TESSERACT: TExt-Supervised SEgmentation for RAdiology with Cross-attentional Transformers

Rishov Paul (vst2hb)[1] and Swakshar Deb (swd9tc)[2]

[1]*Department of Computer Science, University of Virginia*
[2]*Department of Electrical and Computer Engineering, University of Virginia*

Abstract: Volumetric medical image segmentation underpins a wide range of clinical applications, yet existing methods are fundamentally limited in their ability to incorporate natural language as a flexible form of guidance. While recent approaches have explored promptable segmentation using dense annotations or weak supervision cues such as keypoints and bounding boxes, they are not designed to interpret or align free-form textual descriptions with anatomical structures. Critically, current models lack mechanisms for learning explicit semantic correspondences between language and volumetric regions, leaving text-driven 3D segmentation largely unexplored in the literature. In this work, we propose TESSERACT, a novel framework for volumetric segmentation that leverages natural language descriptions to predict anatomical delineations across full 3D scans. TESSERACT processes volumetric data in a slice-wise manner using a 2D encoder-decoder backbone supervised by a segmentation objective, and it aligns each slice with promptable textual context via a cross-attention mechanism between visual features and clinical language embeddings in the latent space. This design enables voxel-level semantic guidance while maintaining spatial consistency throughout the volume. Our method supports flexible, prompt-driven segmentation and demonstrates strong performance in capturing complex anatomical structures based on descriptive text, offering a new paradigm for multimodal medical image understanding.

1. Introduction

Precise segmentation of anatomical structures in volumetric medical images is central to modern clinical workflows, underpinning applications such as computer-assisted diagnosis, radiotherapy planning, and longitudinal disease monitoring [1–3]. Conventional fully-supervised approaches—whether employing volumetric 3D CNNs [4] or slice-wise 2D architectures [5, 6]—focus exclusively on visual signals and ignore rich, structured information readily available in clinical texts. Radiology reports, for instance, often contain spatial and pathological descriptors (e.g., "hyperintense lesion anterior to the left hippocampus") that could serve as strong priors for visual disambiguation. Integrating such semantic context could enhance segmentation performance in rare or ambiguous cases, enable zero-shot segmentation of arbitrary structures, and facilitate natural-language-based interaction with clinicians. Despite these potential benefits, existing segmentation models do not leverage this textual supervision, largely due to the difficulty of aligning sparse language with dense image grids.

Recent advances in vision–language pretraining (e.g., CLIP [7], BLIP-2 [8]) have demonstrated impressive alignment capabilities in natural images and inspired text-conditioned video segmentation models [9, 10]. However, extending these methods to volumetric medical imaging presents unique challenges. First, effective multimodal fusion remains an open problem: global sentence-image embeddings are too coarse to guide fine-grained voxel-level predictions, necessitating new architectures for local cross-modal interaction. Second, medical imaging data are inherently high-dimensional, and naïve application of 3D attention mechanisms incurs prohibitive memory costs, while slice-wise models sacrifice long-range consistency. Third, unlike natural video datasets with clear object categories and motion cues, clinical language is abstract, variable, and often class-ambiguous—demanding models that generalize to open-set prompts without retraining. Thus, there is a need for a scalable, semantically aligned framework capable of fusing language and vision at appropriate spatial resolutions.

Our approach. To address these limitations we introduce **TESSERACT** [1], a novel text-guided medical image segmentation framework that jointly learns aligned image and text representations in a shared latent space (Fig. 1).

[1] **T**ext-**E**nhanced **S**lice-wise **SE**gmentation with **R**epresentational **A**lignment in **C**ross-modal **T**ensors

TESSERACT leverages an efficient 2D UNet-ResNet34 encoder to extract slice-level features and a lightweight BioBERT encoder to embed clinical prompts. A *Text-Guided Cross-Attention* (TGCA) module projects these signals into a common subspace and modulates visual features according to the textual description. The resulting text-conditioned feature maps are decoded to produce a binary or multi-class mask for every slice, which are then stacked to form the volumetric prediction. By operating slice-wise, TESSERACT scales to large 3D scans while retaining global semantic coherence through prompt replication across depth.

Contributions. Our work brings the following contributions to multimodal medical image analysis:

- We present *TESSERACT*, the first end-to-end model for *text-guided volumetric* medical image segmentation, enabling natural-language interaction with 3D scans without requiring any architectural changes to the underlying backbone.
- We introduce the *Text-Guided Cross-Attention* module that aligns slice-wise visual features with BioBERT embeddings, yielding semantically enriched representations that improve robustness to rare anatomies and ambiguous appearances.
- Extensive experiments on publicly available OASIS brain MRI dataset with various text prompt and competitive inference speed on a single GPU.

Taken together, these advances position TESSERACT as a promising step toward context-aware, clinically aligned segmentation systems that can understand and follow free-form radiological descriptions. We release code, pre-trained weights, and evaluation scripts to facilitate further research.

