriend, Chang: your encouragement and care have been a source of strength and comfort throughout this challenging yet rewarding journey, and I am grateful for you. To my close friends: Carlo, Vincent, Borys, Lucas, Antonio, thank you for keeping me grounded, making me laugh, and reminding me to step back and enjoy life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Clinical Significance of Lesion Segmentation

Lesion segmentation, the precise delineation of pathological structures in radiological images, plays a central role in medical image analysis. This thesis focuses on two key applications: tumor segmentation and intracranial hemorrhage segmentation. Under these contexts, accurate localization and volumetric quantification support treatment planning, prognostic evaluation, and longitudinal monitoring. For example, tumor segmentation guides surgical and radiotherapy strategies, while hemorrhage segmentation guides urgent clinical decisions. In addition to the use cases focused on in this thesis, lesion segmentation has broader applications, such as the assessment of pulmonary nodules (during lung cancer screenings) [43], hepatic lesions (prior to resection or ablation) [57], vascular plaques [17], and more.

1.1.1 Intracranial Hemorrhage

Intracranial hemorrhage (ICH) constitutes about 10–15% of all strokes and carries a high early mortality [38]. In clinical settings, rapid volumetric measurement of hematoma size and expansion is essential because larger bleeds predict worse outcomes and guide urgent interventions, such as surgical evacuation or blood pressure control [71, 59]. Furthermore, ICH can occur in different compartments of the cranium, leading to different subtypes (Figure 1.1): intraventricular (IVH),

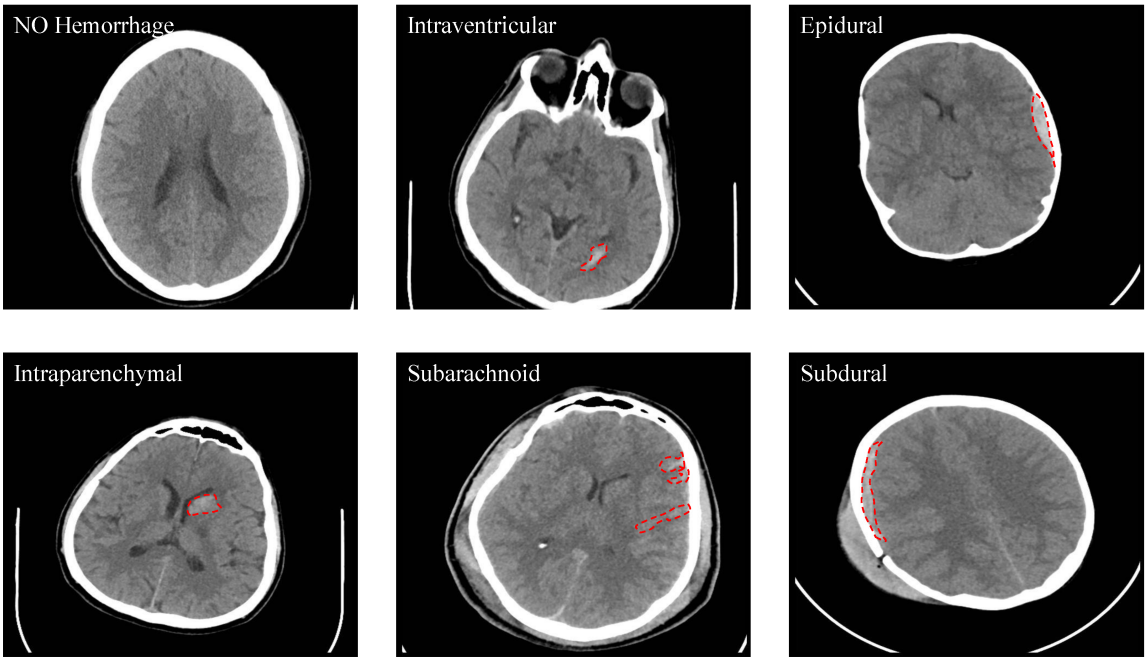


Figure 1.1: Representative CT images of intracranial hemorrhage subtypes [37]

epidural (EDH), intraparenchymal (IPH), subarachnoid (SAH), and subdural (SDH), with each subtype requiring a tailored treatment approach [3]. For example, SAH may require aneurysm repair [1], whereas large intraparenchymal bleeds would need rapid decompressive surgery [20]. Precise localization and quantification of all hemorrhage types are therefore essential for timely clinical decision-making.

1.1.2 Tumor Oncology

Beyond stroke care, lesion segmentation is important in oncology. Delineating tumor margins informs surgical planning, radiation targeting, and volumetric analysis. For example, with brain tumors, segmentation of longitudinal scans allows clinicians to monitor growth or treatment response, informing shifts in therapy if a tumor is not responding [87]. Furthermore, accurately delineating tumor boundaries enables precise measurement of tumor size and shape, which are key prognostic indicators, since larger tumor volumes and irregular shapes correlate with poorer survival and higher recurrence across various cancer types [44, 68, 2].

1.1.3 Medical Education

In addition to its direct clinical utility, manual lesion segmentation can serve as an important tool in medical education. Previous studies have shown that hands-on workshops where medical students and clinicians create 3D segmentation masks significantly improve confidence in interpreting scans and anatomical understanding [41, 16]. Clear manual delineation between different neighboring anatomical structures, as well as between healthy and pathological tissues, requires in-depth knowledge on the radiological features and human anatomy. With suitable interactive guidance, as novice users complete manual segmentation, they can also develop enhanced clinical proficiency in radiological reading and anatomical understanding.

1.2 Challenges

Manual lesion segmentation is time-consuming and challenging. Expert readers may spend hours per case delineating structures slice by slice across hundreds of 2D images to build a full 3D segmentation, navigating through axial, sagittal, and coronal views. Furthermore, developing the required proficiency to manually produce segmentations demands extensive supervised training and repeated practice to build diagnostic confidence [12].

These limitations in manual annotation directly bottleneck the development of automatic segmentation models. Deep-learning approaches rely on large quantities of high-quality, pixel-level masks for training, yet most clinical datasets contain only a handful of fully labeled volumes. As a result, models trained on these few annotations tend to overfit and generalize poorly when applied to images from different scanners, protocols, or patient populations.

1.3 Proposed Solutions

To overcome the scarcity of pixel-level annotations and support both rapid mask creation and user training, we introduce two complementary approaches:

YOLO-URSAM [82]: A fully automated, weakly supervised model that turns inexpensive bounding-box labels into accurate segmentation masks. By combining a YOLOv8 detector with

the Segment Anything Model, a point prompt generator, and an uncertainty-rectification voting scheme, YOLO-URSAM removes the need for costly and difficult to produce pixel-level annotations for deep learning model training.

SAMIRA [83]: A human-in-the-loop VR system with a conversational AI agent that guides users through segmentation mask creation, refinement, and 3D visualization. By combining AI-proposed segmentations with just a handful of user point prompts, it achieves quality on par with fully manual annotations, dramatically reduces annotation time, and provides real-time feedback that teaches users how to segment pathology. In doing so, SAMIRA lowers the time, expertise, and training barriers to generating high-quality pixel-level masks.

1.4 Thesis Organization

Chapter 2 reviews medical imaging fundamentals, deep learning segmentation architectures, foundation models, and interactive AI agent paradigms. **Chapter 3** presents the YOLO-URSAM framework for weakly supervised ICH segmentation, including methodological details and quantitative results. **Chapter 4** details the SAMIRA VR agent system, interaction paradigm evaluations, and user studies on usability and effectiveness across tumor cases. **Chapter 5** discusses the findings, clinical implications, limitations, and directions for future research.

Chapter 2

Background

2.1 Medical Imaging

Radiological imaging uses non-invasive technologies to visualize internal anatomy and pathology. Unlike conventional RGB photographs, most clinical imaging modalities are single-channel (grayscale) with modality-specific contrast and metadata (patient identifiers, voxel size, orientation) stored in file formats, such as DICOM or NIfTI. In this thesis, we focus on Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), each with unique imaging principles, contrast mechanisms, and clinical applications.

2.1.1 Computed Tomography

Computed Tomography (CT) uses a rotating X-ray source and detector array to acquire multiple projections, which are reconstructed (e.g., via the Radon transform) into volumetric maps of X-ray attenuation. More specifically, CT scanners measure the linear attenuation coefficient $\mu(x)$ of tissue, which is converted to Hounsfield Units (HU) via:

$$\mathrm{HU}(x) \;=\; 1000 \cdot \frac{\mu(x) - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}}$$

where water is 0 HU and air is approximately –1000 HU [42]. The image intensity of CT ranges roughly from –1000 HU (air) to +3000 HU (metal) [31]. Due to the large dynamic range, to optimize

contrast for specific tissue types (e.g., bone), a *window width* (WW) and *window level* (WL, also called center) are often selected. The lower (L) and upper (U) HU bounds of the display window are thus given by

$$L = WL - \frac{WW}{2}, \quad U = WL + \frac{WW}{2}.$$

With a typical grayscale colormap, voxel values below L are mapped to black, those above U are mapped to white, and intermediate HUs are scaled linearly between these extremes:

$$I_{\text{disp}}(x) = \begin{cases} 0, & \text{HU}(x) \leq L, \\ \dfrac{\text{HU}(x) - L}{WW}, & L < \text{HU}(x) < U, \\ 1, & \text{HU}(x) \geq U. \end{cases}$$

Often, a *wide* window (e.g., $WW = 2000\,\text{HU}$) is useful for inspecting bone or lung fields, whereas a *narrow* window (e.g., $WW = 80\,\text{HU}$) enhances soft-tissue details. Common presets [64] include: **brain** ($WW = 80\,\text{HU}$, $WL = 40\,\text{HU}$), **bone** ($WW = 1800\,\text{HU}$, $WL = 400\,\text{HU}$), **lung** ($WW = 1500\,\text{HU}$, $WL = -600\,\text{HU}$), and **liver** ($WW = 150\,\text{HU}$, $WL = 30\,\text{HU}$). Besides these preset intensity windows, which are typically programmed in radiological software, for more complex tissue types and pathology, these parameters will need to be tailored accordingly.

2.1.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) leverages the nuclear magnetic resonance of protons in water molecules to generate high–contrast soft-tissue images. In an MRI scanner, the patient lies within a strong, uniform magnetic field created by a large superconducting magnet. A transmit coil then emits short radio-frequency (RF) pulses into tissues, and a receive coil detects the resulting signals [11]. In a static magnetic field B_0, the net magnetization vector $\mathbf{M}$ of spins (the magnetized protons) gets tipped by the RF pulses (called excitation) and then relaxes back to equilibrium [11]. The magnitude of the received voxel-wise MRI signal S, measured at the echo time (TE) with the associated repetition time (TR, the period for the cycle of "excitation-and-relaxation") can be

modeled as

$$S \propto \rho \left(1 - e^{-\,\mathrm{TR}/T_1}\right) e^{-\,\mathrm{TE}/T_2^*},$$

where ρ is proton density, T_1 is the longitudinal relaxation time, and T_2^* is the transverse relaxation time [11]. By varying TR, TE, and the inversion time (TI) if applicable, different tissue contrasts can be achieved [6].

T_1-weighted (T1w) images (Figure 2.1a) use a short TR ($\approx$ 400–600 ms) and short TE ($\approx$ 8–15 ms), providing high signal from soft tissues with short T_1 (e.g., fat) and low signal from free fluid. This contrast is used for general anatomical delineation and evaluation of tissue morphology. **Post-contrast T_1-weighted (T1c)** scans (Figure 2.1b) are acquired by T1w imaging after gadolinium administration. The contrast agent shortens T_1 in regions with disrupted blood–brain or blood–tissue barriers (e.g., tumors, inflammation), causing these areas to appear brighter and improving lesion detection. **T_2-weighted (T2w)** images (Figure 2.1c) use a long TR ($\approx$ 2500–3500 ms) and a long TE ($\approx$ 70–100 ms). Fluid-containing structures (e.g., edema, cysts) produce higher signal, whereas solid tissues appear relatively darker. **FLAIR (Fluid-Attenuated Inversion Recovery)** scans (Figure 2.1d) are acquired using a T2w sequence with TI ($\approx$ 2000–2600 ms) chosen to nullify signal from free fluid. Suppression of fluid signal enhances visibility of lesions next to fluid-filled spaces (e.g., edemas).

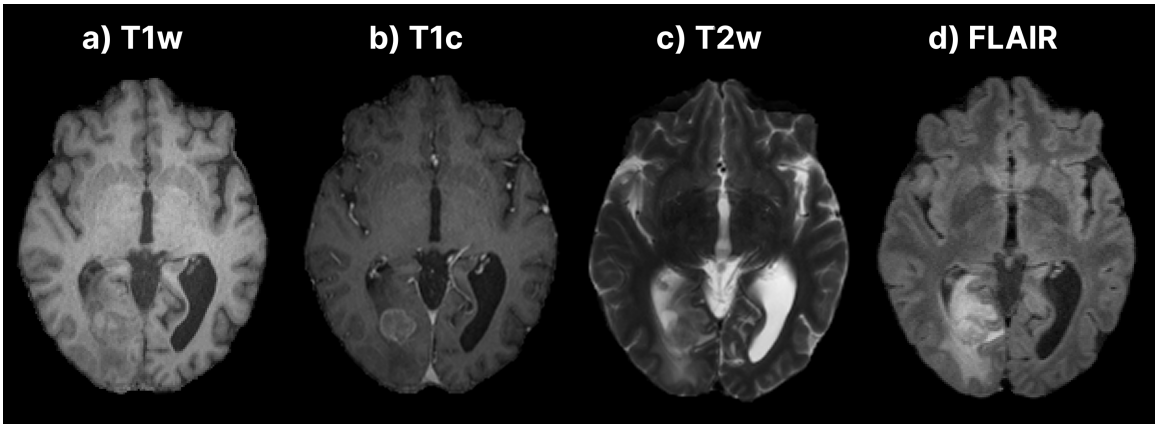


Figure 2.1: Example of four common MRI contrasts in a single pretreatment brain metastasis from the Pretreat-MetsToBrain-Masks cohort [72]. (a) T_1-weighted (T1w) scan highlights normal gray–white matter anatomy; (b) post-contrast T_1-weighted (T1c) shows contrast uptake in the tumor core; (c) T_2-weighted (T2w) accentuates the edema surrounding the tumor; (d) FLAIR suppresses cerebrospinal fluid to reveal the infiltrative edema.

MRI produces grayscale images with intensities in arbitrary units. Typical in-plane resolutions are 0.5–1.0 mm, with a slice thickness of 1–3 mm. Since MRI uses non-ionizing radio waves, it is better suited for longitudinal and pediatric studies compared to CT.

2.1.3 Medical Image Classification & Segmentation

Medical image processing for disease analysis often falls into two general categories: classification and segmentation.

Classification assigns a label $y \in \{1, \ldots, C\}$ to an image under study based on hand-crafted or learned image features. This is deemed a high-level image processing task. Formally, a classifier maps each input image to one of C classes (e.g., "benign" vs. "malignant" in the case of image-based tumor subtyping):

$$\text{Image} \rightarrow \{1, \ldots, C\}$$

Segmentation, which is a mid-level image processing task, divides an input image into M non-overlapping regions $R_1, \ldots, R_M$, assigning each pixel or voxel x exactly one label i (with $i = 1, \ldots, M$). In medical imaging these regions often correspond to organs or general tissue types (e.g., lesions). Classic segmentation algorithms typically operate directly on $I(x)$, the raw pixel- or voxel-wise intensity values or image statistics (e.g., local standard deviations). Some popular classic image segmentation techniques include global thresholding, region growing, and K-means clustering.

Global thresholding involves selecting one or more intensity cutoffs to partition the image's intensity histogram into two or more regions (e.g., binary or multi-level thresholding). Methods like Otsu's [69] automatically find thresholds that best separate intensity value classes.

Region growing starts from seed points and iteratively includes neighbouring pixels/voxels whose image features lie within a tolerance of the region's current mean till achieving pre-defined criteria.

K-means clustering partitions pixels/voxels into K clusters by minimizing the total within-cluster variance

$$J = \sum_{k=1}^{K} \sum_{x \in R_k} \big(I(x) - \mu_k\big)^2.$$

It proceeds in the following steps:

(1) Pick K initial centroids $\mu_1, \ldots, \mu_K$ at random from the intensity range.

(2) For each pixel x, compute the squared difference $(I(x) - \mu_j)^2$ to every centroid μ_j, and assign x to the cluster R_j with the smallest value.

(3) For each cluster j, recompute its centroid μ_j as the mean intensity of all pixels assigned to R_j:

$$\mu_j = \frac{1}{|R_j|} \sum_{x \in R_j} I(x).$$

(4) Repeat steps (2) and (3) until the centroids μ_j stop changing and convergence is reached.

2.1.4 Evaluation Metrics

For image classification/object detection and image segmentation, different metrics are often used to evaluate the relevant algorithms.

Detection/classification metrics: For object detection and image classification, common evaluation metrics include accuracy, precision, recall, specificity, and F1, which are defined based on true

positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) from the classification outcomes. Their detailed definitions are:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} && \text{(fraction of all examples correctly classified)},\\
\text{Precision} &= \frac{TP}{TP + FP} && \text{(of predicted positives, fraction that are true)},\\
\text{Recall} &= \frac{TP}{TP + FN} && \text{(of actual positives, fraction correctly detected)},\\
\text{Specificity} &= \frac{TN}{TN + FP} && \text{(of actual negatives, fraction correctly rejected)},\\
F_1 &= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} && \text{(balance between precision and recall)}.
\end{aligned}$$

Besides these metrics, for object detection, *false positive rate*, *true positive rate*, and the *area under the ROC curve (AUC)* have also been adopted. While the *false positive rate* is defined as $\text{FPR} = \frac{FP}{FP+TN}$ (fraction of negatives labeled positive), the *true positive rate* is the same as recall. The receiver operating characteristic (ROC) curve is a plot of the true positive rate versus the false positive rate under a varied decision threshold, and the area under the ROC curve is

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR},$$

which equals the probability that a randomly chosen positive example is assigned a higher continuous model score (e.g., confidence) than a randomly chosen negative.

Segmentation metrics: For a predicted binary mask P and ground truth mask G, an overlap is often measured with Dice coefficient and/or Intersection of Union (IoU):

$$\text{Dice}(P, G) = \frac{2\,|P \cap G|}{|P| + |G|}, \quad \text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|},$$

both ranging from 0 (no overlap) to 1 (perfect agreement).

2.2 Deep Learning Architectures for Image Segmentation

Deep learning methods use multi-layer neural networks to learn feature representations directly from data. In this thesis, we explore two training paradigms for segmentation models: fully supervised learning using pixel-wise ground-truth masks, and weakly supervised learning using more accessible, coarser annotations such as categorical labels and bounding boxes.

2.2.1 Fully Connected Networks

Fully connected networks, or multilayer perceptrons (MLPs), are the most foundational form of a neural network. These models consist of a series of layers, where each neuron in one layer connects to every neuron in the next layer. During forward passes, information flows from the input through one or more of these layers, each of which applies an affine transformation followed by a point-wise nonlinearity. During training, gradient-based optimization methods are used to adjust the network's weights and biases to minimize a loss function. Despite their simplicity, MLPs highlight the core concept of representation learning, backpropagation, and gradient-based optimization that underlie the more advanced architectures (e.g., CNNs, Transformers) used for image segmentation.

Forward passing: Let $\mathbf{x} \in \mathbb{R}^d$ be the input feature vector. The network consists of L layers of the form

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\,\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}, \qquad \mathbf{h}^{(l)} = \phi\big(\mathbf{z}^{(l)}\big), \quad l = 1, \ldots, L,$$

where $\mathbf{h}^{(0)} = \mathbf{x}$, each $\mathbf{W}^{(l)} \in \mathbb{R}^{h_l \times h_{l-1}}$ is a weight matrix, $\mathbf{b}^{(l)} \in \mathbb{R}^{h_l}$ a bias vector, and $\phi(\cdot)$ a point-wise nonlinearity. Two common choices are

$$\text{Rectified Linear Unit (ReLU):} \quad \phi(x) = \max(0,\, x), \qquad \text{Sigmoid:} \quad \phi(x) = \frac{1}{1 + e^{-x}}.$$

The vector $\mathbf{z}^{(l)}$ is called the pre-activation and $\mathbf{h}^{(l)}$ the activation of layer l.

Output and prediction: For classification into C classes, the final layer's pre-activations $\mathbf{z}^{(L)} \in \mathbb{R}^C$ are converted into probabilities via the softmax:

$$\hat{y}_c = \frac{\exp\big(z_c^{(L)}\big)}{\sum_{k=1}^{C} \exp\big(z_k^{(L)}\big)}, \quad c = 1, \ldots, C.$$

Next, a loss function compares $\hat{\mathbf{y}}$ against the true label.

Loss Functions: Some of the most common loss functions for deep learning-based medical image processing, which can vary depending on the task, are listed below.

Cross-entropy (CE) measures the difference between predicted class probabilities and the one-hot ground-truth labels, making it a common choice for multi-class classification tasks where each image (or pixel) must be assigned to exactly one of C categories. It is defined as:

$$L_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c.$$

where $y \in \{0,1\}^C$ is the one-hot ground-truth and $\hat{y} \in [0,1]^C$ the predicted class-probabilities.

Binary cross-entropy (BCE) evaluates the pixel-wise error between predicted mask probabilities and true binary labels, and is commonly used when training a network to produce a single foreground/background segmentation:

$$L_{\mathrm{BCE}} = -\frac{1}{P}\sum_{p=1}^{P}\big[m_p \log \hat{m}_p + (1-m_p)\log(1-\hat{m}_p)\big].$$

where $m \in \{0,1\}^P$ is the ground-truth mask and $\hat{m} \in [0,1]^P$ the predicted pixel-wise probabilities.