2. Related Works

To incorporate the contextual understanding in volumetric segmentation task, recent studies have begun incorporating textual guidance into segmentation pipelines. Inspired by the success of vision-language models in natural image domains (e.g., CLIP), researchers have proposed adapting language priors to medical segmentation tasks. One such effort is Text-Knowledge-guided Segment Anything Model (TK-SAM) [11], which extends the general-purpose Segment Anything Model (SAM) by incorporating domain-specific text embeddings. By integrating clinical prompts, TK-SAM improves the model's ability to localize anatomical structures without requiring full supervision. However, SAM-based approaches often rely on 2D slice-level inference and lack mechanisms for preserving spatial consistency across 3D volumes. In a different direction, Text-guided Diffusion Models for segmentation [12–14] have emerged as promising tools for label-efficient segmentation. For instance, Enhancing Label-efficient Medical Image Segmentation with Text-guided Diffusion Models [14] uses free-text anatomical descriptions as guidance within a diffusion-based generative framework. By conditioning image generation and segmentation on textual input, the model reduces reliance on dense annotations. Similarly, DiffBoost [12] employs controlled diffusion guided by textual prompts to synthesize anatomically plausible training data, enriching segmentation models under data-scarce conditions. Although these approaches show encouraging results in few-shot and semi-supervised setups. they are primarily mask generation model rather than pure segmentaiton. Moreover, the progressive denoising in the diffusion model [15] hinder their applicability in real time medical segmentation setup due to significant sampling time, which is crucial for real-world diagnosis. TG-LMM (Text-Guided Large Multi-modal Model) [16] represents another major step toward integrating expert knowledge into segmentation networks. It employs pre-trained large language and vision models to fuse textual descriptions with spatial image features, enabling fine-grained semantic alignment. While TG-LMM demonstrates strong performance across multiple datasets, its architecture is not explicitly designed for volumetric processing and lacks strategies for maintaining anatomical coherence across slices.

Despite recent progress, current text-guided segmentation methods face key limitations: *(i)* most are constrained to 2D slice-wise processing and do not generalize well to high-resolution volumetric data; *(ii)* many treat text either as a static global prompt or rely on asymmetric attention, leading to suboptimal feature fusion; *(iii)* multimodal alignment is often coarse and lacks voxel-level precision required for clinical applicability. To address these challenges, we propose TESSERACT—a novel framework for text-guided volumetric segmentation that performs slice-wise processing while maintaining spatial coherence across the 3D scan. TESSERACT learns a shared latent space between clinical descriptions and visual features using cross-attention, enabling precise, prompt-driven segmentation at the voxel level. Unlike prior methods, our model iteratively integrates text-conditioned features throughout the decoder pathway and scales efficiently to volumetric data, offering improved performance and interpretability in complex clinical scenarios.

3. Our Method: TESSERACT

This section introduces TESSERACT, a novel text-guided medical image segmentation model that uniquely learns a multimodal alignment between image and text representations in an integrated latent space. By integrating latent textual guidance into the segmentation network, TESSERACT enables text-conditioned anatomical segmentation, effectively transforming conventional image-only segmentation models into multimodal systems capable of leveraging descriptive clinical context for improved performance and interoperability.

Problem Setup. We formulate our text-guided medical image segmentation task as a joint optimization problem across image and text modalities, specifically enforcing the representation learning in a shared integrated latent space $\mathcal{Z}$. Given a dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$ consisting of N training samples, where each sample is a triplet containing an input image $x_i \in \mathcal{X} \subset \mathbb{R}^{H\times W\times C\times D}$ (with height H, width W, depth D, and C channels), corresponding text description $t_i \in \mathcal{T}$, and ground truth segmentation mask $y_i \in \mathcal{Y}$. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ denote the set of text descriptions, where n is the cardinality of unique anatomical region descriptors in our corpus. Each t_i is a natural language description (a.k.a. *prompt*) encoded as a sequence of tokens $t_i = [w_1, w_2, \ldots, w_m]$, where m represents the maximum sequence length.

The output space $\mathcal{Y} \subset \{0,1\}^{H\times W\times D\times K}$ represents volumetric multi-class segmentation masks with K distinct anatomical classes. The conventional image segmentation paradigm aims to learn a mapping function $f : \mathcal{X} \to \mathcal{Y}$. In contrast, our proposed TESSERACT framework learns a text-conditioned mapping $g : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$, incorporating contextual textual guidance to improve segmentation fidelity and semantic correspondence through cross-modal alignment in the latent space $\mathcal{Z}$.

3.1. Pixel-based Segmentation

Given an input image $x \in \mathcal{X} \subset \mathbb{R}^{H\times W\times C\times D}$, a traditional segmentation model f_θ with parameters θ aims to learn:

$$f_\theta : \mathcal{X} \to \mathcal{Y}, \tag{1}$$

where $\mathcal{Y} \subset \{0,1\}^{H\times W\times D\times K}$ represents the space of volumetric segmentation masks with K anatomical classes. We optimize the f_θ, parameterized by θ by minimizing a pixel-wise loss function $\mathcal{L}_{seg}$ between predicted segmentations $\hat{y} = f_\theta(x)$ and ground truth annotations y:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}_{\text{seg}}(f_\theta(x), y)\right]. \tag{2}$$

For multi-class segmentation with K classes, the pixel-wise cross-entropy loss $\mathcal{L}_{CE}$ is defined as:

$$\mathcal{L}_{CE}(p, y) = -\frac{1}{|\Omega|} \sum_{i\in\Omega} \sum_{k=1}^{K} y_i^k \log(p_i^k), \tag{3}$$

where Ω represents the set of all voxels in the image volume, $y_i^k \in \{0,1\}$ indicates whether voxel i belongs to class k in the ground truth, and p_i^k is the predicted probability of voxel i belonging to class k. The Dice loss $\mathcal{L}_{Dice}$ measures the volumetric overlap between predictions and ground truth:

$$\mathcal{L}_{Dice}(p, y) = 1 - \frac{1}{K} \sum_{k=1}^{K} \frac{2\sum_{i\in\Omega} p_i^k y_i^k}{\sum_{i\in\Omega}(p_i^k)^2 + \sum_{i\in\Omega}(y_i^k)^2 + \varepsilon}, \tag{4}$$

where ε is a small constant added for numerical stability. Defining λ_{CE} and λ_{Dice} as hyperparameters that balance the contribution of each loss component, we formulate the combined segmentation loss as:

$$\mathcal{L}_{seg} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{Dice}\mathcal{L}_{Dice} \tag{5}$$

Traditional volumetric segmentation models operate exclusively in the visual domain, lacking the ability to incorporate rich textual descriptions that could guide the segmentation process []. This fundamental limitation restricts their semantic understanding, flexibility in targeting arbitrary structures, and adaptability to rare anatomical variants. We address these limitations by introducing a novel alignment mechanism that projects both volumetric image data and text descriptions into an integrated latent space $\mathcal{Z}$, enabling seamless integration of multimodal information for improved segmentation performance.

3.2. Slice-wise Volumetric Segmentation with 2D Networks

A key design choice in our TESSERACT framework is to perform volumetric segmentation in a slice-wise manner using 2D networks, which offers significant advantages in terms of computational efficiency, memory usage, and

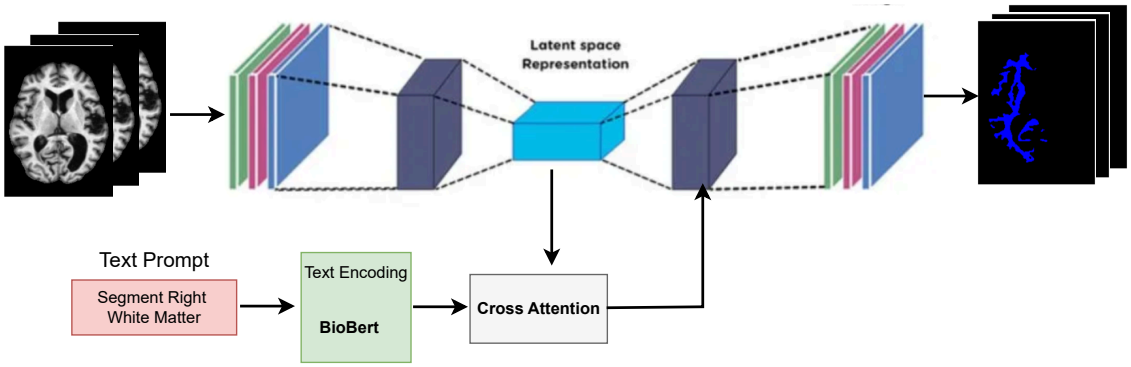

Fig. 1. Overview of the proposed **TESSERACT** framework for text-guided volumetric medical image segmentation. Given a 3D medical image and a corresponding natural language prompt describing a target anatomical structure, the volume is sliced along the depth axis and processed using a 2D encoder-decoder architecture. Visual features from each 2D slice are aligned with textual embeddings using a cross-attention mechanism in a shared latent space. The aligned features are then decoded to generate 2D segmentation masks, which are stacked to produce the final volumetric segmentation output. This multimodal design enables precise, prompt-driven anatomical localization and segmentation.

leveraging pre-trained 2D architectures. We decompose the volumetric segmentation problem into a series of 2D segmentation tasks by transforming the input volume $x \in \mathbb{R}^{H\times W\times C\times D}$ into a set of 2D slices $\{x^d\}_{d=1}^{D}$, where each $x^d \in \mathbb{R}^{H\times W\times C}$ represents a single slice along the depth dimension. For a given slice x^d and text description t, our model produces a 2D segmentation mask $\hat{y}^d \in \{0,1\}^{H\times W\times K}$. We predict the final volumetric segmentation $\hat{y} \in \{0,1\}^{H\times W\times D\times K}$ by stacking the predicted 2D masks along the depth dimension. This approach allows us to effectively process 3D medical images while utilizing the efficiency of 2D convolutional architectures. During training, we reshape the 3D volumetric data as $x_{reshaped} = \text{reshape}(x) \in \mathbb{R}^{B\cdot D\times C\times H\times W}$, where B is the batch size. This reshaping operation effectively treats each slice as an independent 2D image, increasing the effective batch size by a factor of D and enabling efficient processing using 2D convolutional networks.

3.3. Multimodal Latent Space Alignment

We design a cross-attention mechanism that aligns 2D visual features with textual embeddings. Unlike traditional approaches that operate solely in the visual domain [], our model incorporates textual guidance through a Text-Guided Cross-Attention module. Given pretrained encoders for both modalities, we extract deep visual features from a UNet-ResNet34 backbone [] for each 2D slice and textual embeddings from BioBERT [].