Dice loss maximizes the overlap between the predicted mask $\hat{m}$ and ground truth m, making it effective for segmentation tasks:

$$L_{\mathrm{Dice}} = 1 - \frac{2\sum_p \hat{m}_p m_p}{\sum_p \hat{m}_p + \sum_p m_p}.$$

Focal loss [55] down-weighs well-classified (easy) examples to focus learning on hard, misclassified examples, which is particularly useful in object detection or segmentation scenarios with severe class imbalance:

$$L_{\mathrm{FL}} = -\big(1 - p_t\big)^{\gamma} \log(p_t), \quad p_t = \begin{cases} \hat{y} & \text{if } y = 1, \\ 1 - \hat{y} & \text{if } y = 0, \end{cases}$$

where $\gamma > 0$ (e.g., 2) is the focusing parameter.

Mean-squared error (MSE) penalizes the squared difference between continuous predictions and true scalar targets, and is typically used for regression tasks such as IoU estimations or bounding box predictions:

$$L_{\mathrm{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (s_i - \hat{s}_i)^2.$$

where s_i is the true scalar target with prediction $\hat{s}_i$.

Backpropagation: Training minimizes loss $L(\theta)$ with respect to all parameters $\theta = \{ \mathbf{W}^{(l)}, \mathbf{b}^{(l)} \}_{l=1}^{L}$. Backpropagation computes the "error signal" $\delta^{(l)} = \partial L / \partial \mathbf{z}^{(l)}$ at each layer. Starting at the output,

$$\delta^{(L)} = \frac{\partial L}{\partial \mathbf{z}^{(L)}},$$

the model propagates backward for $l = L - 1, \ldots, 1$ by

$$\delta^{(l)} = (\mathbf{W}^{(l+1)})^T \, \delta^{(l+1)} \; \circ \; \phi'\big(\mathbf{z}^{(l)}\big),$$

where "$\circ$" denotes element-wise multiplication and ϕ' is the derivative of the activation. The gradients of the weights and biases then equal:

$$\nabla_{\mathbf{W}^{(l)}} L = \delta^{(l)} \left(\mathbf{h}^{(l-1)}\right)^T, \qquad \nabla_{\mathbf{b}^{(l)}} L = \delta^{(l)}.$$

Optimization: Parameters are updated by a gradient-based optimizer. The generic update to model weights is

$$\theta \leftarrow \theta - \alpha \nabla_\theta L$$

where $\alpha > 0$ is the learning rate (a hyperparameter) that can be applied in several different ways:

Batch gradient descent: the gradient $\nabla_\theta L$ is computed over the entire training set before each update.

Stochastic gradient descent (SGD): the gradient is approximated on a single example or a small "mini-batch" per update. This results in noisier steps and more frequent updates. A widely used version is **SGD with momentum**, which is used in this thesis, and smooths these updates by accumulating past gradients as follows:

$$v^{(t)} = \beta\, v^{(t-1)} + (1-\beta)\, \nabla_\theta L\big(\theta_{t-1}\big), \qquad \theta^{(t)} = \theta^{(t-1)} - \alpha\, v^{(t)},$$

where $\alpha > 0$ is the learning rate and $\beta \in [0, 1)$ is the momentum coefficient (a hyperparameter that is commonly $\beta = 0.9$). Each v ("velocity") term is a running average of past gradients, which smooths out the noise in the updates. More advanced optimizers like Adam [45] build on these ideas by also adapting α per-parameter.

2.2.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a class of deep learning models specifically designed to exploit the spatial structure of data, such as images. Each convolutional layer applies learnable kernels (of size $M \times N$) with configurable stride and padding, followed by a nonlinear activation (e.g., ReLU) and optional pooling for spatial downsampling. Networks are trained end-to-end via backpropagation using optimizers such as SGD or Adam.

Convolutional Layers: Let $\mathbf{I} \in \mathbb{R}^{H \times W}$ denote a single-channel input and $\mathbf{K} \in \mathbb{R}^{M \times N}$ a learnable kernel. The convolutional layer produces an output feature map $\mathbf{S} \in \mathbb{R}^{H' \times W'}$. With stride s

and padding P, the spatial dimensions of $\mathbf{S}$ are

$$H' = \left\lfloor \frac{H - M + 2P}{s} \right\rfloor + 1, \quad W' = \left\lfloor \frac{W - N + 2P}{s} \right\rfloor + 1.$$

Each entry of $\mathbf{S}$ at location (i, j) is given by the discrete convolution

$$S(i,j) \;=\; (\mathbf{I} * \mathbf{K})(i,j) \;=\; \sum_{m=1}^{M} \sum_{n=1}^{N} I(i+m-1,\; j+n-1)\; K(m,n) \quad [30],$$

which re-uses the same weights $\mathbf{K}$ at each spatial location, allowing for parameter sharing and translation equivariance.

Nonlinear Activation: Following each convolution, an element-wise nonlinearity ϕ is applied. The most common choice is the aforementioned Rectified Linear Unit (ReLU).

Pooling Layers: To reduce the spatial size of feature maps and approximate translational invariance, pooling layers (e.g., max- or average-pooling) summarize local neighborhoods [8].

Fully Connected Layers: After several stacks of convolution, activation, and pooling, the resulting feature maps are flattened into a vector $\mathbf{h}_{\text{flat}}$ and passed through one or more fully connected (dense) layers, which synthesize the spatial features into final class scores (or other outputs).

Optimization: The network's parameters θ (all convolutional kernels, biases, and dense-layer weights/biases) are learned by minimizing a suitable loss using gradient-based optimizers such as SGD or Adam.

An exemplary CNN architecture for 0-9 digit recognition is illustrated in Figure 2.2, where successive convolutional and pooling layers learn increasingly abstract representations before a final classification stage.

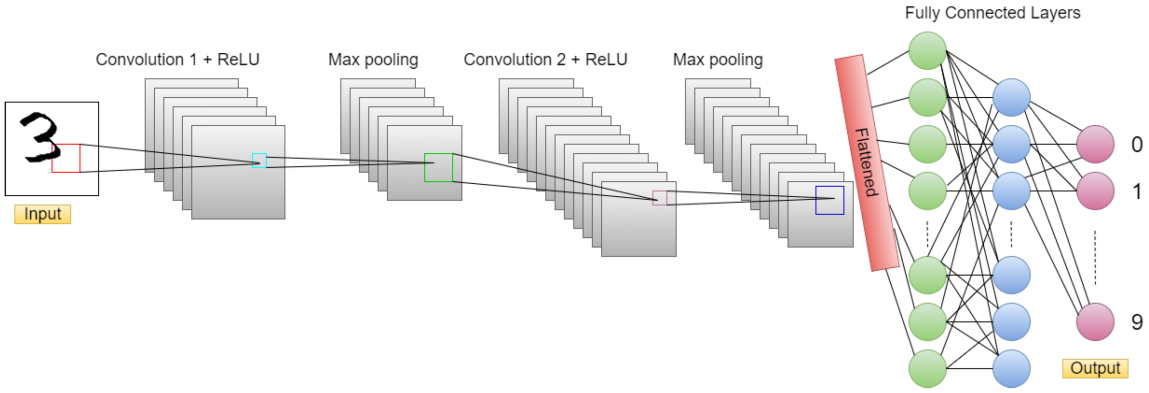


Figure 2.2: Illustrative diagram of a convolutional neural network used for 0-9 digit classification [4]. Convolutional layers extract local features, activation functions (e.g., ReLU) introduce nonlinearity, pooling layers (Max pooling) reduce spatial dimension, and fully connected layers produce classification outputs.

U-Net Architecture: Building on these CNN principles, the vanilla **U-Net** (see Figure 2.3) is first designed for medical image segmentation, and introduces a symmetric *encoder–decoder* ("U"-shaped) design with *skip connections* [77]. The encoder path applies convolutions and downsampling to extract high level, abstract features. The decoder path then upsamples and refines these features back to the original resolution, merging corresponding encoder activations via lateral skip-connections. Formally, if F_l is the feature map from encoder level l and H_{l+1} the upsampled decoder map from level $l+1$, the fused input to the next decoder block is

$$H_l = F_l \parallel H_{l+1},$$

where "$\parallel$" denotes channel-wise concatenation. These skip connections preserve details and improve gradient flow. At the network's output, a 1×1 convolution reduces the final decoder feature map to the desired number of channels (e.g.one for binary segmentation or C for multi-class). A pixel-wise activation (sigmoid for binary, softmax for multi-class), produces probabilities for each pixel in the mask. These probabilities are trained against the actual masks using losses like cross-entropy or Dice loss.

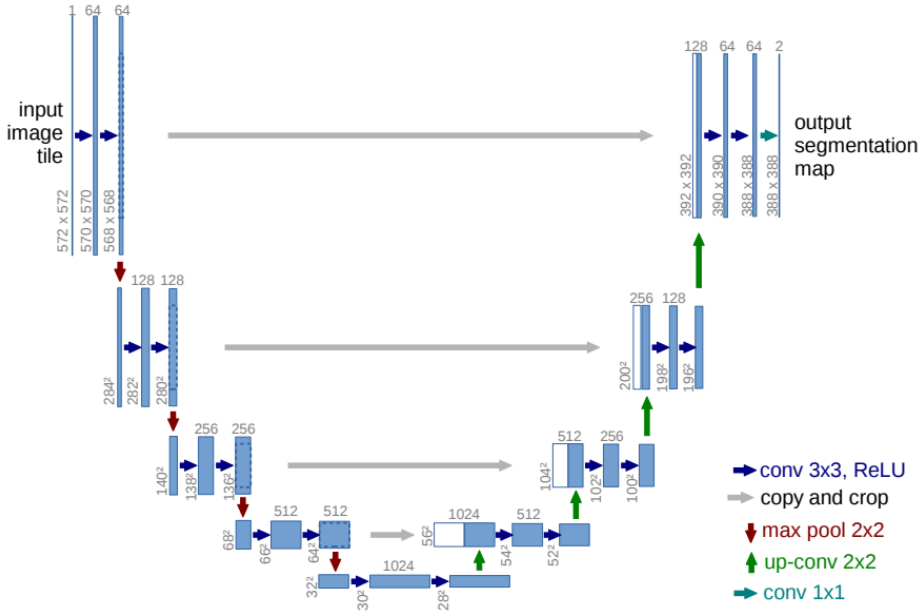


Figure 2.3: U-Net: a symmetric encoder–decoder network with lateral skip connections that concatenate high-resolution encoder features into the decoder, followed by a 1×1 convolution and pixel-wise activation to generate the final segmentation masks [97].

2.2.3 Transformer-Based Models: Vision Transformer and Swin Transformer

While CNNs capture local textures effectively, their fixed receptive field can limit modeling of long-range dependencies [91]. Transformers, on the other hand [89], utilize self-attention mechanisms that globally aggregate information, making them strong candidates for segmentation of structures that have long-range contextual cues. While originally developed for sequence modeling in natural language processing, the Vision Transformer (ViT) [22] demonstrated that an image can be treated as a sequence of patch embeddings, with global self-attention replacing convolutional feature extractors. In self-attention, given query Q, key K, and value V matrices derived from image patch embeddings, the attention output is:

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)\, V,$$

where d is the embedding dimension. This allows each patch token to attend to all others and effectively capture the global context. Cross-attention applies the same softmax-weighted query–key–value operation, but uses queries from one token sequence (e.g., prompt embeddings) and keys/values

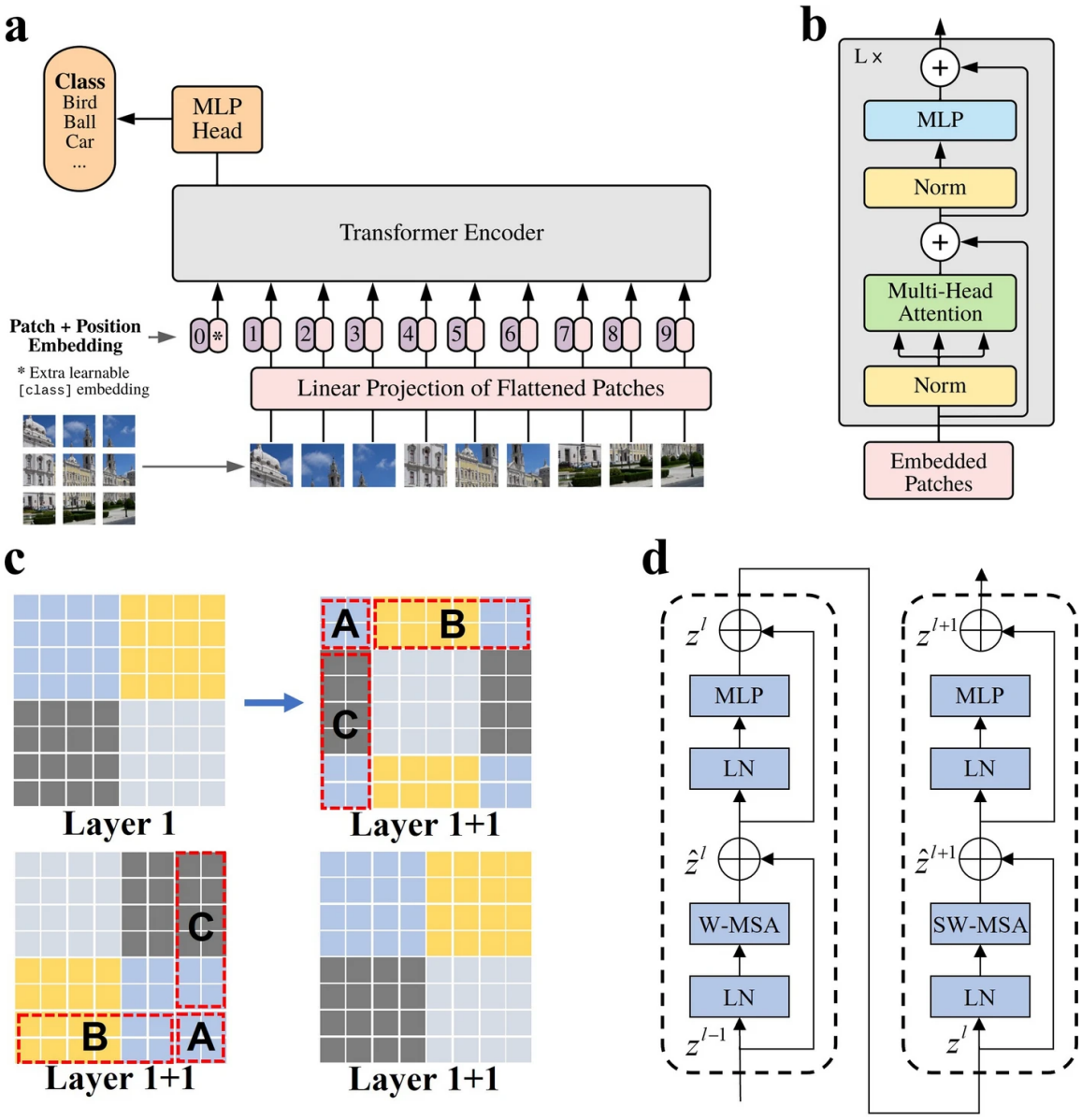


Figure 2.4: Comparison of Vision Transformer (ViT) and Swin Transformer architectures [70]. (**a**) Vision Transformer (ViT) pipeline: the input image is split into fixed-size patches, each flattened and linearly projected, then prepended with a learnable [class] token and positional embeddings before being processed by a stack of global self-attention encoder layers and an MLP classification head. (**b**) Standard transformer encoder block: LayerNorm $\rightarrow$ multi-head self-attention (MSA) with residual connection $\rightarrow$ LayerNorm $\rightarrow$ MLP with residual connection. (**c**) Swin Transformer's shifted-window mechanism. (**d**) Swin Transformer block modules: each stage alternates between a Window-based MSA (W-MSA) sublayer (left) and a Shifted-Window MSA (SW-MSA) sublayer (right), each wrapped with LayerNorm and followed by an MLP with its own residual path to capture both local and cross-window dependencies.

from a different sequence (e.g., image feature vectors), enabling information flow between any two sets of representations. In practice, Transformers use *multi-head attention* to jointly focus on different representation subspaces. Given learned projection matrices $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_k}$ for each head $i = 1, \ldots, h$ and an output projection $W^O \in \mathbb{R}^{hd_k \times d}$, the i-th head is

$$\text{head}_i = \text{Attention}(QW_i^Q,\ KW_i^K,\ VW_i^V),$$

and the multi-head output is

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\ W^O.$$

ViT models pretrained on very large-scale datasets (e.g., JFT-300M [85]) achieved state-of-the-art image classification performance [22]. Applying pure ViT architectures to segmentation, however, requires decoding their 1D sequence outputs back into 2D spatial maps. Early works like SETR simply reshaped the output embeddings to the image grid and topped them with a lightweight decoder [99], while Segmenter introduced both a point-wise linear decoder and a mask-transformer decoder [84]. However, these approaches struggle on smaller medical datasets because standard Transformers lack the locality bias of CNNs (assumption that nearby pixels are most relevant, enforced by small kernels) and the translation equivariance of CNNs (the property that shifting an object in the input yields a correspondingly shifted feature map, enforced by weight-sharing). Without these built-in priors, Transformers must learn patterns from scratch, greatly increasing their data requirements and tendency to overfit. As illustrated in Figure 2.4, the pure ViT model (panel a) treats the image as a sequence of patches with global self-attention (detail in panel b).

Swin-Unet [14] builds on the Swin Transformer backbone (see Figure 2.4c–d), which was designed to address two main drawbacks of standard ViTs: a lack of locality bias and the $O(N^2)$ cost of global self-attention. The Swin Transformer introduces *shifted window* multi-head self-attention (SW-MSA), where self-attention is computed only within small, non-overlapping windows of size $M \times M$, and successive layers "shift" the window partitioning to allow cross-window information flow. This strategy reinforces locality by restricting each head to a compact neighbourhood and reduces attention complexity to $O(N)$, proportional to the number of patches N [58]. Moreover,

the smaller windowed patches (e.g., 4×4) enable finer-grained feature extraction compared to the larger patches used in early ViTs (e.g., 16×16). Swin-Unet replaces both encoder and decoder with hierarchical Swin Transformer blocks. The encoder gradually downsamples the feature maps via patch merging layers, extracting multi-scale representations. The decoder symmetrically upsamples using patch-expanding layers, and lateral skip-connections fuse high-resolution encoder features with upsampled decoder activations. By combining the efficient, window-based attention of Swin with the U-shaped design, Swin-Unet captures both fine local detail and long-range context in a pure-transformer architecture.

Swin-UNETR [34] further extends Swin-Unet to volumetric data. It embeds non-overlapping 3D patches from multi-modal medical scans into tokens, feeds them through a hierarchical Swin Transformer encoder (with shifted-window MSA at each resolution), and restores the original volume resolution via a convolutional decoder linked by skip-connections. This hybrid design has been found to yield strong performance on 3D segmentation benchmarks such as brain tumor delineation [34].

Swin-HGI-SAM [75] is a weakly supervised framework that leverages hierarchical self-attention from a Swin Transformer trained on binary ICH classification. Instead of naively averaging attention maps, each head's weights are scaled by the gradient of the positive ICH score with respect to that head, producing *head-wise gradient-infused* self-attention maps. These maps are then upsampled and fused across the first three Transformer stages to generate a segmentation mask without any voxel-level annotations.

2.3 Foundation Models

Foundation models are large, general-purpose deep neural networks pretrained on extremely large datasets to learn diverse feature representations. Unlike smaller, traditional architectures, which are often developed and trained from scratch for a single task, foundation models are commonly adapted using fine-tuning, where pretrained weights are updated on a smaller task-specific dataset, or applied zero-shot, directly to new tasks without further training. This flexibility dramatically lowers data requirements and development time in task-specific applications.

2.3.1 Segment Anything Model (SAM) and Extensions

The Segment Anything Model (SAM) [46] is a promptable, general-purpose segmentation model with a backbone consisting of three main components: a large ViT-based image encoder, a prompt encoder for manually-selected user inputs, and a lightweight mask decoder with an auxiliary IoU prediction head. The model is trained end-to-end in a fully supervised manner on over a billion masks with a weighted focal + Dice loss (in a 20:1 ratio), and with mean-squared error on the IoU head. Figure 2.5 provides an overview of these components, and how they work together to produce masks.

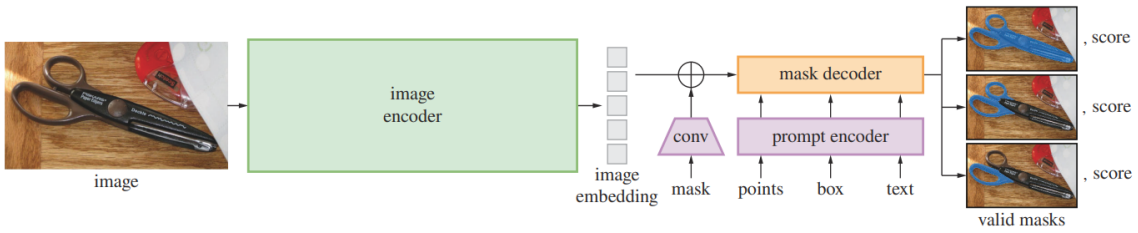


Figure 2.5: Architecture of the Segment Anything Model (SAM) [46]. SAM comprises three main components: (1) a ViT-based image encoder that converts the input image into a spatial feature map; (2) a prompt encoder that projects user interactions (points, bounding boxes or coarse masks) into prompt embeddings; (3) a lightweight mask decoder that fuses image features and prompt embeddings via cross-attention to produce a high-resolution segmentation.

Image Encoder SAM's image encoder E_{img} is a Vision Transformer (ViT) that processes an input image $X \in \mathbb{R}^{3\times H\times W}$ into a spatial feature tensor $F \in \mathbb{R}^{h\times w\times d}$. By using hierarchical patch embeddings and self-attention, this backbone captures local texture and global context across different scales.

Prompt Encoder: The prompt encoder E_{prompt} takes a variable set of sparse prompts (e.g., points/bounding boxes) or dense prompts (e.g., mask inputs) and projects them into a unified token space $\mathbb{R}^{k\times d}$. Each prompt type is first encoded into a fixed-length embedding:

$$P = E_{\text{prompt}}(\{\text{pts, boxes, masks}\}) \in \mathbb{R}^{k\times d},$$

where k depends on the number and kind of prompts. This design enables SAM to flexibly accept any combination of box, point, and mask prompts.

Mask Decoder: A lightweight mask decoder $\mathcal{D}_{\text{mask}}$ fuses image features F with prompt embeddings P via multi-head cross-attention, then uses mask heads to produce a high-resolution segmentation $\hat{Y} \in [0,1]^{H \times W}$, and an IoU head to predict mask quality (trained with mean-squared error against the true IoU).

Prompt Strategies: Effective segmentation with SAM requires choosing prompts that balance user effort and mask quality. Point prompts (positive to include mask, and negative to exclude mask) may produce coarse boundaries when used sparingly, while more points can sharpen more ambiguous edges. Bounding box prompts, which enclose the target object, guide SAM to focus on regions of interest. Hybrid prompts combine a bounding box with a few corrective points, using the box to define the search space and points to correct possible over- or under-segmentation.

MedSAM and SAM2 Innovations: Although zero-shot SAM generalizes well to natural images, medical domains benefit from specialized adaptations. MedSAM [61] fine-tunes the mask decoder on curated medical masks, significantly boosting performance on CT/MRI tasks. More recently, SAM2 [47] further extends SAM to sequential data by adding a memory encoder that propagates masks across video frames (or 3D image slices).

2.3.2 Detection Models for Prompt Generation: YOLOv8

YOLOv8 (see Figure 2.6) builds multi-scale feature maps (P5, P4, P3 for large, medium, and small objects) by fusing convolutional features whose spatial resolutions are 1/32, 1/16, and 1/8 of the input, respectively. Coarser maps are upsampled and added to finer ones so that each output combines local detail with broader context. At each spatial cell j the model predicts class confidence logits s_j and a representation of the bounding box offsets, from which the final box is decoded. YOLOv8 ships in five sizes, from nano to extra-large, with parameter counts rising in both the backbone and detection head, trading off inference speed for accuracy.

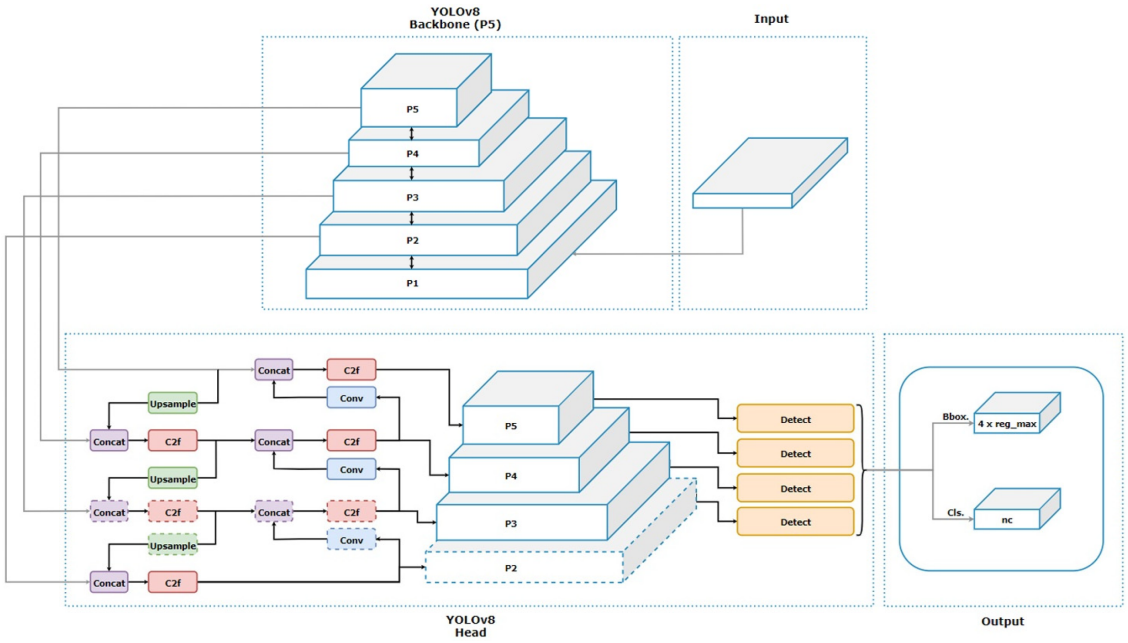


Figure 2.6: YOLOv8 architecture [65]. The backbone computes a full feature pyramid P1–P5, but only P5–P3 are used to produce outputs (P2 is produced by the backbone for completeness and contains fine-grained details, but is not routed into any output). P5–P2 are fused via upsampling, concatenation, and C2f (Conv) modules to yield three detection feature maps. Each map is fed into a 1×1-conv "Detect" head that predicts: (1) bounding box offsets per-class and (2) per-class confidence scores. At inference, these outputs are decoded into bounding boxes, filtered by confidence score, and post-processed with non-maximum suppression to produce the final detections.

Training minimizes a weighted sum of three losses (detection loss):

$$\mathcal{L}_{\text{det}} = B \cdot \left(\lambda_{\text{box}} L_{\text{box}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{dfl}} L_{\text{dfl}}\right),$$

where B is the batch size and $\lambda_{\text{box}}, \lambda_{\text{cls}}, \lambda_{\text{dfl}}$ are hyperparameter gains.

L_{cls} is a classification loss (binary cross-entropy) on the predicted class confidence, encouraging correct class predictions. L_{box} is a localization loss based on Complete IoU (CIoU) [100], encouraging predicted boxes to align tightly with ground truth. L_{dfl} is Distribution Focal Loss [52], which refines bounding box regression by predicting each box-side offset as a discrete distribution.

The CIoU loss combines bounding box overlap, center distance, and aspect ratio consistency. For a predicted box b and ground truth b^*,

$$L_{\text{CIoU}} = 1 - \text{IoU}(b, b^*) + \frac{\rho^2}{c^2} + \alpha\, v,$$

where IoU is the intersection–over–union, ρ^2 is the squared center-point distance, c^2 the squared diagonal of the smallest enclosing box, and

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^*}{h^*} - \arctan\frac{w}{h}\right)^2$$

measures the squared difference in aspect ratios (with w, h and w^*, h^* the predicted and true box width/height, respectively). A trade-off weight $\alpha = \frac{v}{(1-\text{IoU})+v}$ then scales this term based on how large v is relative to the remaining IoU gap.

Distribution Focal Loss (DFL): Instead of predicting each box-side offset as a single real value, DFL treats it as a small classification over integer "bins" (pixel displacements). For side $d \in \{\text{L}, \text{T}, \text{R}, \text{B}\}$ (left, top, right, bottom), let the true offset t_d be, for example, 3.7 pixels from the cell center. We then create a "soft" target that assigns 30% weight to bin 3 and 70% to bin 4 (we linearly interpolate between the two nearest integers). The model outputs a probability for each bin, and we use cross-entropy against this 30/70 target. By framing regression as a tiny two-class (binary) problem per side, DFL reduces quantization error (rounding a continuous value to a discrete

one) and produces sharper, more accurate box boundaries. Mathematically, this can be written as:

$$t_d = \ell + \delta, \quad \ell = \lfloor t_d \rfloor,\ \delta = t_d - \ell, \quad w_\ell = 1 - \delta,\ w_{\ell+1} = \delta,$$

$$L_{\mathrm{DFL},d} = -\big(w_\ell \log \hat{P}_{d,\ell} \ + \ w_{\ell+1} \log \hat{P}_{d,\ell+1}\big), \qquad L_{\mathrm{dfl}} = \tfrac{1}{4} \sum_{d \in \{L,T,R,B\}} L_{\mathrm{DFL},d}.$$

At inference, each spatial cell on P3–P5 yields two outputs: (1) bounding box offsets per-class and (2) per-class confidence scores. These are decoded into bounding boxes, filtered by confidence thresholds, and non-maximum suppression removes overlaps. The remaining boxes can then be used directly as prompts for SAM or converted to point prompts at their centers.

2.3.3 BiomedParse: A Medical Vision–Language Foundation Model

BiomedParse is a unified biomedical foundation model that can generate segmentation masks, bounding boxes, and labels directly from text prompts, pretrained on over 6 million image–mask–text triples. It uses a dual-encoder architecture, with a Focal Transformer backbone (FocalNet) [94] as its image encoder, and PubMedBERT [32] (a BERT variant pretrained on biomedical literature) as its text encoder.

Focal Transformer backbone [94]: FocalNet first splits the input image into non-overlapping patches (patch size = 4), linearly embedding each into a d-dimensional token. It then applies four hierarchical stages (with depths [2,2,6,2]) of *focal self-attention* transformer blocks interleaved with patch-merging downsampling. Focal self-attention attends densely within a local $k \times k$ window and sparsely to progressively pooled "focal" keys at further radii, reducing complexity from $O(N^2)$ to nearly linear while preserving both fine detail and long-range context. At the end of each stage, a LayerNorm produces feature maps, denoted `res2`, `res3`, `res4`, and `res5` from low-level edges/texture to high-level semantic object representations (since each patch-merging stage halves height/width and doubles channels).

PubMedBERT text encoder: PubMedBERT [32] is a domain-adapted BERT (Bidirectional Encoder Representations from Transformers) [21] model pretrained from scratch on millions of PubMed

abstracts and full texts. It retains BERT's architecture (12 Transformer layers, hidden size 768, 12 attention heads) but specializes in medical terminology and syntax. Both the image and text features are projected via small MLPs into a shared d-dimensional feature space for decoding.

Decoder head: Multi-scale features `res2–res5` are merged and upsampled by a pixel decoder into a unified high-resolution embedding. A Transformer predictor then uses cross-attention with object and text queries to output mask logits (upsampled to the input size), bounding-box coordinates, and class scores. During training, a multi-task loss jointly optimizes high-level concept classification, mask supervision and overlap (Dice) objectives:

$$\mathcal{L} = \alpha\, L_{\text{CE}} \;+\; \beta\, L_{\text{BCE}} \;+\; \gamma\, L_{\text{Dice}} \tag{1}$$

At inference, given an input image and an (optional) text prompt, the decoder heads produce: (1) a pixel-wise segmentation mask for the prompted concept, (2) one or more bounding boxes with associated confidence scores for each detected instance of that concept, and (3) a semantic label (class name) for each mask or box.

On a test set of 102,855 triples across nine modalities, BiomedParse outperforms specialized segmentation, detection, and recognition benchmarks [98].

2.3.4 Retrieval-Augmented Generation for Contextual Guidance

Large language and vision–language models excel at generating fluent, context-aware text, but they can hallucinate, inventing incorrect details or failing to reflect up-to-date facts. *Retrieval-Augmented Generation (RAG)* mitigates this by letting the model consult an external "memory" of real examples (e.g., knowledge database) at inference time [51]. RAG proceeds in two phases

1. **Retrieve:** map the user's input to a fixed-length *query vector* $\mathbf{x}_q \in \mathbb{R}^d$, then search a knowledge database of precomputed *document vectors* $\{\mathbf{d}_i\} \subset \mathbb{R}^d$ for the k most similar items.

2. **Generate:** pass both the original input and the retrieved documents to the generative model, prompting it to ground its output in that evidence.

By anchoring generation in real data, RAG reduces hallucinations and allows the retrieval index to be updated independently of the model.

Similarity Metrics: Retrieval can use any similarity function $\mathrm{sim}(\mathbf{x}_q, \mathbf{d})$ that measures how "close" the feature vectors of the query and knowledge base items are. Common choices include:

$$\mathrm{cosine}(\mathbf{x}_q, \mathbf{d}) = \frac{\mathbf{x}_q^\top \mathbf{d}}{\|\mathbf{x}_q\|\|\mathbf{d}\|}, \quad \mathrm{inner\_product}(\mathbf{x}_q, \mathbf{d}) = \mathbf{x}_q^\top \mathbf{d} = \sum_{j=1}^{d} x_{q,j}\, d_j.$$

In our proposed system in Chapter 4, we use the exact inner-product scores, which after normalizing to unit length correspond to the cosine similarity.

Exact Search with FAISS IndexFlatIP: For fast, exact inner-product search over up to hundreds of thousands of feature vectors in the knowledge base, we employ FAISS's [23] `IndexFlatIP` for our system in Chapter 4. This index stores every $\mathbf{d}_i$ directly in memory and, given a query $\mathbf{x}_q$, computes

$$[\, \mathbf{x}_q^\top \mathbf{d}_1, \ldots, \mathbf{x}_q^\top \mathbf{d}_N \,],$$

then returns the top-k scores. Although this brute-force approach requires $O(Nd)$ work per query, it remains sub-millisecond for moderate database sizes on modern hardware, especially using high-level, lower-dimensional representations (e.g., res5 outputs).

For much larger collections, FAISS also provides approximate indexing, such as inverted-file with product quantization (IVFPQ), small-world graph (HNSW), or scalar quantization variants, which all trade a some accuracy for faster, more memory-efficient search. Overall, at query time, RAG systems execute:

$$\underbrace{\text{Encode Input}}_{\to\, \mathbf{x}_q} \xrightarrow{\text{Search}} \underbrace{\{\mathbf{d}_{i_1}, \ldots, \mathbf{d}_{i_k}\}}_{\text{top-}k} \xrightarrow{\text{Prompt Model}} \underbrace{\text{Grounded Response}}_{\text{evidence-based}}.$$

2.4 Virtual Reality Agents for Medical Image Segmentation

2.4.1 Virtual Reality and Mixed Reality

Virtual reality (VR) creates a fully immersive, computer-generated environment in which users can look, move, and interact as if they were present in another space. VR is often discussed as one extreme of the *Reality-Virtuality Continuum* that was introduced by Milgram et al. [63], with the physical real world at the opposite end and various mixes of real and virtual elements (augmented reality, augmented virtuality) in between (see Figure 2.7). In augmented reality (AR), digital content is overlaid onto the real world (e.g., projecting patient anatomy onto a surgeon's view), whereas in augmented virtuality (AV), real-world elements are integrated into a predominantly virtual environment. Between the two extremes of the continuum, all the rest falls under mixed reality (MR), where real and virtual components coexist and interact. In practice, modern VR is delivered through head-mounted displays (HMDs) and natural input devices (controllers or body-tracking), which allow for sensory immersion. This immersive quality can be advantageous in the clinic, where understanding complex 3D anatomy in an unobstructed 3D workspace (with no occlusion from the real world) is critical.

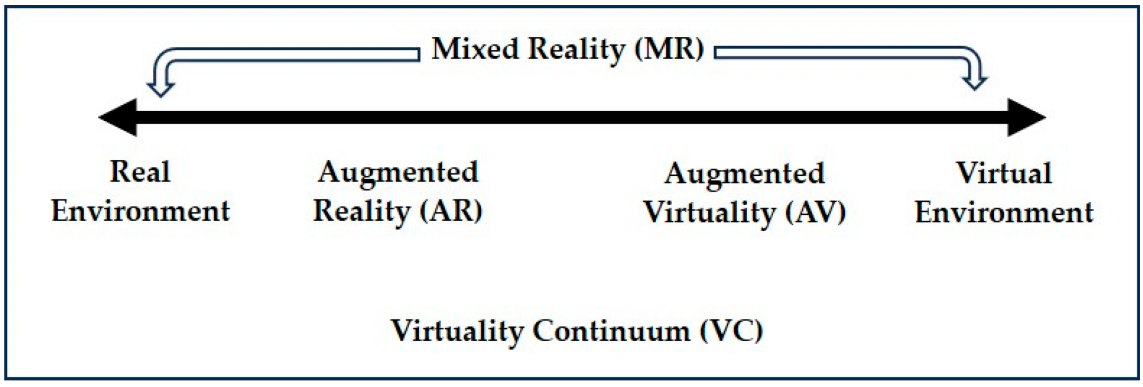


Figure 2.7: The Virtuality Continuum (VC) [63], showing the spectrum from Real Environment to fully Virtual Environment, all encompassed under Mixed Reality (MR) [25].

2.4.2 Medical Image Segmentation in VR

Conventionally, 3D medical image segmentation is performed in a slice-by-slice manner on 2D displays, making it challenging to directly draw or verify 3D geometry of the target anatomical

structure. VR could address this by allowing users to interact with medical images in their native 3D context. For example, Gonzalez et al. developed the *NextMed* platform [39], which automatically segments certain anatomies and renders a 3D model in AR or VR for clinicians to explore. They argue that working directly with life-sized 3D models can facilitate pre-surgical planning and education, and in trials on over 1,000 medical images, *NextMed*'s approach significantly improved doctors' ability to work with and understand patient-specific anatomy. Other researchers have explored more interactive segmentation, such as Gao et al.'s *NUI-VR*^{2} system, which lets users perform 3D volume segmentation in VR using freehand gestures and voice commands instead of a mouse [28] (see Figure 2.8) [28]. This "natural user interface" approach effectively removes the 2D input bottleneck, allowing users to intuitively mark regions in mid-air and even step inside volumes for a better view. Compared to traditional mouse inputs, the gesture-based interface enabled much more precise and complete annotations, with significantly higher segmentation accuracy (i.e., Dice scores). Overall, these systems demonstrate that when the system can employ natural behaviors in VR, their segmentation performance and satisfaction can exceed those from traditional software interfaces.

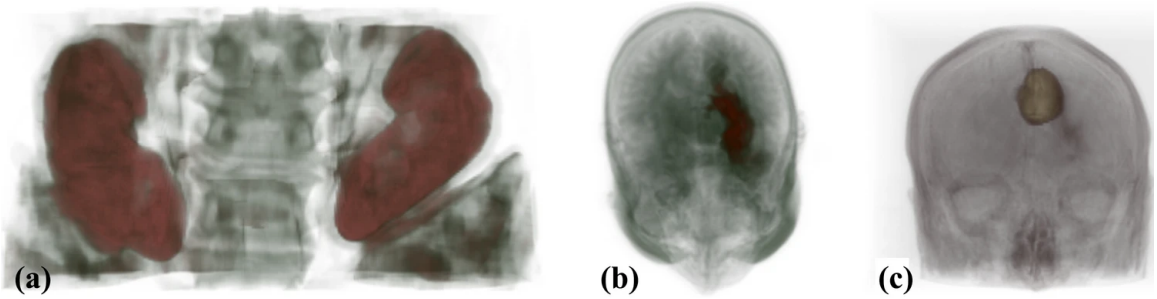


Figure 2.8: VR-based 3D segmentations generated in NUI-VR^2 [28]. Users place 3D "seeds" inside the target region via freehand gestures to create masks. (a) kidneys; (b) left hippocampus; (c) brain tumor.

2.4.3 Interaction Paradigms for Immersive Medical Image Segmentation

Interaction paradigms in VR are an important human-computer-interaction (HCI) research topic, especially applied to medical tasks, where users may need to switch between close-up mask editing/drawing and distant menu interactions. Modern VR head-mounted displays support a range

of input modalities, including handheld controllers, head gaze, and eye tracking, each with unique advantages. In general purpose VR interaction studies, controller-based pointing is often the most precise method for object selection, providing tactile feedback and fine control, whereas gaze-based aiming can be faster, but may suffer from jitter and selection errors [60]. Head-directed pointing (orienting one's head like a laser pointer) requires no extra hardware and imposes minimal physical effort, though it tends to be slightly slower and less precise than using a handheld device [93]. In the medical image segmentation context, these trade-offs become critical, particularly if users must frequently switch their attention between adjusting masks and interacting with a tool-bar or menu interface. This thesis investigates the optimal paradigm that can minimize the friction during such "near-far" attention switching. While Hellum et al. evaluated controller, head-gaze, and eye-tracking for neurosurgical data navigation and annotation in VR [36], no prior study has examined these paradigms specifically for immersive medical image segmentation that involves attention switching between proximal image display and a distal menu interface.

2.4.4 Conversational Assistants vs. Autonomous Agents

Recent advancements in conversational AI have seen the introduction of AI-driven *assistants* in immersive environments, primarily for educational or informational purposes. For example, Liaw et al. developed a voice-based AI assistant for VR sepsis training simulations, which matched the performance of human instructors and led to significantly higher test scores for trainees [53]. Similarly, Chheang et al. introduced a generative AI tutor for anatomical education in VR [18], demonstrating the potential virtual assistants have to support medical learning through dialog. These early systems leverage large language models to provide information and guidance to the user via conversation. However, these assistants provide answers and instructions without actively performing tasks within the virtual environment. In contrast, a true *agent* is more than a conversational assistant: it is an entity that perceives its environment and is capable of executing multi-step tasks in response to high-level user commands with minimal human oversight. Ideally, such agents would interpret user requests (e.g., "locate and segment the tumor"), autonomously execute relevant actions using embedded segmentation algorithms, and dynamically refine results based on user interactions. Furthermore, agents often use Retrieval-Augmented Generation (RAG), allowing them to dynamically

access external, task-relevant knowledge sources. This is particularly useful in medical contexts, where up-to-date case-specific data or reference examples can be essential for accurate segmentation and decision-making support.