To extract the semantic information considering the memory and time efficiency, we develop the multimodal feature alignment at the latent spaces of the visual and text encoders. Let $F_I \in \mathbb{R}^{(B\cdot D)\times C\times H'\times W'}$ represent the visual features from the final encoder layer of our 2D network, where B is the batch size, D is the depth dimension of the original volume, $C = 512$ is the channel dimension, and H', W' are the reduced spatial dimensions. Let $E_T \in \mathbb{R}^{B\times L\times D_t}$ denote the textual embeddings from BioBERT, where L is the sequence length and $D_t = 768$ is the embedding dimension. Since our visual features are processed slice-wise while the text description applies to the entire volume, we replicate the text embeddings for each slice:

$$E_T^{'} \in \mathbb{R}^{(B\cdot D)\times L\times D_t} \tag{6}$$

Our Text-Guided Cross-Attention module transforms these representations as follows $\mathbf{Q} = W_Q F_I, \quad \mathbf{K} = W_K E_T^{'}, \quad \mathbf{V} = W_V E_T^{'}$, where W_Q, W_K, and W_V are learnable projection matrices that transform the visual features and textual embeddings into a common projection space of dimension P. The cross-attention operation is then computed as:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{P}}\right)\mathbf{V} \tag{7}$$

This attention mechanism allows the model to modulate visual features of each slice based on the textual description, effectively creating text-guided feature representations. We obtain the final aligned feature map F_I' by projecting the attention output back to the original feature space and adding it to the input features:

$$F_I' = F_I + W_O(\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \tag{8}$$

where W_O is a learnable projection matrix. This text-conditioned feature map is then passed to the decoder to generate the segmentation for each 2D slice. By integrating textual guidance directly into the visual representation through cross-attention, our model learns to associate spatial regions with their semantic descriptions, enabling precise text-guided anatomical segmentation across the entire volume.

3.4. Text-Enhanced Decoder Architecture

Following the multimodal latent space alignment, we integrate the text-modulated features into a 2D decoder architecture to generate the predicted segmentation masks for each slice. We follow a skip connection-based encoder-decoder structure. The encoder extracts hierarchical features at different resolutions, while the decoder progressively upsamples these features to restore the original spatial dimensions.

Let $\{F_1, F_2, \ldots, F_L\}$ denote the feature maps extracted at each level of the encoder, where F_L represents the deepest features that undergo text-guided cross-attention. The text-enhanced features F_L' are then propagated through the decoder pathway as:

$$D_i = \mathcal{U}_i([D_{i+1}, F_i]), \quad \text{for } i = L-1, \ldots, 1 \tag{9}$$

where $D_L = F_L'$ and $\mathcal{U}_i$ represents the decoder block at level i, which includes upsampling operations followed by convolutional layers, and $[,]$ denotes channel-wise concatenation of feature maps from the decoder path and corresponding encoder path via skip connections. This architecture allows the text-guided features to influence the entire decoding process while preserving fine spatial details from the encoder pathway. We obtain the segmentation mask for each 2D slice by applying a segmentation head to the output of the decoder:

$$\hat{y}^d = \sigma(H(D_1)) \tag{10}$$

where H is a 1×1 convolution that maps the decoder output to the desired number of classes, and σ is the sigmoid activation function that produces probability maps. The final volumetric segmentation is reconstructed by stacking these 2D predictions:

$$\hat{y} = \text{stack}(\{\hat{y}^1, \hat{y}^2, \ldots, \hat{y}^D\}) \tag{11}$$

3.5. Network Loss

We formulate the total loss of TESSERACT by combining the previously defined binary cross-entropy (BCE) and Dice loss components. For a given batch of image slices, text descriptions, and ground truth masks, the loss function is defined as:

$$\mathcal{L}(x,t,y) = \mathcal{L}_{BCE}(g(x,t),y) + \mathcal{L}_{Dice}(g(x,t),y), \tag{12}$$

where $g(x,t)$ represents the model's predicted segmentation given the image x and text description t, and $\mathcal{L}_{BCE}$ and $\mathcal{L}_{Dice}$ correspond to the losses defined in Equations (3) and (4), respectively. This combined loss function balances pixel-wise classification accuracy with volumetric overlap measurement, providing a robust training objective for our text-guided segmentation task.

Tractable solution. *To establish a tractable framework for text-guided volumetric segmentation, we formulate the problem as a binary segmentation task with a one-to-one correspondence between text prompts and target anatomical structures, where $K = 1$ and each input text-image pair (x_i, t_i) is associated with a single binary volumetric mask $y_i \in \{0,1\}^{H \times W \times D}$ indicating the presence or absence of the described anatomical entity at each voxel location.*

4. Experimental Evaluation

4.1. Dataset: 3D brain MRIs segmentations

We include 414 subjects of public T1-weighted longitudinal brain MRIs with manually delineated anatomical structures from Open Access Series of Imaging Studies (OASIS) []. Due to the difficulty of preserving the diffeomorphic property across individual subjects with large age variations, we carefully evaluate images from subjects aged from 60 to 90. All MRIs were pre-processed as $160 \times 192 \times 224$, $1.25mm^3$ isotropic voxels, and underwent skull-stripping, intensity normalization, bias field correction and pre-alignment with affine transformations.

Split	2 Brain Region Model	4 Brain Region Model
Train	573	1147
Val	82	164
Test	164	328
Total	**819**	**1639**

Table 1. Dataset summary for 2 and 4 Brain Region Model

4.2. Implementation Details

We implement TESSERACT with a 2D UNet architecture using a ResNet-34 encoder pretrained on ImageNet [17]. We freeze the encoder parameters during training to preserve the pretrained visual representations. For the text encoder, we employ BioBERT [], a domain-specific variant of BERT pretrained on biomedical corpora, which provides contextualized embeddings for anatomical descriptions. We also keep the BioBERT model frozen during training to leverage its rich linguistic representations. We implement the text-guided cross-attention module with a projection dimension of $P = 128$, which empirically provides a good balance between computational efficiency and representational capacity. To handle 3D volumetric data with our 2D architecture, we reshape the input volumes by treating the depth dimension as part of the batch dimension, effectively processing each slice independently while sharing the same text embedding across all slices of a volume.