Despite advances in conversational assistants, no existing VR system currently provides an agent for medical image segmentation. Our work aims to bridge this gap by developing the first immersive, conversational segmentation agent, designed to minimize manual segmentation burdens, maintain expert oversight, and provide information using RAG to support medical segmentation education.

2.4.5 Evaluation of Usability and Human Factors

Evaluating an immersive VR system, particularly one that integrates a segmentation agent, requires careful consideration of task performance and users' subjective experiences. Two widely accepted instruments for this purpose are the System Usability Scale (SUS) [10] and the NASA Task Load Index (NASA-TLX) [33] questionnaires. The SUS contains ten statements regarding the system's ease of use, learnability, and integration of functions, where after interacting with the system, users indicate their agreement with each statement on a five-point scale. The final SUS is a summary of all ten responses, which add up to 100. A SUS score above 68 would indicate that the software system under study is sufficiently intuitive and approachable [9]. The NASA-TLX, on the other hand, measures perceived workload across six dimensions: mental demand, physical demand, temporal demand, effort, performance satisfaction, and frustration, each rated on a 0 to 20 point scale. NASA-TLX reveals whether the system imposes excessive cognitive strain or frustration.

Chapter 3

Weakly Supervised Intracranial Hemorrhage Segmentation with YOLO and an Uncertainty Rectified Segment Anything Model

A version of this chapter has been presented at the Stroke Workshop on Imaging and Treatment CHallenges (SWITCH) at the Medical Image Computing and Computer Assisted Interventions (MICCAI 2024) Conference. The conference proceedings has been published in *Image Analysis in Stroke Diagnosis and Interventions*:

- Spiegler P, Rasoulian A, Xiao Y. Weakly supervised intracranial hemorrhage segmentation with YOLO and an uncertainty rectified segment anything model. *Image Analysis in Stroke Diagnosis and Interventions. ISLES SWITCH 2024*; Springer; 2024. LNCS 15408, pp. 12–21 [82].

3.1 Introduction

Intracranial hemorrhage (ICH) accounts for 10-15% of all stroke cases and carries a significant risk of mortality [38]. Hemorrhage volume, which can rapidly expand within the first few hours, is a key predictor of treatment outcomes and potential complications [71]. Precise localization and quantification of the five ICH subtypes, including intraventricular (IVH), intraparenchymal (IPH), subarachnoid (SAH), epidural (EDH), and subdural (SDH), are therefore essential for tailoring treatment strategies and minimizing adverse events [3]. While supervised deep learning (DL) models have demonstrated excellent potential in automating ICH assessment [35], their success heavily relies on large datasets with pixel-level annotations (ground-truth masks) and poor segmentation accuracy is observed with smaller training datasets [37]. However, large training datasets containing high-quality ground-truth masks are difficult to obtain due to high demands in time, labor, and domain expertise. Together with scarce public ICH segmentation datasets, this bottleneck poses great challenges in developing automatic ICH quantification algorithms to better facilitate the care and management of the condition.

To overcome the aforementioned issue, weakly supervised learning approaches [75, 74] have emerged as a promising alternative. These methods leverage more economic ground truths, such as categorical labels, bounding boxes, or coarse masks to train segmentation models, bypassing the requirement of refined masks for fully supervised and semi-supervised approaches. While most existing literature is dedicated to ICH detection, ICH segmentation using weakly supervised methods remains under-explored. However, limited prior explorations exist leveraging explainable AI methods for weakly-supervised stroke segmentation, including class-activation maps (CAM) [92] and self-attention maps [75], providing encouraging results. Recent developments in foundation models, such as the Segment Anything Model (SAM) [46] have shown great potential to mitigate the segmentation ground truth bottleneck, but have not been explored for improving weakly supervised ICH segmentation. Therefore, we propose a novel weakly supervised ICH segmentation technique that incorporates automatic box and point prompt generation with SAM to allow for ICH detection and segmentation on CT scans. We have three main contributions. **First**, we leveraged a finetuned YOLOv8 model and a novel morphology-based method to automatically generate box

and point prompts, respectively, for SAM. **Second**, to enhance segmentation accuracy with SAM, we employed an uncertainty rectification approach to account for uncertainty in prompt generation. **Lastly**, we explored the impacts of different prompt types for our proposed framework in ICH segmentation and compared it against state-of-the-art (SOTA) supervised and weakly supervised techniques.

3.2 Related Works

ICH segmentation methods still primarily rely on fully supervised approaches [50, 15, 49, 19] and often with in-house datasets. More recently, semi-supervised techniques [90] have also been proposed for ICH quantification. However, refined segmentation ground truths are still crucial for their success, and more practical weakly supervised methods are gaining interest. In the limited prior works in this direction, most have relied on categorical labels as weak ground truths. For example, Wu et al. [92] proposed to use refined CAM results and representation learning to achieve ischemic stroke lesion segmentation, achieving a 0.3827 mean Dice score on multi-spectral MRIs. Later, from a binary classification CNN, Nemcek et al. [67] detected the location of ICH as bounding boxes in axial brain CT slices using the local extrema of derived attention maps, with a mean Dice of 0.58 for the lesion bounding boxes. Recently, Rasoulian et al. [75] utilized Head-Wise Gradient-Infused Self-Attention Maps from a Swin Transformer (Swin-HGI-SAM) trained on binary labels (ICH vs. no ICH) to obtain ICH segmentation, which obtained a mean Dice score of 0.438 on CT scans. The recent introduction of SAM [46], which allows interactive prompting in the forms of bounding boxes and/or points for zero-shot segmentation has attracted significant attention. However, its performance on CT-based ICH quantification and as an integrated solution allowing full automation in weakly supervised segmentation is yet to be explored. Furthermore, YOLO models [86] have been employed for ICH detection [24], but no reports have investigated their potential to facilitate the automation of SAM in ICH segmentation thus far.

3.3 Methods and Materials

3.3.1 Dataset and Preprocessing

For our study, we used the public Brain Hemorrhage Extended (BHX) dataset [76], which includes bounding box annotations for ICH along with their corresponding lesion subtypes, and the manually labeled PhysioNet CT dataset [37], which includes manual ICH segmentations. While 4607 CT slices and 5543 bounding boxes from the BHX dataset (containing the ICH subtypes and healthy scans) were employed to train and validate the YOLO model for lesion bounding box detection, the PhysioNet ICH segmentation dataset, which has 2814 CT slices with 318 mask-annotated ICH slices, was reserved as an independent test set to evaluate ICH segmentation with SAM. In addition, for our selected fully supervised baseline methods (more details in Section 3.3), subject-wise five-fold cross-validation was used on the manually segmented PhysioNet dataset to provide segmentation results for all cases and ensure that no slices from the same subject exist across different folds. As CT scans typically have a high dynamic range, for each CT slice, brain, subdural, and bone windows were created based on previous guidelines [27] and stacked together to form a composite RGB image, which was normalized to the range of [0,1] in each channel to facilitate training.

3.3.2 Uncertainty-Rectified YOLO-SAM Models

We propose YOLO-URSAM, a novel weakly supervised framework for ICH segmentation, where the YOLOv8 model [86] provides several prompts for SAM to perform ICH segmentation. Here, we built three YOLO-SAM variants, including YOLO-SAM-BBox, YOLO-SAM-Point, and YOLO-SAM-PointBBox, which perform ICH segmentation using bounding box prompts, point prompts, and combinations of bounding boxes and point prompts, respectively. These models each employ an uncertainty rectification strategy that combines 10 SAM outputs based on their 10 respective perturbed prompts. The detailed procedure of our methods is described below and shown in Fig. 3.1.

YOLO Detection: The preprocessed CT slices are passed to YOLOv8, which outputs the bounding boxes and associated lesion types for detected ICH. Then, the corner coordinates of the predicted

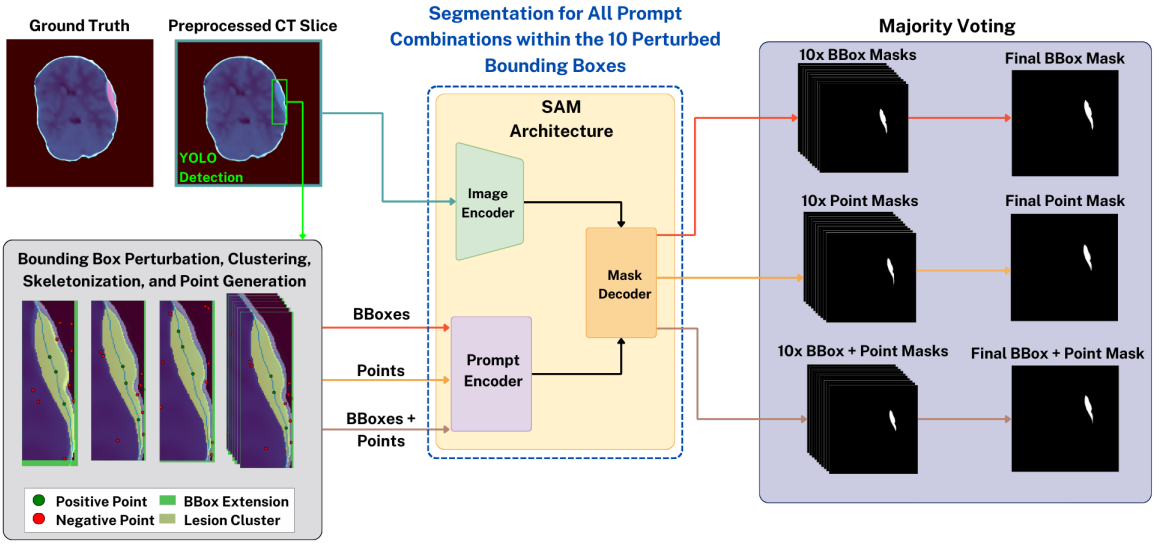


Figure 3.1: Workflow of the proposed weakly supervised ICH segmentation method.

bounding boxes are recorded to serve as the basis for automatic point prompt generation.

Bounding Box Perturbations: To enhance segmentation robustness and facilitate downstream uncertainty rectification in SAM's outputs, we introduce a method involving bounding box perturbation. Specifically, each YOLO-predicted bounding box is perturbed 10 times by randomly increasing its size by 1-4 pixels on each side. These perturbed boxes are recorded for the next step.

Clustering and Point Prompt Generation: Next, to strengthen the prompts' efficacy for SAM, leveraging the predicted ICH bounding box from YOLO, we introduce a novel point prompt generation method for the lesion and background based on a tailored tissue clustering solution and morphological analysis. To delineate lesions in proximity to the skull (e.g. SDH hemorrhage) for SAM, BET [81] skull-stripping is first applied to the entire CT image. Then, within the ICH bounding box for the skull-stripped RGB composite CT slice, K-means clustering is applied for tissue classification. Here, we use four clusters ($K = 4$) regardless of hemorrhage sub-types. If any residual skull is present, in decreasing order of the Hounsfield unit (HU) value, we must account for 1) residual skull tissue 2) ICH 3) healthy brain tissue and 4) dark background; if not, we can expect the brightest cluster to be assigned to the lesion and the remaining 3 clusters to be assigned to the rest. Then, an algorithm is devised to automatically identify the lesion cluster out of the four

(YOLO-Clustering). The resulting simple tissue clustering is obtained by first inspecting whether the cluster with the highest average HU value corresponds to the brightest signals in the bone window channel, which represents the residual bone. If not, the cluster is selected as the lesion cluster. Otherwise, the algorithm picks the cluster with the second-highest average HU value. Finally, on the K-means-based lesion clusters, skeletonization is performed to extract the skeleton of the shapes. From these skeletons, positive ICH point prompts are sampled. Then, from each of the three other clusters, negative points are sampled for SAM segmentation.

SAM Segmentation with Uncertainty Rectification: For each of the 10 perturbed bounding boxes, each combination of generated prompts (bounding box, points, and point-box) are passed to SAM's prompt encoder along with the input image to produce a segmentation sample (Fig. 3.1). For each YOLO-SAM variant, their final segmentation is obtained via majority voting based on 10 segmentation samples from the associated prompt type. This voting mechanism ensures the robustness of ICH segmentation against network-related prompt instability and SAM's potential sensitivity to these variations, further improving segmentation quality.

3.3.3 Baseline Models and Ablation Study

To validate our proposed method, we compared its performance against the SOTA weakly supervised and fully supervised segmentation techniques for ICH segmentation. With an open-source repository and good performance, we chose the recent Swin-HGI-SAM [75] as our weakly supervised baseline. In terms of baseline methods with full supervision, we selected the popular UNet [77] and Swin-UNETR models [34], which have demonstrated strong performance in a wide range of medical image segmentation tasks. For the UNet model, we implemented the architecture from the manually segmented PhysioNet CT data paper [37], with four hierarchical layers in the encoding and decoding paths. For the Swin-UNETR model [34], we also adopted four hierarchical levels to be consistent with the UNet model.

While the SAM model [46] allows both bounding boxes and/or points as interactive prompts to generate segmentation results, the robustness and accuracy of individual prompt types and their combined usage still require further investigation. Therefore, besides comparison with the baseline models, we also performed an ablation study on the impact of prompt types for the target task

(YOLO-SAM-Point, YOLO-SAM-BBox, YOLO-SAM-PointBBox).

3.3.4 Model Training & Evaluation Metrics

The YOLOv8-m model pretrained on the MS COCO dataset [54] in our YOLO-SAM variants was finetuned on the BHX dataset [76], with 3685 CT-slice images and 4479 bounding box labels for the training set, as well as 922 CT-slice images with 1064 labels for the validation set. We used the default YOLOv8 configuration (batch size=16, patience=100) during training. For the first 10,000 iterations, the AdamW optimizer was used with a learning rate of 0.00111 (calculated by a fitting equation using the number of bbox classes, which was 5 for each ICH subtype) and a momentum of 0.9. For the remaining iterations (after epoch 44), the SGD optimizer with an initial learning rate of 0.01 and momentum of 0.9 was used. We trained the weakly supervised Swin-HGI-SAM model [75] with the RSNA 2019 Brain CT hemorrhage dataset [27] (90%:10% data split for training vs. validation) following the details from the original publication. As for the supervised baselines (UNet and Swin-UNETR), subject-wise five-fold cross-validation was employed exclusively on the manually segmented PhysioNet dataset, using the AdamW optimizer with an initial learning rate of 0.001 as well as a loss function based on Dice coefficient and cross-entropy. We conducted all model training on a desktop computer with an Intel Core i9 CPU and an NVIDIA GeForce RTX 3090 GPU. *After model training, all evaluations were based on the manually segmented PhysioNet dataset in a slice-wise manner using the default YOLO confidence threshold of 0.25*. As ICH detection is a crucial component of our method, besides segmentation, we also evaluated the binary ICH detection performance (ICH vs. no ICH) for all DL models with accuracy, precision, recall, AUC, F1-score, and specificity. Note that a YOLO prediction was considered a true positive if it correctly identified a slice containing ICH, irrespective of the predicted subtype. For UNet and Swin-UNETR, a true-positive detection was defined as a slice with ICH segmentation that contains more than 10 pixels. In terms of segmentation, we computed the Dice coefficient and Intersection over Union (IoU) for all proposed and baseline models. Paired two-sample t-tests were then used to compare the Dice and IoU scores between the proposed method and the baselines, with $p < 0.05$ indicating a statistically significant difference.

3.4 Results

3.4.1 Detection Performance

The ICH detection performance for all models is listed in Table 3.1. Similar to Swin-HGI-SAM, the YOLOv8-m model demonstrated superior detection performance for most metrics compared to the U-Net and Swin-UNETR models, particularly in precision (0.665 vs. 0.239 and 0.253), AUC (0.796 vs. 0.757 and 0.764), F1-score (0.645 vs. 0.373 and 0.374), and specificity (0.966 vs. 0.612 and 0.622). This highlights the potential of using bounding box localization models such as YOLO to achieve superior performance compared to mask-trained approaches on limited data (UNet, Swin-UNETR) and competitive performance with models trained on substantially more binary labels (Swin-HGI-SAM). However, a weakness of the YOLOv8-m model is its lower slice-wise recall compared to Swin-HGI-SAM (0.626 vs. 0.791), indicating that Swin-HGI-SAM will more reliably detect true positives. For all other detection metrics, YOLOv8-m demonstrated comparable but marginally weaker detection performance, likely due to the smaller number of training samples (4607 bounding box annotated CT slices for YOLO versus 677523 binary-labelled CT slices for Swin-HGI-SAM).

3.4.2 Segmentation Performance

The ICH segmentation results are shown in Table 3.2, with qualitative outcomes demonstrated in Fig. 3.2. Table 3.2 shows that point prompts, hybrid point and bounding box prompts, as well as

Table 3.1: Detection Performance of Different Methods

Metric	Swin-HGI-SAM	U-Net	Swin-UNETR	YOLOv8-m
Accuracy	0.950	0.647	0.655	0.933
Precision	0.765	0.239	0.253	0.665
Recall	0.791	0.901	0.907	0.626
AUC	0.880	0.757	0.764	0.796
F1-score	0.767	0.373	0.374	0.645
Specificity	0.969	0.612	0.622	0.966

simple tissue clustering within the YOLO bounding box (YOLO-Clustering) yielded significantly higher segmentation performance than Swin-HGI-SAM, UNet, Swin-UNETR and using bounding box prompts alone ($p < 0.005$). It is also shown that hybrid prompts have improved performance over point prompts and YOLO-Clustering on average, though not statistically significant ($p > 0.05$). While YOLO-Clustering had good segmentation quality, it also had higher standard error than SAM with hybrid and point prompts, highlighting the point prompt's better precision and reliability. Finally, while YOLO-SAM-BBox does not show significantly higher Dice and IoU scores than UNet ($p = 0.0853$) or Swin-UNETR ($p = 0.768$), it significantly outperforms Swin-HGI-SAM ($p < 0.005$).

3.5 Discussion

Our YOLO-SAM framework that integrates YOLOv8-m, a novel point-prompt generator, and SAM with uncertainty rectification has demonstrated great performance in weakly supervised ICH segmentation, particularly with the hybrid prompts. The superior performance over existing weakly supervised and fully supervised methods can be explained by the incorporation of the power of the foundation models and spatial information represented by the bounding box ground truths. It is important to acknowledge that the poor performance of fully supervised DL models, such as UNet and Swin-UNETR can also be partially due to the low number of ground-truth mask labels. Despite this success, the slice-wise recall metric for our YOLO model lagged behind the Swin-HGI-SAM, suggesting a potential compromise in the model's ability to detect all ICH slices. However, after investigating this further on a patient-wise basis, the recall metric was calculated at 0.9714, with 34 out of 35 patients with hemorrhage having had at least one slice detected. In a clinical setting, the proportion of true positive ICH cases would therefore be much higher than the reported slice-wise recall metric. Our ablation study showed that hybrid prompts offered better performance than points or bounding boxes. This observation echoes previous reports [61] and could be explained by the lack of robustness when capturing thin, elongated, and curved structures (e.g., IPH subtype) with bounding boxes by SAM. Finally, while MedSAM [61] has gained great popularity in the community, its adoption in our YOLO-MedSAM-BBox model resulted in inferior segmentation outcomes (Dice= 0.412±0.018, IoU=0.298±015). This is consistent with other reports of SAM

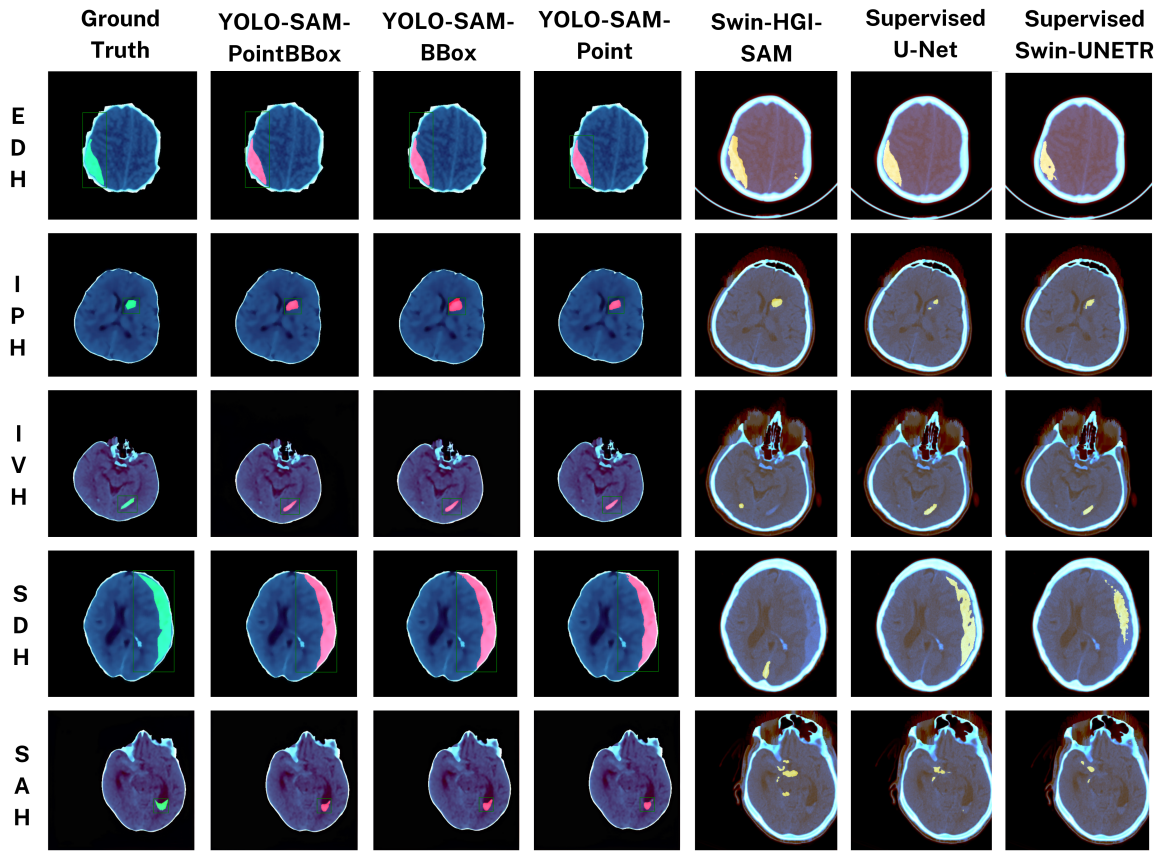


Figure 3.2: Qualitative segmentation results on different ICH subtypes

Table 3.2: Segmentation Performance of Different Models (mean ± standard error)

Model	**Dice**	**IoU**
Swin-HGI-SAM	0.403 ± 0.014	0.283 ± 0.011
Fully supervised U-Net	0.388 ± 0.019	0.297 ± 0.016
Fully supervised Swin-UNETR	0.428 ± 0.018	0.330 ± 0.011
YOLO-Clustering	0.625 ± 0.020	0.506 ± 0.019
YOLO-SAM-BBox	0.562 ± 0.020	0.445 ± 0.018
YOLO-SAM-Point	$\mathbf{0.627 \pm 0.018}$	$\mathbf{0.506 \pm 0.017}$
YOLO-SAM-PointBBox	$\mathbf{0.629 \pm 0.018}$	$\mathbf{0.508 \pm 0.017}$

outperforming MedSAM on certain medical image segmentation tasks [56] and may be due to a lack of public datasets for training MedSAM on ICH tasks, as Dice loss was used in training the model [61].