We train the model using the Adam optimizer [] with a learning rate of 1×10^{-5} and weight decay of 1×10^{-5}. We employ early stopping with a patience of 5 epochs to prevent overfitting, monitoring the validation loss as the stopping criterion. The batch size is set to 1 for volumetric inputs, which effectively becomes D (the depth dimension) for the 2D slices after reshaping.

During inference, our model accepts a volumetric image and a textual description of the target anatomical structure. We encode the text using BioBERT, and process each slice of the volume through our 2D network, where the resulting embeddings guide the segmentation process through the cross-attention mechanism. We then reconstruct the 3D segmentation mask from the slice-wise predictions, producing a volumetric segmentation mask that corresponds to the described anatomical region.

4.3. Experiments

Initially, we employed the SAMWISE [18] model to evaluate brain segmentation guided by text prompts. In the first stage, we conducted inference using the pretrained model without any fine-tuning. However, as the model had been trained primarily on natural images, it failed to segment any brain regions—even when prompted to identify the largest anatomical structures. Subsequently, we attempted to fine-tune the model on our curated dataset comprising annotations for 35 distinct brain regions. Despite our efforts, the training proved computationally resource intensive, requiring approximately three days to complete a single epoch. To reduce the training burden, we further restricted the task to segmenting only four brain regions. Nevertheless, this configuration also resulted in prolonged training times and yielded unsatisfactory inference performance, with the model failing to segment even these limited targets accurately. These findings suggest that SAMWISE, in its current form, is not suitable for detailed brain segmentation tasks without significant architectural or training modifications.

Following the limitations observed with SAMWISE, we proceeded to evaluate our custom-designed model, TESSERACT, under multiple dataset configurations. Specifically, we trained two distinct versions of the model, each corresponding to a different segmentation task. The first configuration focused on segmenting only the two largest brain regions: Left Cerebral White Matter and Right Cerebral White Matter. In the second configuration, the model was trained to segment four regions, adding Left Thalamus and Right Thalamus to the previous two. This experimental setup was designed to systematically assess the model's performance as the complexity of the segmentation task increased. By incrementally exposing the model to a greater number of anatomical structures, we aimed to evaluate its robustness and generalization capability in handling more fine-grained and cognitively demanding segmentation scenarios. To assess the performance of the two model configurations, we report both the Dice Similarity Coefficient and Binary Cross Entropy (BCE) loss. Additionally, we provide region-specific evaluation metrics to identify which brain structures posed greater challenges for the model to segment accurately.

We experimented with two different text embedding models to encode the text prompts, which were then used to guide the segmentation model. A representative example of a text prompt is *"segment Right Cerebral White Matter."* Initially, we employed the fastText [19] model for this purpose. However, we observed that fastText produced nearly identical embedding vectors for all four distinct text prompts. As a result, the model was unable to differentiate between the brain regions specified by the prompts and instead attempted to segment a generalized region encompassing all structures. In contrast, when using the BioBERT model, we obtained more semantically

Metric	2-Region	4-Region
BCE Loss	0.0531	0.0515
Dice Score	0.9035	0.6412
Dice Loss	0.0965	0.3588
Total Loss	0.1496	0.4103

Table 2. Comparison of Evaluation Metrics for TESSERACT Model Trained on Two vs. Four Brain Regions

Model	Brain Part	Samples	Dice Score	Dice Loss	BCE Loss	Total Loss
2-Region	Left-Cerebral-White-Matter	83	0.8851	0.1149	0.0624	0.1773
	Right-Cerebral-White-Matter	75	0.9239	0.0761	0.0428	0.1189
4-Region	Left-Cerebral-White-Matter	83	0.8632	0.1368	0.0633	0.2001
	Right-Cerebral-White-Matter	75	0.8056	0.1944	0.0858	0.2803
	Left-Thalamus	84	0.8933	0.1067	0.0044	0.1111
	Right-Thalamus	81	0.0002	0.9998	0.0565	1.0563

Table 3. Comparison on Separate Brain Parts Test Metrics for TESSERACT Model Trained on Two vs. Four Brain Regions

distinct embeddings for each prompt. This helped the model better match each text prompt to the correct brain region, making the guided segmentation more accurate.

4.4. Results

To evaluate the segmentation performance of our proposed model TESSERACT, we trained and tested it under two dataset configurations: one involving only the two largest brain regions (Left and Right Cerebral White Matter), and the other extending to four regions by including the Left and Right Thalamus. Table 4.4 presents the aggregate evaluation metrics—Dice Score, Dice Loss, Binary Cross Entropy (BCE) Loss, and Total Loss—across both configurations.

TESSERACT achieved a significantly higher Dice Score of 0.9035 and a lower Total Loss of 0.1496 when trained on the two-region setup. In contrast, the four-region model experienced a substantial performance drop, with the Dice Score decreasing to 0.6412 and the Total Loss increasing to 0.4103. This highlights a clear trade-off between task complexity and segmentation accuracy.

To gain deeper insight, Table 2 reports region-wise performance. In the two-region model, the Right Cerebral White Matter achieved the highest Dice Score (0.9239) and the lowest Total Loss (0.1189), indicating excellent segmentation fidelity. The Left Cerebral White Matter also maintained strong performance, with a Dice Score of 0.8851. In the four-region setting, however, segmentation performance varied significantly by region. The Left Thalamus achieved a Dice Score of 0.8933, closely matching the performance in the simpler configuration. This suggests that some anatomical regions, despite being newly introduced, were well-learned by the model. On the other hand, the Right Thalamus showed an almost complete segmentation failure with a Dice Score of only 0.0002, indicating that the model struggled severely with this region, likely due to either its anatomical ambiguity or fewer representative samples.

Notably, the Right Cerebral White Matter Dice Score decreased from 0.9239 to 0.8056, and the BCE Loss nearly doubled, suggesting that model performance for previously well-learned regions can degrade when additional segmentation tasks are introduced. These results underscore the importance of embedding quality. Our initial experiments using fastText embeddings led to indistinguishable prompt representations, which impaired region-specific segmentation. Replacing fastText with BioBERT enabled the model to better align text prompts with specific anatomical targets, significantly improving guided segmentation outcomes.

As shown in Fig. 2, the segmentation output for the first 2-region sample demonstrates clear boundaries. In Fig. 3, we observe a similar trend. For the 4-region setup, Fig. 4 through Fig. 7 display varied performance, with Fig. 7 indicating segmentation failure in one of the regions.

5. Future Work

While **TESSERACT** shows strong potential for text-guided volumetric segmentation, several areas remain open for future exploration:

- **Enhanced Text-Visual Alignment:** Although BioBERT embeddings improved semantic understanding,

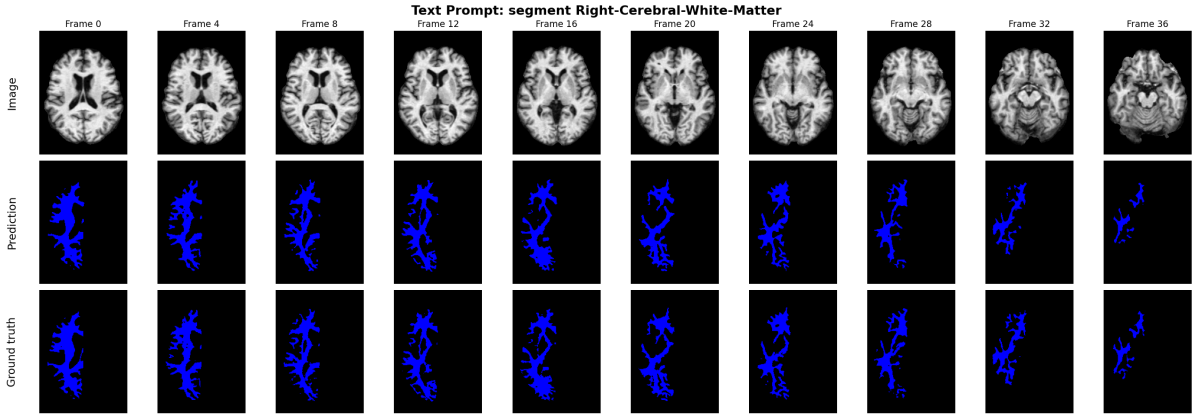

Fig. 2. Segmentation Output - TESSERACT trained on 2 Brain Regions Example 1

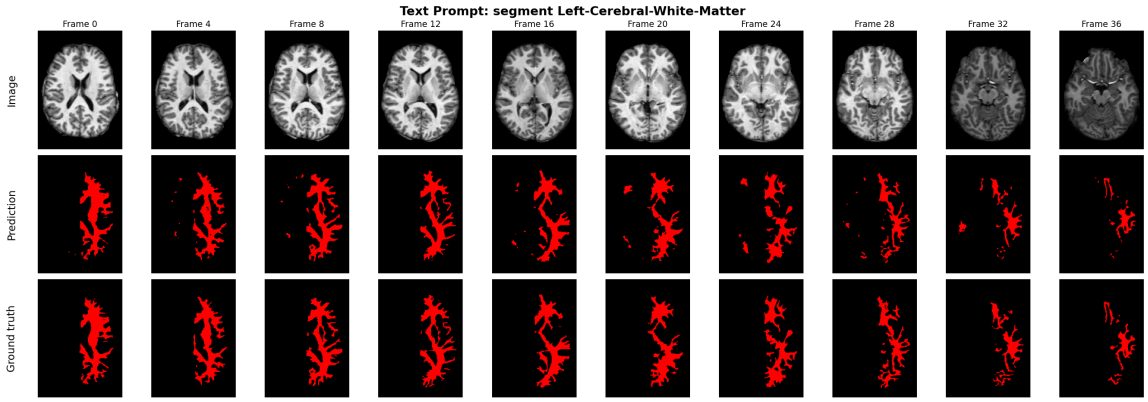

Fig. 3. Segmentation Output - TESSERACT trained on 2 Brain Regions Example 2

further research is needed to more tightly align the textual embeddings with the visual feature space. This could involve training a joint embedding space or incorporating contrastive learning techniques to improve multimodal fusion.