3.6 Conclusion

In conclusion, we have proposed a novel weakly supervised ICH segmentation technique that uses YOLO and an uncertainty-rectified SAM. In addition to bounding boxes provided via YOLO, our morphology-based point prompt generation was proven to offer enhanced segmentation performance. Thorough assessments have revealed its superior performance over SOTA weakly supervised and fully supervised baselines while maintaining strong ICH detection capabilities.

Chapter 4

Towards user-centered interactive medical image segmentation in VR with an assistive AI agent

A version of this chapter has been submitted to Springer Nature's *Virtual Reality* journal and is available online as a preprint:

- Spiegler P, Harirpoush A, Xiao Y. Towards user-centered interactive medical image segmentation in VR with an assistive AI agent. *arXiv preprint arXiv:2505.07214*; 2025 [83].

4.1 Introduction

Medical image segmentation is a critical task in clinical diagnosis and treatment planning, particularly for identifying and quantifying abnormalities such as tumors, stroke lesions, and other pathological anomalies. The process involves producing segmentation masks that delineate and highlight regions of interest to provide a basis for further analysis, treatment decisions, and longitudinal tracking. However, traditional workflows for medical image segmentation are time-consuming and labor-intensive, typically requiring experts to manually annotate up to hundreds of 2D slices to isolate structures within 3D MRI or CT scans. Furthermore, while segmenting pathologies (e.g., tumor) accurately is itself a demanding skill, it often requires extensive hours of supervision and training to develop diagnostic confidence and anatomical precision [12]. Finally, visualizations of these annotations are similarly challenging: clinicians either scroll through superimposed binary masks on 2D slices across the axial, sagittal, and coronal planes, or view 3D reconstructions rendered on flat screens, both of which lack real spatial context and a true sense of scale.

While virtual reality (VR) can offer more intuitive 3D medical data visualization and interaction, especially under high spatial constraints (e.g., in the clinic), recent developments in foundation artificial intelligence (AI) models, such as vision-language models (VLMs) have demonstrated early promise to further enhance the efficiency, accuracy, and interactability for tasks in VR in the form of AI agents [48, 5]. As an alternative to conventional brush painting-based segmentation paradigms, imagine a workflow that unites the efficiency of AI with the spatial interaction advantages of a virtual environment: a user reviews a brain tumor MRI in VR assisted by an AI agent. First, the agent guides them towards a representative tumor slice. When the user confirms this slice contains a tumor, they issue a simple voice command, triggering the AI agent to highlight the tumor and provide a detailed patient-specific radiological description. Crucially, this interaction goes beyond the role of a typical conversational assistant: rather than passively responding to dialogue, the AI agent actively executes a sequence of actions, from volumetric segmentation to case-based guidance and full 3D rendering with minimal input. If necessary, the user can easily refine the segmentation using natural inputs, including head pointing, gaze, or handheld controllers, and finally, the corrected mask is rendered in 3D at true scale, offering enhanced spatial interpretation of the tumor's

dimensions. Notably, as current deep learning algorithms for radiological segmentation still require human quality assurance [13, 88, 40], such a human-in-the-loop approach preserves user oversight while dramatically improving speed, accuracy, and spatial understanding through immersive AI assistance.

In our work, we present such a VR-based, AI-assisted medical image segmentation system that supports medical image review for both diagnostic decision-making and education. To the best of our knowledge, this system is the first of its kind, not only automating segmentation across image slices via a conversational AI agent, but also investigating optimal interaction paradigms for human-in-the-loop mask refinement. Our novel contributions are as follows: **First**, we systematically investigate optimal interaction paradigms for human-in-the-loop segmentation refinement that transitions between proximal corrections and distal menu interactions, comparing natural inputs from handheld controllers, head pointing, and gaze. **Second**, we introduce a new 3D segmentation algorithm based on the BiomedParse and SAM2 foundation models which reduces mask drift from noise during slice-to-slice annotation propagation in medical images. **Finally**, we present SAMIRA, a conversational AI agent that assists with segmenting 3D radiological scans via voice commands, provides interactive guidance, supports iterative refinement to preserve expert oversight, and enables life-scale 3D visualization of results in immersive VR.

4.2 Related Work

Although research on AI assistants for clinical training in VR is still in its infancy, early work has begun to demonstrate its potential to support clinical workflows and healthcare education. For example, Liaw et al. [53] developed a conversational AI assistant within a VR simulation for sepsis training, which matched human-controlled scenarios in clinical and communication performance and led to significantly higher test scores. Furthermore, Chheang et al. [18] introduced a generative AI-based virtual assistant in a VR anatomy education environment, highlighting the potential of such assistants to support anatomy education. Although these systems leverage large language models (LLMs) to provide educational support via dialogue, they primarily function as virtual assistants focused on delivering information. In contrast, AI agents should autonomously execute

sequential tasks in response to high-level user prompts, for example, as in our scenario, performing segmentation, providing clinical context via text and images, and rendering 3D models. Retrieval-Augmented Generation (RAG) often powers such agents by retrieving relevant examples from a structured knowledge database and combining similarity search results with generative models to offer refined responses to queries. This contextualized guidance is particularly useful for medical tasks, where visual patterns vary across cases. To our knowledge, no prior work has integrated AI agents into immersive VR for interactive 3D radiological segmentation.

Recent development in foundation models has paved the way for interactive medical image analysis with multi-modal inputs. One such model, BiomedParse[98], is a Transformer-based vision-language model that uses separate encoders for medical images and clinical text data. Trained on 1.2 million paired 2D radiological images and reports, it supports detection, classification, and segmentation of 82 clinical concepts (e.g. tumor) across 9 imaging modalities using natural language prompts. On the other hand, SAM2[47] is a foundation model for segmenting objects in images and video using user-provided point and/or box prompts that enables prompt-based segmentation mask refinement and leverages a memory encoder to propagate masks across video frames. Although developed for natural image domains, SAM2's ability to handle sequential image data positions itself as a promising tool for 3D medical imaging segmentation, where volumetric scans, such as CT or MRI can be viewed as series of 2D images analogous to videos [78, 101]. This opens up the possibility of combining BiomedParse's language-driven label mask generation with SAM2's interactive label refinement and propagation capabilities to enable human-in-the-loop segmentation workflows for 3D medical imaging, an underexplored area, particularly in immersive VR environments. While prior VR segmentation systems have allowed users to paint regions of interest using hand gestures [28, 96] or have used predefined anatomy-specific deep learning models [29], none have integrated foundation segmentation models with agent-driven guidance to support interactive refinement.

In immersive medical image segmentation with SAM2, users will need to frequently switch their visual and motor attention between the image being annotated and a spatially decoupled user interface (UI) menu for triggering actions (e.g., toggling between negative/positive prompts, resetting prompts, etc.), thus creating a dual-focus challenge. Previously, Rashid et al. [73] compared

proximal (on-device) versus distal (remote) widget selection in distributed user interfaces and found that proximal methods were faster and preferred for complex, multi-step tasks, while distal methods yielded lower error rates for simpler interactions, suggesting that high-precision tasks like segmentation prompt placement may benefit from proximal selections and that simpler menu selections may benefit from being performed at a distance. In our case, the interactive refinement of AI-predicted segmentations with SAM2 in VR necessitates identifying optimal interaction paradigms for completing the task under attention-switching. *However, the ideal interaction paradigm (e.g., controller, head pointing, or eye-tracking) for attention switching between proximal and distal displays remains unexplored.* Sidenmark et al. [80] conducted a head-mounted VR study comparing gaze, head, and controller pointing for dynamically revealed target selection, showing that both gaze- and controller-based pointing significantly outperformed head pointing in terms of speed and precision, though they did not explore static dual-panel switching scenarios. Luro et al. [60] compared eye tracking with hand-controller aiming tasks in VR and reported that controllers achieved the highest placement accuracy, while gaze-based selection felt more natural, but suffered from increased selection accuracy variability. Studies of multi-depth targeting, such as the experiment of Schultheis et al. [26] revealed that eye-based selection can achieve higher throughput across varying depth planes, but controller input offers more stable performance when precision is critical. Furthermore, Xu et al.'s evaluation of text-selection techniques in VR [93] demonstrated that head-pointing with click confirmation strikes a balance between speed and accuracy, ranking just behind controller pointing in speed while maintaining a minimal task load, positioning it as a possibly viable paradigm for attention switching interactions. With these previous insights, we will investigate the optimal interaction paradigms for point placement and menu selection under attention switching conditions for our application.

4.3 System Overview

To address the aforementioned issues, we present *SAMIRA* (Segmentation Assistant for Multimodal Interaction and Radiological Analysis), a novel conversational agent designed to support human-in-the-loop 3D medical image segmentation by generating segmentation masks, enabling

efficient mask refinement, and providing radiological guidance through speech and reference images. Figure 4.1 illustrates the system's key components during the segmentation of a liver tumor in a CT scan.

All VR development was conducted in Unity 2022.3.f1 on a desktop equipped with an NVIDIA RTX 3090 and an HTC Vive Pro Eye VR headset. In the VR system, medical image slices are displayed on a virtual panel anchored to a VR controller in the user's hand, allowing flexible repositioning for detailed inspection. Users navigate through image slices by rotating the thumbpad on the HTC Vive controller, clockwise to advance and counter-clockwise to reverse. Each 60° rotation corresponds to a single slice, reinforced by a haptic pulse that provides tactile feedback at each transition.

The menu interface is structured into three panels (see Figure 4.1): on the left panel, AI-driven guidance is provided as text and synthesized speech for the given segmentation task. In the middle panel, interactive controls with functional buttons allow users to issue voice commands, refine predicted masks, propagate segmentations across slices, and render results in 3D. On the right panel, reference images are retrieved in real time based on the user's currently viewed slice based on RAG from an existing knowledge repository. On this panel, the system presents anatomically similar image examples with and without the target structure (e.g., tumor), helping users locate pathology and build a clearer understanding of its visual characteristics.

Using the AI agent's guidance and slice-scrolling mechanism, users can find the target pathology, initiate segmentation via voice commands, refine masks using efficient natural inputs, then render a final segmentation into a true-to-scale 3D visualization to gain spatial understanding of pathologies.

All real-time inference (e.g., RAG retrieval), segmentation generation, point-prompt-based segmentation refinement, mask propagation, and 3D mesh creation are realized by SAMIRA and coordinated via two dedicated Python WebSocket servers: a local rendering server and a dedicated inference server. This is to ensure responsive, low-latency AI assistance and integration between Unity and the supporting AI models. However, future similar systems can benefit from cloud-based setups.

Local Rendering Server: The server ran on the same desktop as the Unity client. It receives binary

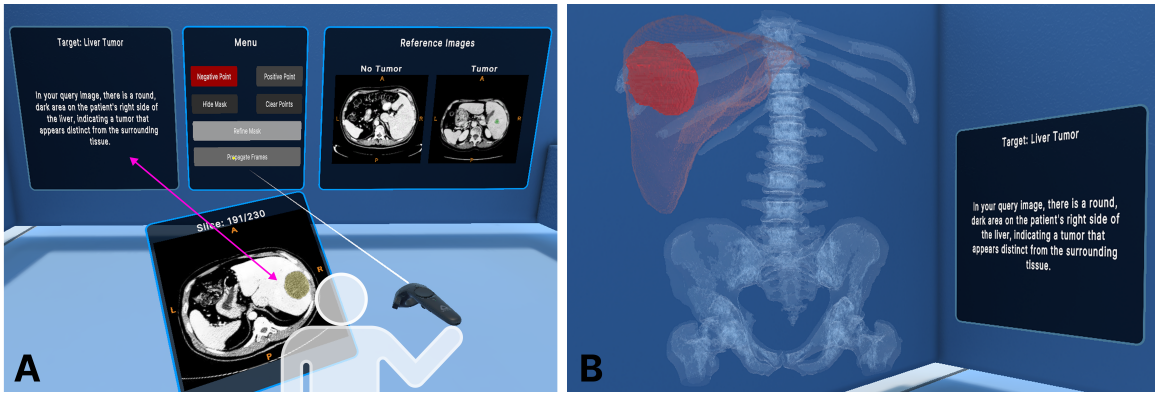


Figure 4.1: A. The AI agent generates an initial segmentation of a liver tumor in CT and provides guidance using reference images and patient-specific pathology explanations. B. The final refined 3D visualization is rendered as a large, spherical, high-contrast liver tumor in red, overlaid on anatomical structures, to scale. Few refinements are expected due to the simple shape.

segmentation masks, scales the lesions using medical image metadata, and returns a `.obj` mesh for VR visualization.

Dedicated Inference Server: A separate Ubuntu 22.04.2 LTS server was set up with a dedicated NVIDIA RTX 3090 to host SAMIRA's RAG and foundational segmentation models (BiomedParse, SAM2). It handles voice commands, reference retrieval, textual guidance, text-based segmentation generation, mask refinement, and mask propagation across frames (with SAM2).

4.4 Methods and Materials

4.4.1 Interaction Paradigm Evaluation

Before evaluating the full VR system with our AI agent, SAMIRA, we aimed to first reveal the optimal interaction paradigm for point prompt placement and menu selection under attention switching between the proximal handheld image display and the distal menu interface. To accomplish this, we designed an experiment to perform SAM2-based segmentation mask refinement using three distinct paradigms: *Controller*, *Head Pointing*, and *Eye Tracking*. The detailed setup is shown in Fig. 4.2. As a potential middle ground between *Controller* and *Eye tracking*, *we hypothesize that Head Pointing is the optimal interaction paradigm for our intended application, with the consideration of task load, accuracy, and efficiency*.

Controller pointing is a staple method for object and menu selection in many VR applications. We employed standard controller-based ray-casting, where a visible ray extends from the tip of the controller to intercept with the image under analysis and the menu for point prompt annotation and button clicking, respectively. Here, confirmation of selection is achieved by pressing the trigger button of the controller.

Head Pointing adopts an invisible ray forward from the center of the user's headset, aligning with the head's orientation. The interception of the ray and the image/menu is represented by a visible dot, and the controller trigger button is used for placement/selection confirmation.

Eye Tracking utilizes the integrated eye-tracking hardware of the HTC Vive Pro Eye headset to cast an invisible ray based on the user's gaze direction. Similar to head pointing, a visible dot is used to indicate the gaze point on the image under analysis and the menu. Based on previous studies [36, 66], we continue to use the trigger button to confirm selection. To mitigate the adverse impacts of natural eye jitter on selection precision and user experience, we apply exponential smoothing to the normalized 3D gaze vectors during every frame using a smoothing factor of $\alpha = 0.2$. The gaze ray is computed as follows:

$$\mathbf{r}_{\text{smooth}} = (1 - \alpha) \cdot \mathbf{r}_{\text{previous}} + \alpha \cdot \mathbf{r}_{\text{current}} \tag{2}$$

where $\alpha = 0.2$ is the smoothing factor, $\mathbf{r}_{\text{previous}}$ is the smoothed direction from the previous frame, and $\mathbf{r}_{\text{current}}$ is the normalized average of the left and right eye gaze direction vectors, originating at the midpoint of the two eyes.

The overall interaction paradigm evaluation workflow is summarized in Figure 4.2F, where the user begins at the middle slice of a segmented 3D scan and scrolls through the volume to identify and correct segmentation errors. Three fixed large panels (Figure 4.2C) are positioned three meters in front of the user, different from the panel displays for the full interactive segmentation workflow (Figure 4.1): the left panel displays the current interaction mode (controller, head pointing, or eye tracking), the central panel provides interactive controls (i.e., functional buttons), and the right panel shows the ground truth segmentation (red, 40% opacity overlaid on the image) for the current

slice under study. Here, the ground truth segmentation is used to define a consistent reference for refinement and objective evaluation of segmentation accuracy across participants in the user study. Users were instructed to correct the provided masks until they visually matched the ground truths. The central panel includes six buttons. "Positive Point" and "Negative Point" buttons allow users to place point prompts that add missing regions or remove excess segmentation, respectively. The "Hide Mask" button toggles the visibility of both the segmentation under work and ground truth to better assess tissue boundaries. The "Clear Points" button erases all placed point prompts without modifying the current mask. The "Refine Mask" button sends the current slice and point prompts to the inference server to update the segmentation. Finally, "Complete Plan" finalizes segmentation refinement and advances to the next interaction paradigm (*Controller*, *Head Pointing*, or *Eye Tracking*), selected randomly, allowing within-participant comparisons of performance across paradigms.

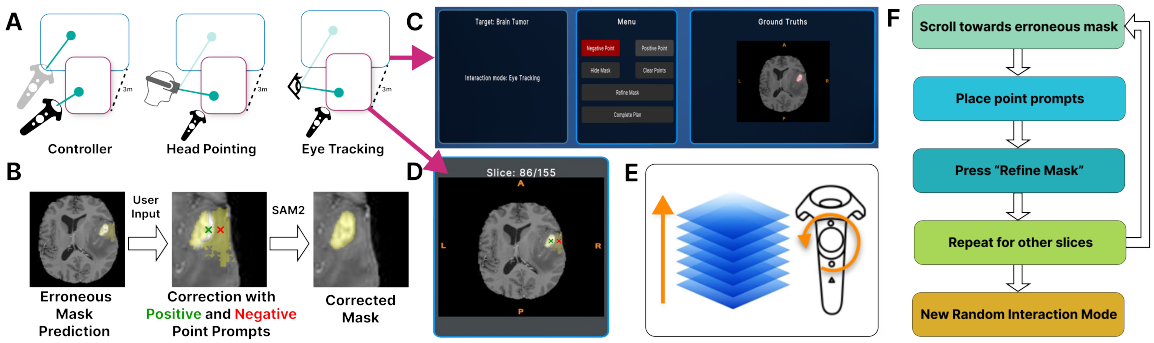


Figure 4.2: System setup for user interaction paradigm evaluation under attention switching. A. Three interaction paradigms: controller ray, head pointing, and eye tracking. B. Users correct erroneous masks using positive (green) and negative (red) point prompts, refined by SAM2. C. In-VR interface for prompt selection and real-time ground truth reference. D. Medical image display with current slice and segmentation overlay. E. Controller-based slice scrolling. F. Interaction paradigm evaluation segmentation workflow.

To evaluate interaction paradigms under attention-switching for AI-facilitated segmentation refinement, we used a T1c MRI scan from the publicly available BRATS brain tumor dataset [62], which was intensity-normalized and converted into a sequence of 155 axial JPEG image slices. All slices along with 44 slice-wise ground truth masks and 16 intentionally corrupted slice-wise masks were included. The corrupted masks represented common failure cases, such as incomplete tumor coverage, over-segmentation, missing regions, or false positives. The quality of the initial segmentation, measured using the 3D Dice score, was assessed at 0.91. 3D Dice is a standard metric for

volumetric medical segmentation that quantifies the spatial overlap between the predicted mask A and the ground truth B as:

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

A Dice score of 1 indicates perfect agreement; 0 indicates no overlap.

4.4.2 SAMIRA - User Interface and Workflow

The full workflow of our interactive medical image segmentation leverages SAMIRA, our assistive AI agent via the VR user interface described in Section 4.3. The interface enables SAMIRA to operate on unseen cases without ground truth segmentations, and supplies reference images and case-relevant descriptive guidance in real time. This supports the potential for future clinical and educational applications beyond the controlled evaluation settings showcased in our study.