- **Scaling to Finer Anatomical Structures:** The current setup primarily focuses on larger and more distinguishable brain regions. Future iterations of **TESSERACT** should address segmentation of finer, more complex anatomical structures, which may require improved prompt granularity and targeted feature sensitivity.
- **Multi-Region Prompt Segmentation:** Presently, the model handles one region per prompt. Extending the system to support simultaneous multi-region segmentation from a compound or structured prompt could significantly increase clinical utility. This may involve developing a prompt parser or attention mechanisms capable of associating multiple text spans with distinct spatial regions.

By addressing these areas, future versions of **TESSERACT** can become more generalizable, efficient, and clinically applicable across a wider range of medical imaging tasks.

6. Conclusion

In this work, we proposed TESSERACT, an efficient and easy-to-train text-guided brain segmentation model that leverages natural language text prompts and medical image representations. By incorporating textual guidance into a slice-wise 2D segmentation framework, TESSERACT achieves lower inference time compared to more computationally intensive models such as TextDiff [14], Diffboost [12]. Our method demonstrates the feasibility of prompt-driven anatomical segmentation, opening the door to flexible and semantically interpretable clinical applications. Despite its promising performance, several avenues remain for future research. Further improvements are needed to better align textual embeddings with visual features in the shared latent space. Additionally, expanding the anatomical coverage by including a broader set of structures and enabling simultaneous segmentation

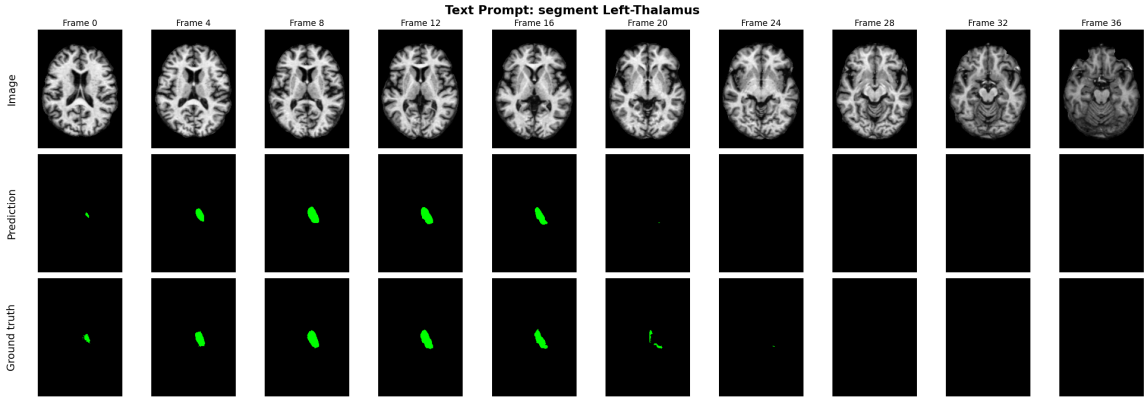

Fig. 4. Segmentation Output - TESSERACT trained on 4 Brain Regions Example 1

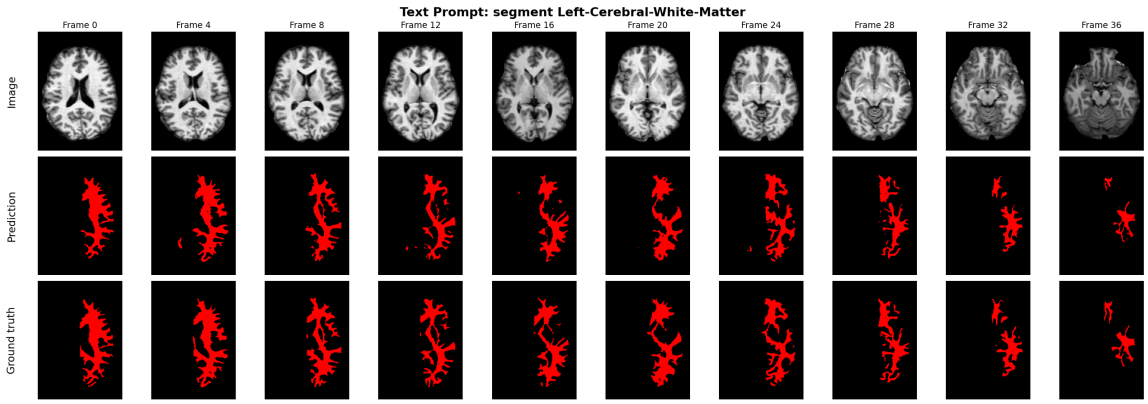

Fig. 5. Segmentation Output - TESSERACT trained on 4 Brain Regions Example 2

of multiple regions through complex prompts represent important next steps. Overall, TESSERACT provides a strong foundation for advancing text guided segmentation in volumetric medical imaging.