For demonstration and evaluation purposes, the segmentation workflow was applied to a brain tumor MRI scan from another dataset, Pretreat-MetsToBrain-Masks [72], and a liver tumor CT scan from the LiTS dataset [7]. These two datasets were deliberately chosen to assess generalization beyond the training distribution of the underlying models, since the model we are introducing, BiomedParse was trained on the BRATS dataset used in the interaction paradigm test. While the examples focused on tumors, the workflow is compatible with any of the 82 clinical targets supported by BiomedParse. The brain tumor example was selected as the more difficult case, with branching regions (see Figure 4.3E), whereas the liver tumor case was expected to be easier to segment, with a smoother, spherical structure in which less refinement is expected (see Figure 4.1).

The full workflow is illustrated in Figure 4.3 and consists of the following sequential steps:

Step 1. Initial Contextualization: A handheld image display begins at the middle slice of the brain or the liver scan. The AI agent introduces general guidance for segmenting the clinical target on the left panel (Figure 4.3A) using a RAG pipeline. To ground users' visual understanding, the system retrieves and displays a positive example (with pathology) and a negative example (without pathology) on the right panel, allowing for contrastive comparison. Both examples bear anatomical similarity to the image slice under study. Then, general pathology explanations are delivered via Google's Text-To-Speech API in a female voice, and are displayed as text on the left panel. At this

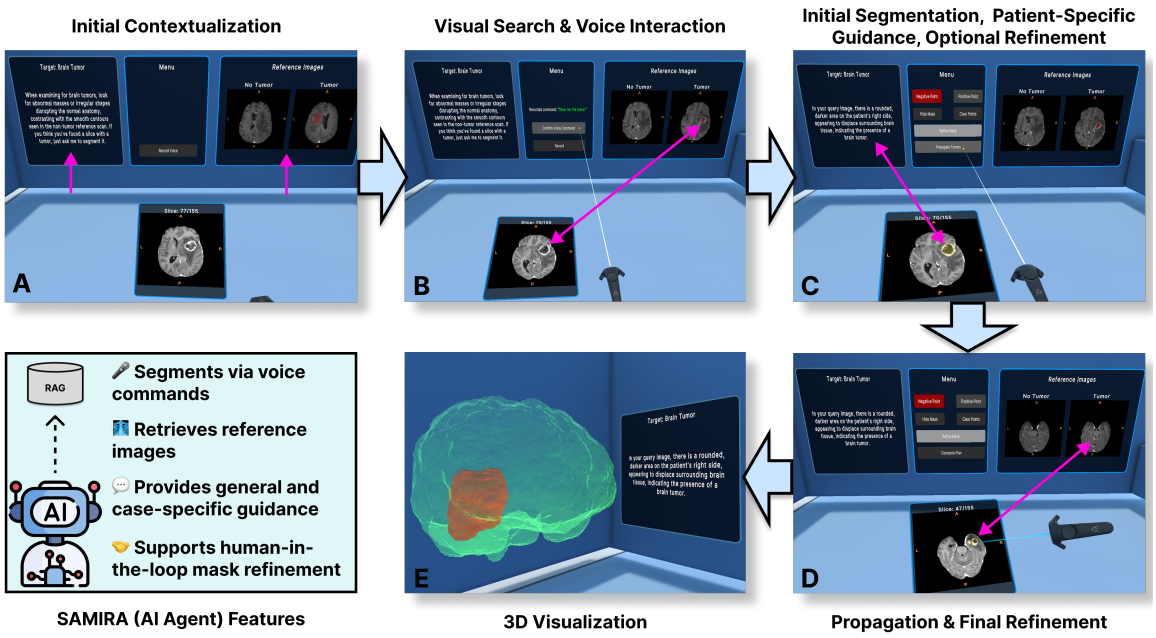


Figure 4.3: Demonstration of workflow for the proposed AI-assisted interactive medical image segmentation in VR. Users begin by reviewing AI-generated textual guidance and visually similar reference slices (A), then navigate the volume to find the tumor. Once found, they issue a voice command to segment the tumor (B). Next, the agent predicts a mask and a patient-specific description of the tumor (C). If necessary, users can edit this mask, then propagate it across frames. Finally, users review all predicted masks and place point prompts to refine the masks (D). Upon completion, the final segmented structure is rendered in true 3D scale over the patient's anatomy (E).

stage, the middle panel has one button that says “Record Voice”.

Step 2. Visual Search: Users scroll through the image slices in search of the target region and the agent continuously updates the right panel reference panel with visually similar tumor and non-tumor cases retrieved from other patients. The contextual guidance in Step 1 helps less experienced users with deciding whether the current slice contains the pathology of interest (see Figure 4.3B).

Step 3. Voice Interaction: Users initiate segmentation by clicking the “Record Voice” button and issuing natural language requests to the AI agent. Spoken input is transcribed in real time using the Microsoft Azure Speech SDK, with recognized commands displayed on the middle panel (Figure 4.3B). After voice recognition, users can press the “Record Voice” button again to re-record, or press “Confirm Voice Command” to pass their command to the inference server.

Step 4. Initial Segmentation and Patient-Specific Guidance: The AI agent returns an initial segmentation mask at the selected 2D slice and uses its RAG pipeline to provide additional spoken, case-specific diagnostic context (Figure 4.3C). The middle panel updates with the same refinement options as the *Interaction Paradigm Evaluation* interface and an additional button “Propagate Frames”.

Step 5. Refinement and Propagation: Image slice scrolling is disabled at this stage, forcing the users to focus on the selected 2D slice of interest to achieve the best segmentation accuracy via point-prompt-based refinement if necessary. This is because it will be used to seed the automatic SAM2-based label propagation to obtain the full 3D segmentation. Once satisfied with the current slice, the users initiate mask propagation across slices by pressing the “Propagate Frames” button. As the inference server returns produced segmentation masks in sequence, the image display automatically updates the slices in the order that the masks are produced in real-time, enabling users to visually assess the segmentation results and quickly judge whether additional refinements will be needed.

Step 6. Final Refinement of Predicted Masks: Users review the propagated masks across slices, applying additional point-prompt-based refinements wherever segmentation errors remain (Figure 4.3D). During this stage, the “Propagate Frames” button is replaced by a “Complete Plan” button. When pressed, a confirmation screen appears to prevent accidental submission, ensuring that users have fully completed their corrections before finalizing the segmentation plan. Upon completion,

the final segmentation masks are saved.

Step 7. 3D Visualization: Upon task completion, a request is sent to the local Python-based web-socket server with the path to the saved segmentation result. The server uses the *Marching Cubes* algorithm to extract a polygonal mesh from the 3D segmentation and generates a corresponding `.obj` file representing the tumor pathology. To ensure anatomically accurate scaling, the voxel resolutions are extracted from the medical image's metadata and applied during mesh generation. The resulting mesh is then loaded back into Unity at runtime and rendered at true-to-life scale (Figure 4.3E). For brain tumor visualization, a threshold of 500 is applied to extract the brain surface from the scan, while liver tumor visualization uses a threshold of 150 to capture relevant anatomical structures such as the spine and ribcage. Additionally, liver tissue from the same patient was rendered using data from the LiTS dataset. In the future, it can easily be generated with liver segmentation models, or even interactively with our workflow.

4.4.3 SAMIRA - Segmentation Algorithm and Retrieval-Augmented Generation

SAMIRA leverages AI foundation models and Retrieval-Augmented Generation to assist the designated interactive medical image segmentation task.

Interactive segmentation with foundation models

We proposed a novel deep learning-based, speech-initiated interactive segmentation method for 3D medical images as illustrated in Figure 4.4. This method forms a key function of SAMIRA and relies on two complementary foundation models. The first model, *BiomedParse*, generates the initial segmentation mask in response to user-issued voice commands. Spoken prompts are first transcribed using the Azure Speech AI service, producing a text prompt that is passed to BiomedParse along with the image slice. Upon receiving the textual prompt and corresponding image slice, BiomedParse produces a binary mask of the target structure described in the user's voice command. The second model, SAM2, is used for interactive refinement and multi-slice propagation. Users can iteratively correct the mask using positive and/or negative point prompts, which are passed to SAM2 for real-time refinement. Once satisfied, the refined mask is propagated bi-directionally

through the volume using a modified version of SAM2's memory mechanism with a novel propagation termination criterion: if the inter-slice Intersection-over-Union (IoU) between the current mask (Mask_t) and previous mask (Mask_{t-1}) fell below 0.3, propagation is halted in that direction, under the assumption that mask changes should be relatively smooth between neighboring image slices. Inter-slice IoU measures how much two sequentially predicted binary masks overlap, defined as the ratio between the area of their intersection and the area of their union:

$$\text{IoU} = \frac{|\text{Mask}_t \cap \text{Mask}_{t-1}|}{|\text{Mask}_t \cup \text{Mask}_{t-1}|}$$

While point-prompt-based revision with SAM2 helps ensure the accuracy of the segmentation, it is desirable to keep the needs at minimum for the efficiency of the workflow. Thus, to gauge the baseline performance of our proposed segmentation method in the absence of prompt-based revision, we conducted a standalone ablation study using the Pretreat-MetsToBrain-Masks [72] and LiTS [7] datasets, by comparing the accuracy of a "BiomedParse seeding + SAM2 propagation with IoU-based early stopping" pipeline versus a "BiomedParse seeding + original SAM2 propagation" one. To accomplish this, we randomly selected 40 volumes (20 brain tumor MRI scans and 20 liver tumor CT scans) from the datasets, excluding cases used in the full workflow study. For each volume, a slice with visible tumor was manually selected and submitted to the inference server along with a natural language prompt —"show me the brain tumor" for brain cases and "show me the liver tumor" for liver cases. Here, BiomedParse generates an initial tumor mask for the selected slice, which was then propagated bi-directionally (in the superior and inferior directions) using SAM2. Using the paired sample Wilcoxon signed-rank test, results indicate that the break condition significantly improved accuracy for liver tumor CT scans ($p = 0.0024 < 0.05$), while having a smaller but still significant effect for brain tumor MRIs ($p = 0.0039 < 0.05$), as summarized in Table 4.1. The stronger effect observed in CT scans may be attributed to the higher levels of noise and lower soft-tissue contrast, which can cause SAM2, originally trained on natural images, to mistake noise for anatomical structures. The break condition prevents propagation of spurious segmentations across slices, which is particularly helpful in noisier CTs. Compared to recent SAM2-based methods for 3D medical segmentation [78], our approach introduces both a language-based initialization and a

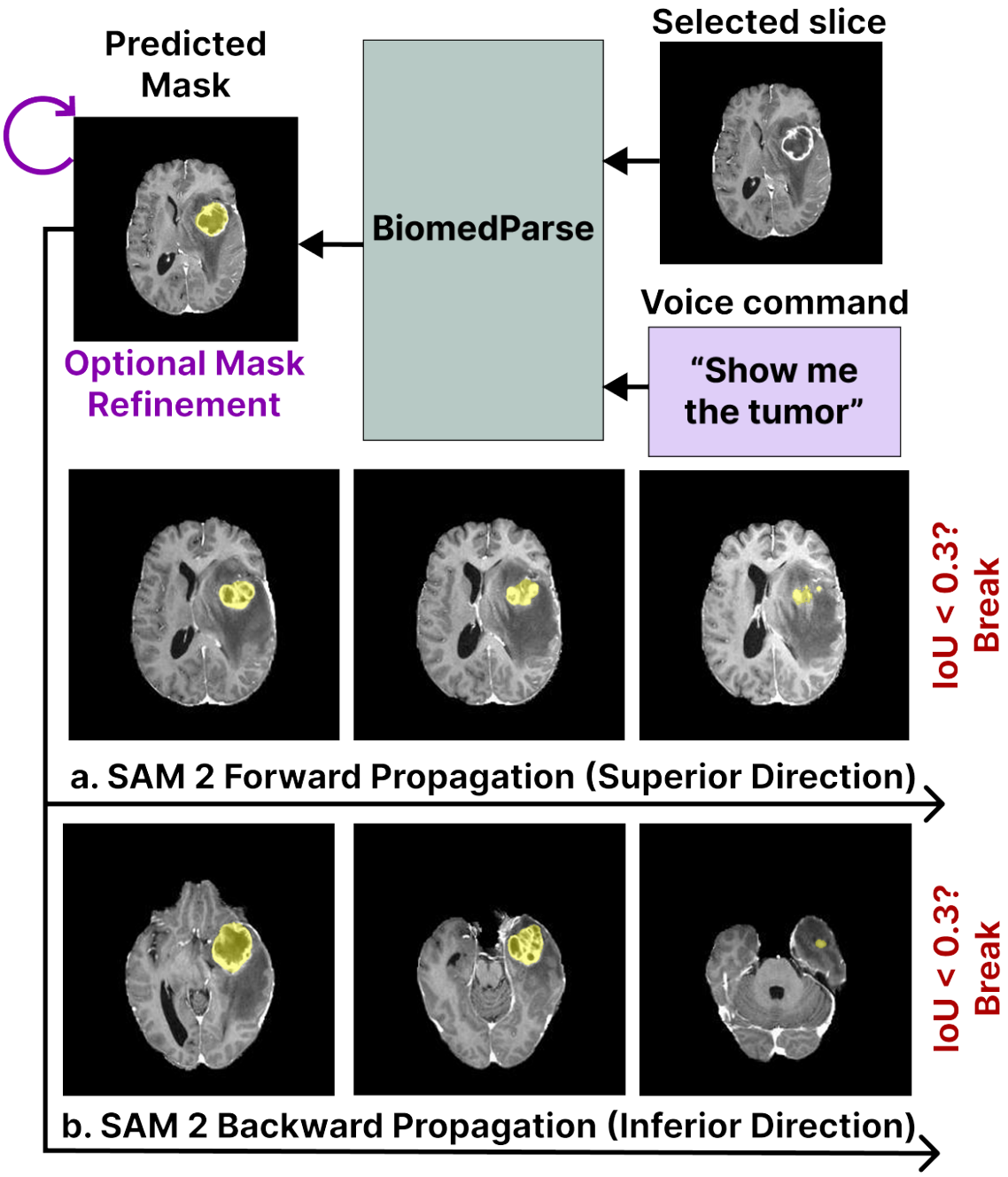


Figure 4.4: SAMIRA's segmentation module for mask prediction, refinement, and propagation across frames. After a voice command initiates initial tumor segmentation via BiomedParse, the user may optionally refine the mask through point prompts. The mask is then propagated slice-wise using SAM2, first superiorly (a) and then inferiorly (b), with propagation automatically terminating when inter-slice Intersection-over-Union (IoU) falls below 0.3 to prevent segmentation drift.

propagation stopping rule, enhancing accuracy and reducing user burden. These findings support the inclusion of the IoU threshold as an essential mechanism for robust slice-wise mask propagation

in clinical datasets.

Table 4.1: Mean 3D Dice scores for automatic SAM2 propagation with and without the IoU break condition.

Modality	Target	IoU Break	3D Dice (%, Mean ± Std)
MRI	Brain Tumor	**True**	**87.41 ± 10.77**†
		False	87.28 ± 10.73
CT	Liver Tumor	**True**	**73.31 ± 13.04**†
		False	68.94 ± 16.30

†Statistically significant improvement compared to no IoU break condition

RAG-based guidance system

In addition to interactive segmentation, to provide multi-modal, context-aware guidance during segmentation, SAMIRA employs a RAG framework that integrates image similarity search with generative language modeling. At its core is FAISS [23], a library developed by Meta for fast approximate nearest-neighbor search on high-dimensional vectors. FAISS enables real-time retrieval of anatomically similar tumor and non-tumor slices from large medical image datasets, based on high-dimensional vectors from the high-level feature maps output by the Res5 layer of BiomedParse's image encoder. These retrieved examples are used in two ways: first, to provide general contextual grounding using representative healthy and pathological slices (RAG Request 1) and second, to generate query-specific guidance based on the user's spoken prompt and the current image slice (RAG Request 2).

Knowledge Database Construction: We constructed two FAISS databases, one for brain MRIs and one for liver CTs, which contain a total of 30,845 brain slices (6,766 tumor, 24,079 without tumor) and 57,193 liver slices (6,982 tumor, 50,210 without tumor), drawn from 199 and 127 patients, respectively. Embedding vectors were computed using BiomedParse (specifically its high-level res5 output), normalized, and stored for efficient similarity search. These databases can easily be expanded in the future with a larger variety of targets and cases.

RAG Request 1. Initial Contextualization: When a new scan of interest is loaded in the system, SAMIRA uses the encoding vector of the middle slice to retrieve two visually similar reference

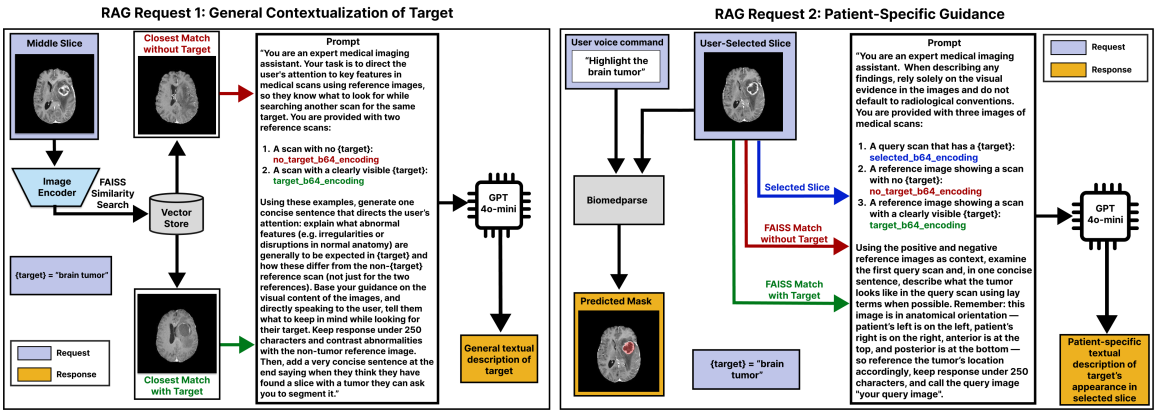


Figure 4.5: Retrieval-Augmented Generation (RAG) pipelines for multimodal guidance. (Left) To support initial understanding, the system retrieves two anatomically similar reference slices—one with and one without the target pathology—and uses them to generate a general description of the abnormality. (Right) After the user selects a slice and issues a voice command, the system compares visual features of the patient's scan to healthy and pathological reference images. Guided by shared and differing features, the agent describes what the abnormality likely looks like in the selected slice.

images from the knowledge database: one with and one without the target pathology (i.e., tumor). These contrastive examples help ground the task as the non-tumor slice illustrates normal anatomy, while the tumor case highlights typical features of the target pathology. Then, both are passed to GPT-4o-mini, chosen for its fast inference time, which generates a general explanation of what to look for during visual exploration of the target. Figure 4.5 illustrates this process, showing how supplying reference images with a carefully engineered prompt can support contextual grounding to guide the user.

RAG Request 2. Query-Specific Interpretation: While general contextualization offers users an overview of what the pathology typically looks like, it does not account for anatomical variation in individual cases being investigated. Therefore, query-specific interpretation is used to provide more personalized, patient-specific guidance. When the user issues a voice command to segment a structure (e.g., "highlight the brain tumor"), the selected image slice is used to query the knowledge database, retrieving two visually similar reference slices: one with the target pathology (positive example) and one without (negative example). These references, *along with the user's query slice*, are passed to GPT-4o-mini in another structured multi-modal prompt, illustrated in Figure 4.5. By explicitly contrasting the *query* with both healthy and pathological references, the agent can generate

targeted explanations grounded in the anatomy of the target image under study.

4.4.4 User Study and Evaluation Metrics

We conducted two separate user studies for the developed system. While the first study aimed to identify the optimal user interaction paradigm for SAM2-based segmentation refinement, the second study assessed the full workflow of the proposed system. All user studies were conducted under institutional ethics approval and after all participants provided informed consent.

User-Interaction Paradigm Evaluation

We recruited 15 participants (age = 26.5 ± 2.68 years; 6 female & 9 male) to evaluate the three interaction paradigms (*Controller*, *Head Pointing*, and *Eye Tracking*). Participants rated their familiarity with VR, human anatomy, medical imaging modalities, and medical image segmentation on a 1–5 scale (1 = unfamiliar, 5 = familiar). On average, they reported familiarity with VR (4.2 ± 0.94), and between neutral and somewhat familiarity with anatomy (3.47 ± 1.24), imaging modalities (3.53 ± 1.30), and medical image segmentation (3.73 ± 1.27). Prior to the study, participants completed a guided hands-on tutorial in the VR environment, which included practice with segmentation mask refinement on a lung CT scan from the LCTSC dataset [95] and menu interaction using each of the three interaction paradigms. The tutorial session also served to calibrate the eye-tracking system. Following this, users began the interaction test, correcting the erroneous BRATS brain tumor MRI slices with each interaction paradigm as described in Section IV.A. For each interaction paradigm trial, task completion time, segmentation accuracy (with 3D Dice score), perceived task load, point prompts placed, and point prompts erased were recorded.

To assess perceived task load, we used the *NASA Task Load Index (NASA-TLX)* [33], a validated six-item instrument measuring mental demand, physical demand, temporal demand, effort, frustration, and perceived performance. Each item was rated in the range of 1-21, which was scaled to 0–100 for analysis, and finally all items were averaged to produce a total task load score. To compare overall performance across interaction paradigms, we computed a composite interaction score for each of the 45 user-paradigm trials (15 participants × 3 interaction paradigms). This score quantifies trade-offs between segmentation accuracy, task load, and completion time. Z-scores for

accuracy, NASA-TLX, and completion time were calculated across all trials to ensure standardized comparison. For each trial i, the composite score was calculated as:

$$\text{Composite}_i = z_{\text{accuracy}} - z_{\text{NASA}} - z_{\text{time}}$$

This score treats accuracy as beneficial and both task load and time as costs, giving equal weight to each. Mean and standard deviation of composite scores were then computed per interaction paradigm to summarize performance. Finally, for additional self-reported confirmation, we asked users to rank their preferred interaction paradigm.

Full Workflow Study of SAMIRA

To evaluate the complete AI-assisted segmentation system, including conversational interaction, mask refinement, and 3D visualization, we recruited 19 participants (age = 26.8 ± 3.63 years; 8 female & 11 male). On average, they again reported familiarity with VR (4.21 ± 1.18) and between neutral to somewhat familiarity with anatomy (3.42 ± 1.22), imaging modalities (3.79 ± 1.18), and medical image segmentation (3.74 ± 1.37). After a brief tutorial, where participants practiced the full workflow by segmenting a lung CT scan from the LCTSC dataset [95], participants completed the two segmentation tasks: brain tumor in MRI and liver tumor in CT. The order of these was randomized across participants to minimize order effects. Outcome measures included segmentation accuracy (3D Dice scores before and after segmentation refinement), task completion time, and user experience metrics. Task load was assessed using NASA-TLX, similarly to the interaction paradigm evaluation. To assess the overall usability, we also administered the *System Usability Scale (SUS)* [10], a widely used 10-item questionnaire that yields a total score from 0 to 100. Scores above 68 are considered to indicate good usability [9]. Additionally, we developed a custom 9-item questionnaire to evaluate users' perceptions of AI agent guidance, reference images, 3D visualization, and user confidence in the workflow. The full list of questions are provided in Figure 4.7. Items used a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). Finally, we included an open-ended section, where users could freely mention what they liked and/or disliked about the system, and further elaborate their semi-quantitative evaluations.

Statistical Analysis

For the interaction paradigm study, differences in NASA-TLX scores, completion times, and Dice scores across three user-interaction paradigms were evaluated using Kruskal–Wallis tests. For the full workflow study, we compared the 3D Dice scores before and after user refinement using Wilcoxon rank-sum tests. SUS scores were tested against the usability benchmark of 68 using a one-sample t-test. Custom user questionnaire responses were tested against the neutral value of 3 using Wilcoxon signed rank tests. For all thee statistical tests, a statistical significance was confirmed with $p < 0.05$.

4.5 Results

4.5.1 Interaction Paradigm Evaluation

Table 4.2: Comparison of interaction paradigms for segmentation refinement. Values are shown as mean ± standard deviation. The best score is in bold fonts. NASA-TLX is out of 100.

Interaction Mode	3D Dice (%)	Time (s)	NASA-TLX	Composite Score
Controller	**99.25 ± 00.25**	**220.3 ± 79.3**	18.8 ± 14.5	**0.51 ± 1.91**
Head Pointing	99.21 ± 0.30	248.8 ± 78.5	**16.8 ± 13.9**	0.20 ± 1.56
Eye Tracking	99.13 ± 0.46	251.1 ± 78.7	26.6 ± 15.4	-0.71 ± 1.77

Accuracy, Completion Time, and NASA-TLX

The metrics for evaluating the three interaction paradigms are shown in Table 4.2, with non-significant differences in overall scores between groups ($p > 0.05$). For **3D Dice (segmentation accuracy)**, all paradigms yielded highly accurate refined masks, significantly above the starting 3D Dice score of 0.91 ($p < 0.05$). The *Controller* paradigm had the highest accuracy score with the least variability ($99.25 \pm 0.25\%$), followed by *Head Pointing* ($99.21 \pm 0.30\%$), then *Eye Tracking* ($99.13 \pm 0.46\%$). For **completion time**, the ranking remained the same: *Controller* was the fastest ($220.3 \pm 79.3s$) followed by *Head Pointing* ($248.8 \pm 78.5s$) and *Eye Tracking* ($251.1 \pm 78.7s$). The overall **NASA-TLX** scores (out of 100) slightly favored *Head Pointing*, which had the lowest

overall task load (16.8 ± 13.9), followed by *Controller* (18.8 ± 14.5) and *Eye Tracking* (26.6 ± 15.4). While overall task load did not demonstrate significant differences, one sub-item showed a significant effect: mental demand was significantly lower for *Head Pointing* (16.7 ± 22.5) compared to *Controller* (30.0 ± 23.9, $p = 0.0123$) and *Eye Tracking* (38.7 ± 25.5, $p = 0.0329$), as illustrated in Figure 4.6. No other NASA-TLX sub-items' differences reached statistical significance.

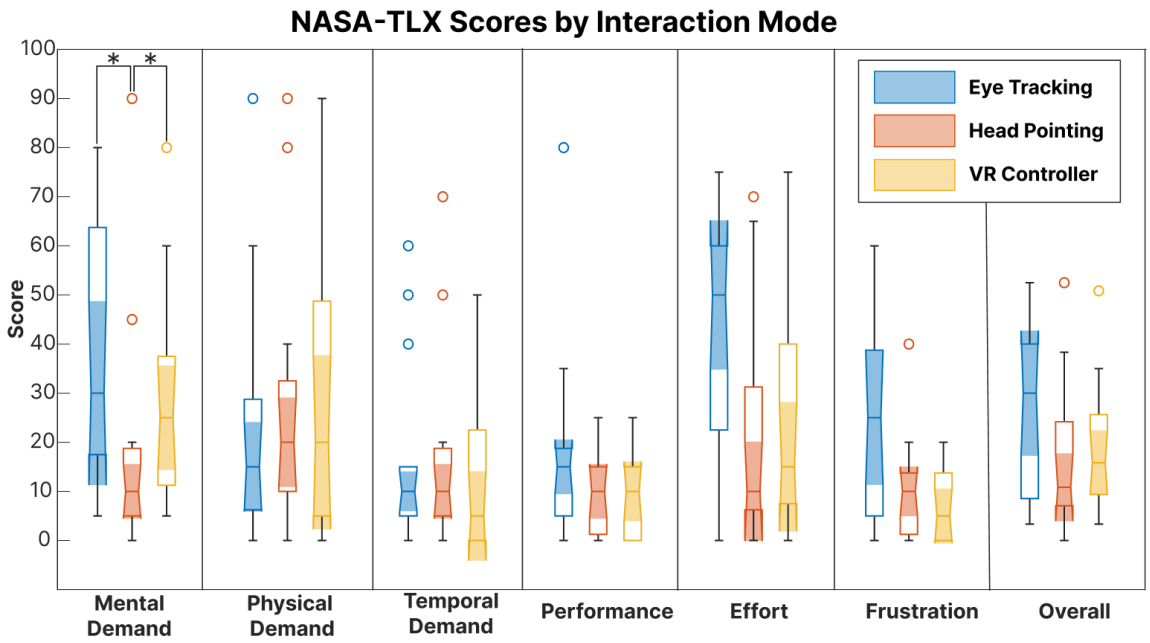


Figure 4.6: Boxplots of NASA Task Load Index (TLX) for different interaction paradigms, including Controller, Head Pointing, and Eye tracking. Significant pair-wise differences are marked with "*".

Composite Scores

The resulting mean composite scores that promote a balance between performance, speed, and user effort were 0.51 ± 1.91 for *Controller*, 0.20 ± 1.56 for *Head Pointing*, and -0.71 ± 1.77 for *Eye Tracking*. Among the paradigms, *Controller* achieved the highest mean composite score, driven by a favorable combination of high segmentation accuracy (99.25%), low task load (NASA-TLX = 18.8), and the fastest completion time (220.3 s). *Head Pointing* followed closely, while *Eye Tracking* lagged behind due to relatively higher task load and slower performance.

Point Prompt Efficiency

To complement the accuracy, task completion time, and NASA-TLX, we also compared the three interaction paradigms in terms of the ratio between the number of confirmed point prompts for segmentation refinement and the cleared points (i.e., points-per-clear efficiency, the higher the better). This ratio reveals the robustness of these paradigms with the joint consideration of their inherent precision, efficiency, and task load. Overall, *Controller* yielded the highest points-per-clear efficiency (24.33 ± 24.91), followed by *Head Pointing* (16.08 ± 6.17) and *Eye tracking* (14.67 ± 8.72), suggesting that the *Controller* condition was more robust for point prompt placement (less corrections required) and eye tracking was more prone to errors.

User preferences

Finally, when asking the participants to rank their preferred interaction paradigms, 7 of 15 participants selected *Controller*, 6 selected *Head Pointing*, and only 2 selected *Eye tracking* as their top choices. These results, in addition to the close composite scores, suggest that both *Controller* and *Head Pointing* are viable and well-received paradigms, while *Eye tracking* is comparatively less favored. The close ranking between *Controller* and *Head Pointing* indicates that either approach could be suitable for deployment in the full workflow. However, to reduce variability in downstream evaluation, we selected *Controller* as the primary interaction paradigm for the subsequent full workflow study due to its marginally superior results.

4.5.2 Full Segmentation Workflow with SAMIRA

Segmentation accuracy

In terms of segmentation accuracy (3D Dice), the workflow yielded high scores. For the **brain MRI data**, refined masks ($94.92 \pm 0.52\%$) showed significantly higher accuracy than the user's propagated unrefined masks ($90.53 \pm 10.51\%$, $p < 0.0001$). For the **liver CT data**, Dice scores were comparable between unrefined (95.47 ± 0.39) and refined (95.46 ± 0.40) masks, with no significant difference, ultimately indicating high starting accuracies are preserved (Table 4.3). This high starting accuracy is likely attributed to the liver tumor case being larger and more spherical

shaped than the brain tumor case that appears as a mass centrally, but splits into multiple lobes in the superior and inferior regions. The lower difficulty of the liver tumor case is further reflected in the shorter completion time (liver: 279.6 ± 109.4s vs. brain: 361.7 ± 144.3 s) and fewer point prompts placed (liver: 8.8 ± 6.5 vs. brain: 27.7 ± 19.9). The brain tumor case likely demanded more user input due to its aforementioned irregular shape.

Semi-quantitative questionnaire results

Participants rated the overall workflow with SAMIRA as highly usable. The **System Usability Scale (SUS)** score was 90.00 ± 8.98, which is significantly higher than the benchmark of 68 ($p < 0.001$) and corresponds to an 'A' usability score [9]. Meanwhile, the **NASA-TLX** scores (out of 100) indicate low to moderate task loads across sub-items. Descriptive statistics are as follows: Mental Demand (31.84 ± 28.10), Physical Demand (12.11 ± 14.27), Temporal Demand (13.68 ± 18.25), Performance (20.00 ± 24.72), Effort (25.26 ± 15.50), Frustration (7.11 ± 8.22), and Overall task load (18.33 ± 11.93).

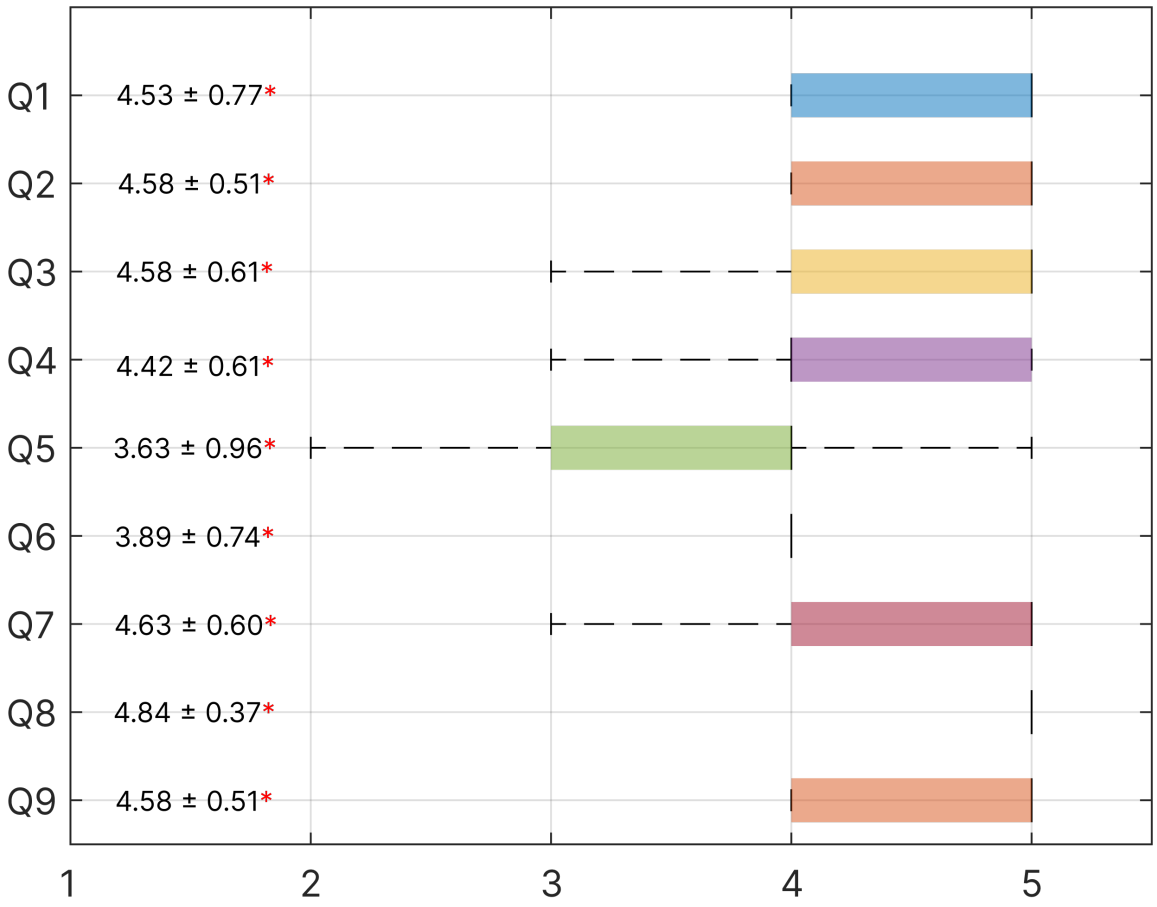
Table 4.3: 3D Dice scores before and after refinement. Asterisks denote statistically significant changes.

Tumor	Unrefined	Refined	p-value
Brain	90.53 ± 10.51	94.92 ± 0.52	1.19×10^{-5}*
Liver	95.47 ± 0.39	95.46 ± 0.40	0.3968

Finally, all responses on the **custom questionnaire** were significantly above the neutral midpoint of 3 ($p < 0.05$), indicating favorable perceptions of the system's AI agent integration, its ability to support pathology learning, the helpfulness of reference images, and the usefulness of 3D visualization (see Figure 4.7).

Qualitative feedback results

Participants' written feedback, where they were asked to describe what they liked or disliked about the system, further supports the system's perceived usability and educational value. *16 out of 19 participants* described the system as "easy to use," "intuitive," or "clear," highlighting its



Q1. I found the system provides an immersive experience.
Q2. I found that the AI system is nicely integrated in the workflow.
Q3. I found the AI system instrumental for me to complete the segmentation tasks.
Q4. Overall, the system offers me a good understanding of the pathology to segment.
Q5. I found the reference images helpful in making my judgement for segmentation.
Q6. I found the AI-generated description of pathology helpful for me.
Q7. I found the 3D visualization of the segmentation result helpful.
Q8. Overall, I think the workflow is well designed.
Q9. Overall, I think the system improves my confidence for performing medical image segmentation tasks.

Figure 4.7: Boxplots of custom user experience questionnaire results (mean ± std on left, values significantly above 3 with red asterisk) for the full workflow with SAMIRA.

low learning curve and smooth integration with the VR environment. One participant noted, "After you learn it, it is easy to use, fast, and interactive," while another stated, "The system is very user friendly and well integrated with the HTC Vive controls." These perceptions align with the high overall scores on SUS Question 3 ("I thought the system was easy to use," 4.58 ± 0.51) and

Question 5 ("I found the various functions in this system were well-integrated," 4.68 ± 0.48). AI voice interaction was also frequently praised. *10 participants* explicitly mentioned that the voice command feature improved usability, with one stating, "The voice commands are well integrated and the propagation is very helpful for identifying the whole tumor." Another remarked, "The AI was well integrated and didn't feel intrusive—more like a helpful assistant." The system was also perceived as supportive of learning and decision-making. *7 participants* reported increased confidence during segmentation tasks, attributing this to guidance from the AI agent and the real-time visualization tools. For example, one wrote, "It was so useful and made me confident to do the task", echoing high SUS Question 9 results ("I felt very confident using this system," 4.74 ± 0.45).

While overall impressions were positive, some participants suggested providing additional medical context, and a few expressed awkwardness with voice inputs. Nonetheless, the qualitative feedback reinforces the questionnaire findings and confirms that SAMIRA's design successfully balances guidance, autonomy, and interpretability in a VR environment.

4.6 Discussion

The findings of our experiment provided partial support for our hypothesis that *Head Pointing* would provide the optimal trade-off between segmentation accuracy, efficiency, and task load. While *Head Pointing* did show the lowest mental demand, *Controller*-based input achieved slightly better overall performance, as reflected in its higher mean accuracy and efficiency metrics and its highest average composite score. Furthermore, all paradigms yielded excellent Dice scores following segmentation correction, but *Controller* and *Head Pointing* outperformed *Eye Tracking* in terms of accuracy, task load, and completion time, echoing the findings of Xu et al.'s evaluation of text-selection techniques [93], where head-pointing and controller performance were close. Despite composite scores favoring controller-based pointing overall, *Head Pointing* emerged as a lower mental effort alternative—especially for applications where users may need to work with just one hand or seek a cognitively lighter interaction. *Eye Tracking*, while promising in theory, remains less favored for segmentation refinement tasks, where precision and visual stability are critical, despite

the damping function (Equation 2) we employed to improve precision and user-experience. In future deployments, the system can allow users to select their preferred interaction paradigm, offering flexibility. This flexibility would be warranted, since all three paradigms were chosen as favorites across the different users.

Overall, our findings suggest that AI-assisted segmentation in VR is not only technically viable, but also educationally and ergonomically impactful. Across studies, users were able to generate high-quality segmentations without domain expertise and the system demonstrated that it promotes understanding, confidence, and informed interaction. In the full workflow, participants rated the system highly on both usability and interpretability of the task. As seen in Figure 4.7, users strongly agreed that the AI agent helped them complete tasks (Q3: 4.58±0.61), supported understanding of the pathology (Q4: 4.42±0.61), and improved their confidence in performing segmentation tasks (Q9: 4.58±0.51). The RAG mechanism played a key role here. By comparing the queried radiological slice to both healthy and pathological references, users received contextualized, anatomy-specific guidance that was grounded in real cases. This was especially reflected in high agreement with Q8, where workflow design was rated highest (4.84±0.37). Interestingly, the AI-generated textual explanations (Q6) and reference images (Q5) received slightly lower scores (3.89±0.74 and 3.63±0.96, respectively), possibly reflecting uncertainty and confusion in some participants, who in general did not have strong familiarity with human anatomy (3.42 ± 1.22, 1 = unfamiliar, 5 = familiar). To improve comfort and presence during voice interaction, future versions of the system could feature a visual avatar for the assistant, helping to reduce the slight awkwardness some users felt when speaking to a disembodied voice.

Furthermore, it was evident in the segmentation results that users demonstrated a clear understanding of when to intervene and when to trust the AI system. For example, with the simpler liver tumor CT case, users made minimal edits, and the refined masks were similar to unrefined ones. Yet importantly, performance did not degrade after user interaction, indicating that users did not over-correct or introduce noise. This suggests a healthy level of trust and restraint, and a true understanding of segmentation quality based on visual features and reference images. In contrast, the more challenging brain tumor MRI case, which included branching outer boundaries, showed a significant accuracy improvement after refinement. The 3D Dice score rose from $90.53 \pm 10.51\%$

to $94.92 \pm 0.52\%$ post-interaction ($p < 0.001$), demonstrating that users could identify areas needing correction and effectively apply point-based refinements. This reinforces that users not only understand segmentation correctness, but can also meaningfully enhance AI outputs in cases where human expertise is still necessary.

Together, these results position our system not only as a segmentation tool, but as a supportive assistant capable of accelerating workflows, teaching radiological features of pathology, and fostering trust with clinical AI. The positive responses to confidence and task understanding, combined with high segmentation accuracy, suggest that such a tool has potential in both clinical workflows and medical education. Future work may explore expanding the agent to work with more radiological concepts, or adaptive RAG responses based on user skill level (i.e. beginners versus experienced radiologists).

4.7 Conclusion

We introduced a novel VR system for interactive medical image segmentation that integrates foundation models with attention-switching interaction and a supportive conversational AI agent, SAMIRA. At the core of our method is a novel segmentation algorithm that combines BiomedParse's language-driven detection with a medical-image-adapted SAM2 model. To adapt SAM2 for clinical imaging, we introduced a novel IoU-based stopping criterion in its memory mechanism to prevent drift across noisy or low-contrast slices. Our findings show that this criterion can significantly improve segmentation quality, and that by using it with SAMIRA, users can efficiently achieve high segmentation accuracy with minimal effort across all interaction paradigms. Of these interaction paradigms, *Controller* pointing offers the best overall balance of accuracy, speed, and task load, closely followed by head-pointing. More importantly, users demonstrated a clear ability to interpret and refine AI outputs based on its generated guidance, engaging critically with reference images, contextual explanations, and 3D visualization.