References

1. Hui Lin, Haonan Xiao, Lei Dong, Kevin Boon-Keng Teo, Wei Zou, Jing Cai, and Taoran Li. Deep learning for automatic target volume segmentation in radiation therapy: a review. *Quantitative Imaging in Medicine and Surgery*, 11(12):4847, 2021.
2. Imad Eddine Toubal. *Volumetric medical image segmentation with deep learning pipelines*. PhD thesis, University of Missouri–Columbia, 2020.
3. Carolin M Sauer, Katrin Heider, Jelena Belic, Samantha E Boyle, James A Hall, Dominique-Laurent Couturier, Angela An, Aadhitthya Vijayaraghavan, Marika AV Reinius, Karen Hosking, et al. Longitudinal monitoring of disease burden and response using ctdna from dried blood spots in xenograft models. *EMBO Molecular Medicine*, 14(8):e15729, 2022.
4. Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.
5. Gustav Müller-Franzes, Firas Khader, Robert Siepmann, Tianyu Han, Jakob Nikolas Kather, Sven Nebelung, and Daniel Truhn. Medical slice transformer: Improved diagnosis and explainability on 3d medical images with dinov2. *arXiv preprint arXiv:2411.15802*, 2024.
6. Shuming Zhang, Hao Wang, Suqing Tian, Xuyang Zhang, Jiaqi Li, Runhong Lei, Mingze Gao, Chunlei Liu, Li Yang, Xinfang Bi, et al. A slice classification model-facilitated 3d encoder–decoder network for segmenting organs at risk in head and neck cancer. *Journal of radiation research*, 62(1):94–103, 2021.
7. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
8. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

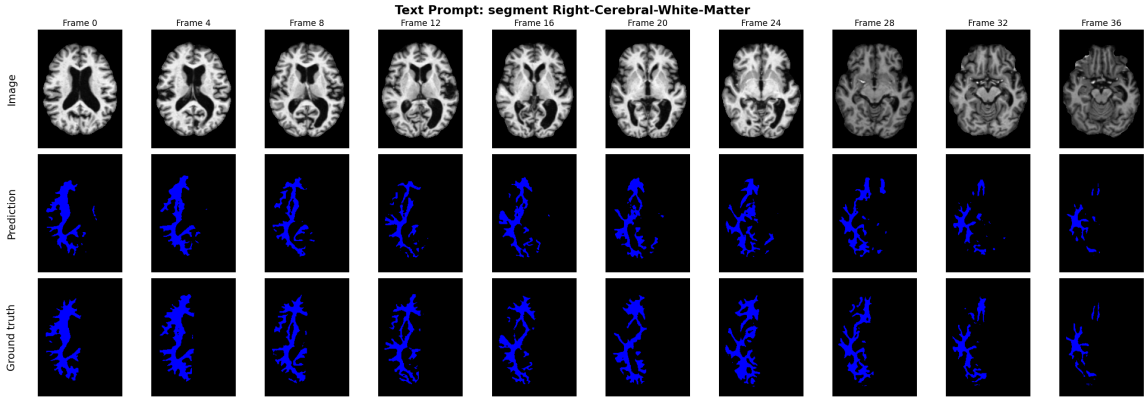

Fig. 6. Segmentation Output - TESSERACT trained on 4 Brain Regions Example 3

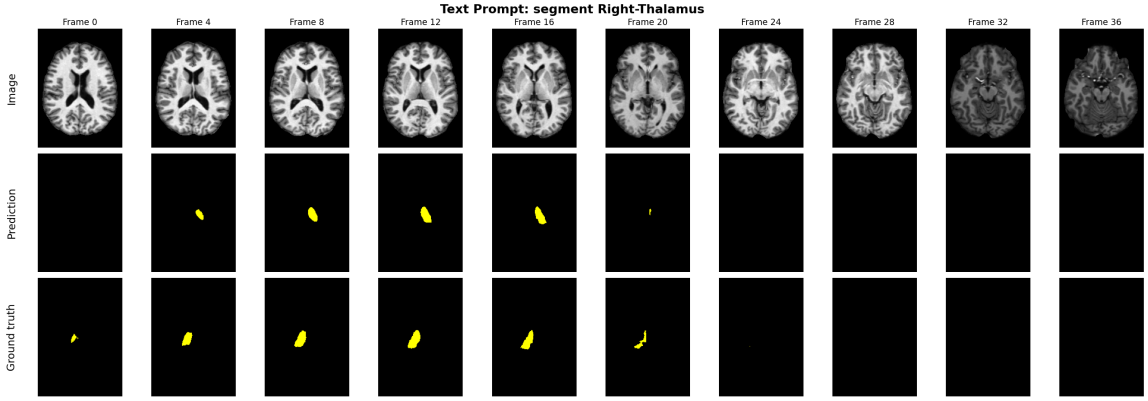

Fig. 7. Segmentation Output - TESSERACT trained on 4 Brain Regions Example 4

9. Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. *arXiv preprint arXiv:2411.17646*, 2024.
10. Thiago F Rangel, Jose Alexandre F Diniz-Filho, and Luis Mauricio Bini. Sam: a comprehensive application for spatial analysis in macroecology. *Ecography*, 33(1):46–50, 2010.
11. Young Woon Kim, Hyunjun Cho, Sung-Jea Ko, and Seung-Won Jung. Text knowledge-guided segment anything model for medical image segmentation. In *2024 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, pages 1–4. IEEE, 2024.
12. Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Diffboost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*, 2024.
13. Zhiwei Dong, Genji Yuan, Zhen Hua, and Jinjiang Li. Diffusion model-based text-guided enhancement network for medical image segmentation. *Expert Systems with Applications*, 249:123549, 2024.
14. Chun-Mei Feng. Enhancing label-efficient medical image segmentation with text-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–262. Springer, 2024.
15. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
16. Yihao Zhao, Enhao Zhong, Cuiyun Yuan, Yang Li, Man Zhao, Chunxia Li, Jun Hu, and Chenbin Liu. Tg-lmm: Enhancing medical image segmentation accuracy through text-guided large multi-modal model. *arXiv preprint arXiv:2409.03412*, 2024.
17. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
18. Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation, 2025.
19. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2017.