Importantly, this system is generalizable to any 3D medical image aligned with BiomedParse's training scope and can be expanded by enriching its RAG knowledge database. Its modular design, high usability, and adaptability to various interaction styles position it as a powerful tool for clinical

workflows, but also offers insights for future HCI research in intelligent, immersive medical systems with clinical AI agents.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis presents two complementary frameworks to lower the barriers to high-quality medical image segmentation by investigating foundation deep learning models and human-centered interaction:

First, the YOLO-URSAM weakly supervised ICH segmentation pipeline leverages a finetuned YOLOv8 detector to generate box- and morphology-based point prompts, and applies an uncertainty rectified Segment Anything Model to produce accurate intracranial hemorrhage masks without pixel-level labels. On public CT datasets, YOLO-URSAM significantly outperformed both weakly supervised baselines (e.g., Swin-HGI-SAM) and popular fully supervised models (U-Net, Swin-UNETR), achieving a Dice score of approximately 0.63 versus 0.40–0.43 ($p < 0.005$) while needing only bounding-box annotations. Second, the SAMIRA interactive 3D VR segmentation system introduced a conversational AI agent combining a vision–language foundation model (Biomed-Parse) for speech-initiated mask seeding with SAM2 for point-prompt refinement and slice-wise propagation. Meanwhile, retrieval augmented generation provides case-specific guidance via contrastive reference images and patient-tailored explanations. In our user study (N=19), SAMIRA achieved high usability (SUS=90.0±9.0) and improved Dice from 0.91 to 0.95 ($p < 0.001$) from

human-in-the-loop refinement on challenging brain tumor cases, all with minimal user effort. Together, these contributions demonstrate that weak supervision can approach fully supervised accuracy with just coarse labels, and that immersive, agent-driven workflows can facilitate segmentation creation. Furthermore, from a clinical standpoint, YOLO-URSAM has the potential to support more rapid hemorrhage quantification in stroke settings, and SAMIRA's immersive, agent-driven workflow can educate trainees and shorten segmentation times while preserving expert oversight.

5.2 Future Work

In the future, it is possible to extend YOLO-URSAM and SAMIRA to other pathologies and modalities. This could be achieved by enriching prompting strategies and the RAG knowledge base, respectively. As the RAG index grows beyond hundreds of thousands of vectors, we will evaluate FAISS's scalable approximate indexing methods (e.g. IVF-PQ, HNSW) to maintain sub-millisecond retrieval times, since responsiveness is critical in immersive environments. Furthermore, it would be valuable to conduct clinical studies with radiologists and surgeons to assess integration of the real world workflow, diagnostic accuracy, and time saved. For YOLO-URSAM specifically, it would be beneficial to perform a sensitivity analysis of the majority voting scheme, which could involve systematically varying the majority-voting threshold (e.g. $k = 3, 5, 7, 9$), and determining its effect on segmentation accuracy and runtime. Furthermore, investigating 3D SAM adaptations in YOLO-URSAM's framework (e.g. 3D Transformers or CNN-Transformer hybrids) [79] may demonstrate improvements in performance compared to using vanilla SAM. For SAMIRA, it would be insightful to investigate the system with a cohort of medical students to measure improvements in anatomical education. Additionally, extending SAMIRA to work with 3D ultrasound could be particularly useful for training, considering the modality's popularity in clinics.

Appendix A

Ethics Approval: Towards user-centered interactive medical image segmentation in VR with an assistive AI agent

A.1 Ethics Approval Form

CERTIFICATION OF ETHICAL ACCEPTABILITY
FOR RESEARCH INVOLVING HUMAN SUBJECTS

Name of Applicant:	Dr. Yiming Xiao
Department:	Gina Cody School of Engineering and Computer Science\Computer Science and Software Engineering
Agency:	Concordia University
Title of Project:	Evaluating Interaction Paradigms for an AI Agent in VR for Medical Image Segmentation
Certification Number:	30021356

Valid From: January 30, 2025 To: January 29, 2026

The members of the University Human Research Ethics Committee have examined the application for a grant to support the above-named project, and consider the experimental procedures, as outlined by the applicant, to be acceptable on ethical grounds for research involving human subjects.

Richard DeMont

Dr. Richard DeMont, Chair, University Human Research Ethics Committee

Bibliography

[1] Robert J Adams, Lawrence J Appel, Lynne T Braun, Seemant Chaturvedi, Mark A Creager, Antonio Culebras, Robert H Eckel, Robert G Hart, Judith A Hinchey, Virginia J Howard, et al. Aha/asa guideline. *Guidelines for early management of adults with ischemic stroke*, 2007:38, 2010.

[2] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014.

[3] Maria I. Aguilar and Thomas G. Brott. Update in intracerebral hemorrhage. *The Neurohospitalist*, 1(3):148–159, 2011.

[4] Ali Al Bataineh, Devinder Kaur, Mahmood Al-khassaweneh, and Esraa Al-sharoa. Automated cnn architectural design: A simple and efficient methodology for computer vision tasks. *Mathematics*, 11(5):1141, 2023.

[5] Majid Behravan, Krešimir Matković, and Denis Gračanin. Generative ai for context-aware 3d object creation using vision-language models in augmented reality. In *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pages 73–81. IEEE, 2025.

[6] Matt A Bernstein, Kevin F King, and Xiaohong Joe Zhou. *Handbook of MRI pulse sequences*. Elsevier, 2004.

[7] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.

[8] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[9] John Brooke. Sus: a retrospective. *Journal of usability studies*, 8(2), 2013.

[10] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189 (194):4–7, 1996.

[11] Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[12] Michael A Bruno, Evan A Walker, and Hani H Abujudeh. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676, 2015. doi: 10.1148/rg.2015150023.

[13] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis*, 71: 102062, 2021.

[14] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[15] Peter D Chang, Edward Kuoy, Jack Grinband, Brent D Weinberg, Matthew Thompson, Richelle Homo, Jefferson Chen, Hermelinda Abcede, Mohammad Shafie, Leo Sugrue, et al. Hybrid 3d/2d convolutional neural network for hemorrhage evaluation on head ct. *American Journal of Neuroradiology*, 39(9):1609–1616, 2018.

[16] Soudeh Chegini, Clare Schilling, Matthew Clarkson, and Eddie Edwards. 63. a training tool for clinicians in segmenting medical images to make 3d models. *British Journal of Oral and Maxillofacial Surgery*, 60(10):e107–e108, 2022.

[17] Jieyu Cheng, He Li, Feng Xiao, Aaron Fenster, Xuming Zhang, Xiaoling He, Ling Li, and Mingyue Ding. Fully automatic plaque segmentation in 3-d carotid ultrasound images. *Ultrasound in medicine & biology*, 39(12):2431–2446, 2013.

[18] Vuthea Chheang, Shayla Sharmin, Rommy Márquez-Hernández, et al. Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments. In *Proc. IEEE Int. Conf. on Artificial Intelligence and Virtual Reality (AIVR)*, pages 21–30, 2024. doi: 10.1109/AIxVR59861.2024.00011.

[19] Junghwan Cho, Ki-Su Park, Manohar Karki, Eunmi Lee, Seokhwan Ko, Jong Kun Kim, Dongeun Lee, Jaeyoung Choe, Jeongwoo Son, Myungsoo Kim, et al. Improving sensitivity on identification and delineation of intracranial hemorrhage lesion using cascaded deep learning models. *Journal of digital imaging*, 32:450–461, 2019.

[20] Airton Leonardo de Oliveira Manoel. Surgery for spontaneous intracerebral hemorrhage. *Critical Care*, 24(1):45, 2020.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

[24] Ömer Faruk Ertuğrul and Muhammed Fatih Akıl. Detecting hemorrhage types and bounding box of hemorrhage by deep learning. *Biomedical Signal Processing and Control*, 71:103085, 2022.

[25] Athanasios Evagelou, Alexandros Kleftodimos, Magdalini Grigoriou, and Georgios Lappas. Augmented reality for natural heritage education: A design framework for enhancing indoor experiences. *Heritage*, 8(6):191, 2025.

[26] Ajoy S Fernandes, T Scott Murdison, and Michael J Proulx. Looking in depth: Targeting by eye and controller input for multi-depth target placement. *International Journal of Human–Computer Interaction*, pages 1–16, 2024.

[27] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

[28] Yi Gao, Cheng Chang, Xiaxia Yu, Pengjin Pang, Nian Xiong, and Chuan Huang. A vr-based volumetric medical image segmentation and visualization system with natural human interaction. *Virtual Reality*, 26(2):415–424, 2022.

[29] Santiago González Izard, Ramiro Sánchez Torres, Oscar Alonso Plaza, Juan Antonio Juanes Mendez, and Francisco José García-Peñalvo. Nextmed: automatic imaging segmentation, 3d reconstruction, and 3d model visualization platform using augmented and virtual reality. *Sensors*, 20(10):2962, 2020.

[30] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[31] Kyle Greenway, Henry Knipe, Frank Gaillard, Arlene Campos, et al. Hounsfield unit. https://doi.org/10.53347/rID-38181, 2024. URL https://radiopaedia.org/articles/38181. Reference article, Radiopaedia.org (Accessed on 28 Jul 2025).

[32] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[33] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[34] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[35] JJ Heit, H Coelho, FO Lima, M Granja, A Aghaebrahim, R Hanel, K Kwok, H Haerian, CW Cereda, C Venkatasubramanian, et al. Automated cerebral hemorrhage detection using rapid. *American Journal of Neuroradiology*, 42(2):273–278, 2021.

[36] Owen Hellum, Marta Kersten-Oertel, and Yiming Xiao. Assessment of user-interaction strategies for neurosurgical data navigation and annotation in virtual reality. *Virtual Reality*, 27(2):1345–1355, 2023.

[37] Murtadha D. Hssayeni, Muayad S. Croock, Aymen D. Salman, Hassan Falah Al-khafaji, Zakaria A. Yahya, and Behnaz Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1), 2020. ISSN 2306-5729.

[38] David J. Seiffge Isabel C. Hostettler and David J. Werring. Intracerebral hemorrhage: an update on diagnosis and treatment. *Expert Review of Neurotherapeutics*, 19(7):679–694, 2019. PMID: 31188036.

[39] Santiago González Izard, Ramiro Sánchez Torres, Óscar Alonso Plaza, Juan Antonio Juanes Méndez, and Francisco José García-Peñalvo. Nextmed: automatic imaging segmentation, 3d reconstruction, and 3d model visualization platform using augmented and virtual reality. *Sensors (Basel, Switzerland)*, 20(10):2962, 2020.

[40] Xiyao Jin, Yao Hao, Jessica Hilliard, Zhehao Zhang, Maria A Thomas, Hua Li, Abhinav K Jha, and Geoffrey D Hugo. A quality assurance framework for routine monitoring of deep learning cardiac substructure computed tomography segmentation models in radiotherapy. *Medical physics*, 51(4):2741–2758, 2024.

[41] Maria-Ruxandra Jinga, Rachel BY Lee, Kai Lok Chan, Prabhvir S Marway, Krishan Nandapalan, Kawal Rhode, Christopher Kui, and Matthew Lee. Assessing the impact of 3d image segmentation workshops on anatomical education and image interpretation: A prospective pilot study. *Anatomical Sciences Education*, 16(6):1024–1032, 2023.

[42] Willi A Kalender. *Computed tomography: fundamentals, system technology, image quality, applications*. John Wiley & Sons, 2011.

[43] Mehr Kashyap, Xi Wang, Neil Panjwani, Mohammad Hasan, Qin Zhang, Charles Huang, Karl Bush, Alexander Chin, Lucas K Vitzthum, Peng Dong, et al. Automated deep learning–based detection and segmentation of lung tumors at ct imaging. *Radiology*, 314(1):e233029, 2025.

[44] JH Kim, KH Jun, and HM Chin. Prognostic value of calculated tumor volume in patients with gastric cancer. *Clin Surg. 2021; 6*, 3373.

[45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[47] Alexander Kirillov et al. Sam 2: Segment anything, any time, all at once. *arXiv preprint arXiv:2404.07143*, 2024.

[48] Mikhail Konenkov, Artem Lykov, Daria Trinitatova, and Dzmitry Tsetserukou. Vr-gpt: Visual language model for intelligent virtual reality applications. *arXiv preprint arXiv:2405.11537*, 2024.

[49] Weicheng Kuo, Christian Häne, Esther Yuh, Pratik Mukherjee, and Jitendra Malik. Cost-sensitive active learning for intracranial hemorrhage detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pages 715–723. Springer, 2018.

[50] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H Tajmir, Claude E Guerrier, Sarah A Ebert, Stuart R Pomerantz, Javier M Romero, Shahmir Kamalian, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature biomedical engineering*, 3(3):173–182, 2019.

[51] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[52] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in neural information processing systems*, 33:21002–21012, 2020.

[53] Sok Ying Liaw, Jian Zhi Tan, Khairul Dzakirin Bin Rusli, Rabindra Ratan, Wentao Zhou, Siriwan Lim, Tang Ching Lau, Betsy Seah, and Wei Ling Chua. Artificial intelligence versus human-controlled doctor in virtual reality simulation for sepsis team training: randomized controlled study. *Journal of medical Internet research*, 25:e47748, 2023.

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[55] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[56] Hong Liu, Haosen Yang, Paul J van Diest, Josien PW Pluim, and Mitko Veta. Wsi-sam: Multi-resolution segment anything model (sam) for histopathology whole-slide images. *arXiv preprint arXiv:2403.09257*, 2024.

[57] Xiang Liu, Rui Wang, Zemin Zhu, Kexin Wang, Yue Gao, Jialun Li, Yaofeng Zhang, Xiangpeng Wang, Xiaodong Zhang, and Xiaoying Wang. Automatic segmentation of hepatic metastases on dwi images based on a deep learning method: assessment of tumor treatment response according to the recist 1.1 criteria. *BMC cancer*, 22(1):1285, 2022.

[58] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[59] Melissa A LoPresti, Samuel S Bruce, Elvis Camacho, Sudkir Kunchala, Byron G Dubois, Eliza Bruce, Geoff Appelboom, and E Sander Connolly Jr. Hematoma volume as the major determinant of outcomes after intracerebral hemorrhage. *Journal of the neurological sciences*, 345(1-2):3–7, 2014.

[60] Francisco Lopez Luro and Veronica Sundstedt. A comparative study of eye tracking and hand controller for aiming tasks in virtual reality. In *Proceedings of the 11th ACM Symposium on eye tracking research & applications*, pages 1–9, 2019.

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

[62] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[63] Paul Milgram, Haruo Takemura, Akira Utsumi, and Fumio Kishino. Augmented reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and telepresence technologies*, volume 2351, pages 282–292. Spie, 1995.

[64] A. Murphy, J. Feger, M. Ismail, et al. Windowing (ct). Radiopaedia.org reference article, 2025. URL https://radiopaedia.org/articles/52108. Accessed 22 July 2025.

[65] Ain Atiqa Mustapha, Sarah 'Atifah Saruchi, Heru Supriyono, and Mahmud Iwan Solihin. A hybrid deep learning model for waste detection and classification utilizing yolov8 and cnn. *Engineering Proceedings*, 84(1):82, 2025.

[66] Aunnoy K Mutasim, Anil Ufuk Batmaz, and Wolfgang Stuerzlinger. Pinch, click, or dwell: Comparing different selection techniques for eye-gaze-based pointing in virtual reality. In *Acm symposium on eye tracking research and applications*, pages 1–7, 2021.

[67] Jakub Nemcek, Tomas Vicar, and Roman Jakubicek. Weakly supervised deep learning-based intracranial hemorrhage localization. *arXiv preprint arXiv:2105.00781*, 2021.

[68] Takeru Ohtaka, Ken Ando, Takahiro Oike, Shin-ei Noda, Takuya Kaminuma, Kazutoshi Murata, and Tatsuya Ohno. The prognostic effect of tumor volume, reduction ratio, and cumulative doses on external beam radiotherapy with central-shielding method and image-guided adaptive brachytherapy for cervical cancer. *Frontiers in Oncology*, 14:1366777, 2024.

[69] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[70] Qiumei Pu, Zuoxin Xi, Shuai Yin, Zhe Zhao, and Lina Zhao. Advantages of transformer and its application for medical image segmentation: a survey. *BioMedical engineering online*, 23 (1):14, 2024.

[71] AI Qureshi and YY Palesch. Antihypertensive treatment of acute cerebral hemorrhage (atach) ii: design, methods, and rationale. *Neurocritical care*, 15:559–576, 2011.

[72] Divya Ramakrishnan, Leon Jekel, Saahil Chadha, Anastasia Janas, Harrison Moy, Nazanin Maleki, Matthew Sala, Manpreet Kaur, Gabriel Cassinelli Petersen, Sara Merkaj, et al. A large open access dataset of brain metastasis 3d segmentations on mri with clinical and imaging information. *Scientific Data*, 11(1):254, 2024.

[73] Umar Rashid, Jarmo Kauko, Jonna Häkkilä, and Aaron Quigley. Proximal and distal selection of widgets: designing distributed ui for mobile interaction with large display. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 495–498, 2011.

[74] Amirhossein Rasoulian, Soorena Salari, and Yiming Xiao. Weakly supervised intracranial hemorrhage segmentation using hierarchical combination of attention maps from a swin transformer. In *Machine Learning in Clinical Neuroimaging*, pages 63–72, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-17899-3.

[75] Amirhossein Rasoulian, Soorena Salari, and Yiming Xiao. Weakly Supervised Intracranial Hemorrhage Segmentation using Head-Wise Gradient-Infused Self-Attention Maps from a Swin Transformer in Categorical Learning. *arXiv (Cornell University)*, 1 2023.

[76] Eduardo Reis, Felipe Nascimento, Mateus Aranha, Fernando Secol, Birajara Machado, Marcelo Felix, Anouk Stein, and Edson Amaro. Brain hemorrhage extended (bhx): Bounding box extrapolation from thick to thin slice ct images. 07 2020.

[77] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[78] Chuyun Shen, Wenhao Li, Yuhang Shi, and Xiangfeng Wang. Interactive 3d medical image segmentation with sam 2, 2025. URL https://arxiv.org/abs/2408.02635.

[79] Yiqing Shen, Jingxing Li, Xinyuan Shao, Blanca Inigo Romillo, Ankush Jindal, David Dreizin, and Mathias Unberath. Fastsam3d: An efficient segment anything model for 3d

volumetric medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 542–552. Springer, 2024.

[80] Ludwig Sidenmark, Franziska Prummer, Joshua Newn, and Hans Gellersen. Comparing gaze, head and controller selection of dynamically revealed targets in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4740–4750, 2023. doi: 10.1109/TVCG.2023.3320235.

[81] Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3): 143–155, 2002.

[82] Pascal Spiegler, Amirhossein Rasoulian, and Yiming Xiao. Weakly supervised intracranial hemorrhage segmentation with yolo and an uncertainty rectified segment anything model. In *MICCAI Challenge on Ischemic Stroke Lesion Segmentation*, pages 12–21. Springer, 2024.

[83] Pascal Spiegler, Arash Harirpoush, and Yiming Xiao. Towards user-centered interactive medical image segmentation in vr with an assistive ai agent. *arXiv preprint arXiv:2505.07214*, 2025.

[84] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[85] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[86] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, 2023.

[87] IJ Torres, AJ Mundt, PJ Sweeney, S Llanes-Macy, L Dunaway, M Castillo, and RL Macdonald. A longitudinal neuropsychological study of partial brain radiation in adults with brain tumors. *Neurology*, 60(7):1113–1118, 2003.

[88] Constanza Vásquez-Venegas, Camilo G Sotomayor, Baltasar Ramos, Víctor Castañeda, Gonzalo Pereira, Guillermo Cabrera-Vives, and Steffen Härtel. Human-in-the-loop—a deep learning strategy in combination with a patient-specific gaussian mixture model leads to the fast characterization of volumetric ground-glass opacity and consolidation in the computed tomography scans of covid-19 patients. *Journal of clinical medicine*, 13(17):5231, 2024.

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[90] Justin L Wang, Hassan Farooq, Hanqi Zhuang, and Ali K Ibrahim. Segmentation of intracranial hemorrhage using semi-supervised multi-task attention-based u-net. *Applied Sciences*, 10(9):3297, 2020.

[91] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[92] Kai Wu, Bowen Du, Man Luo, Hongkai Wen, Yiran Shen, and Jianfeng Feng. Weakly supervised brain lesion segmentation via attentional representation learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 211–219. Springer, 2019.

[93] Wenge Xu, Xuanru Meng, Kangyou Yu, Sayan Sarcar, and Hai-Ning Liang. Evaluation of text selection techniques in virtual reality head-mounted displays. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 131–140. IEEE, 2022.

[94] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022.

[95] Jinzhong Yang, Greg Sharp, Harini Veeraraghavan, Wouter Van Elmpt, Andre Dekker, Tim Lustberg, and Mark Gooding. Data from lung ct segmentation challenge 2017 (lctsc), 2017. URL https://www.cancerimagingarchive.net/collection/lctsc/.

[96] Shea B Yonker, Oleksandr O Korshak, Timothy Hedstrom, Alexander Wu, Siddharth Atre, and Jürgen P Schulze. 3d medical image segmentation in virtual reality. *Electronic Imaging*, 31:1–6, 2019.

[97] Chen Zhang, Xiangyao Deng, and Sai Ho Ling. Next-gen medical imaging: U-net evolution and the rise of transformers. *Sensors*, 24(14):4668, 2024.

[98] Theodore Zhao, Yu Gu, Jianwei Yang, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *Nature Methods*, 22:166–176, 2025. doi: 10.1038/s41592-024-02499-w.

[99] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[100] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

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