Exploring Med-VLM for Abdominal CT Organ Segmentation on the BTCV Dataset

Ramin Mohammadi
UT Austin
ramin.mohammadi@utexas.edu

Abstract

Accurate segmentation of abdominal organs in computed tomography (CT) imaging is essential for diagnosis, surgical planning, and downstream clinical tasks. Recent vision–language models (VLMs) have demonstrated strong capabilities in grounding textual descriptions to visual features, suggesting their potential for medical image segmentation where anatomical context is often described linguistically. This work explores the feasibility of applying Med-VLM, a vision–language segmentation framework, to the BTCV abdominal CT dataset for spleen and liver segmentation. I implement both large and small variants of Med-VLM and compare them to a 2D U-Net baseline under controlled preprocessing and training conditions. Med-VLM performs text-guided, single-organ segmentation using binary masks and organ-specific prompts, whereas the U-Net performs multi-organ segmentation using non-binary masks. All models exhibit substantial overfitting given the limited slice-level supervision, yet Med-VLM consistently achieves higher validation Dice scores than the U-Net, suggesting possible benefits from pretrained encoders, more specifically vision transformer image encoders and BERT like text encoders, and text conditioning. My findings highlight both the promise and current limitations of VLM-based medical segmentation, particularly under resource-constrained training regimes.

1 Introduction

Computed tomography (CT) is widely used in clinical workflows for diagnosing abdominal diseases, monitoring treatment response, and guiding interventional procedures. A key step in many of these applications is the reliable segmentation of organs of interest, such as the liver and spleen. Manual delineation, however, is time-consuming and subject to inter-observer variability, motivating the need for robust automated segmentation systems.

Deep learning has enabled substantial progress in medical image segmentation, with architectures such as U-Net and its variants achieving state-of-the-art performance across many benchmark datasets. More recently, vision–language models (VLMs) and multimodal transformers have emerged as powerful tools capable of integrating textual descriptions with visual information. In medical imaging, this opens the possibility of using structured anatomical text prompts to guide segmentation, enabling models to leverage high-level contextual knowledge that purely image-based networks cannot access.

Med-VLM (Zhao et al., 2025) is a recently proposed vision–language segmentation framework that conditions organ masks on natural-language descriptions of anatomical location. By fusing image and text embeddings through cross-attention, the model aims to improve organ localization and boundary precision. While Med-VLM demonstrates strong results on large multi-organ CT datasets, its generalization characteristics and performance under limited training resources are not yet well understood.

This work investigates the application of Med-VLM to the BTCV abdominal CT dataset, focusing on spleen and liver segmentation. I implement both the large and small variants of Med-VLM based on the methodology described in the original paper and compare them against a 2D U-Net baseline. The analysis highlights how model design choices—including text prompting and the use of pretrained encoders—affect performance in a controlled experimental setting.

The goal of this study is not to reproduce state-of-the-art results, but to evaluate the practical feasibility, strengths, and limitations of VLM-driven segmentation pipelines for abdominal CT imaging.

2 Related Work

2.1 Med-VLM

Med-VLM is a vision-language segmentation model that aims at increasing accuracy on medical CT organ segmentation by incorporating a clinical expert text description of the organ's spatial context within the human anatomy (Zhao et al., 2025). Their study focuses on the impact of segmentation performance by mixing this textual description with the image embedding. The model architecture consists of 4 main parts: an image encoder, text decoder, feature mixer, and mask decoder.

The image encoder is made up of a pretrained encoder (ViT vision transformer) and LoRA (Low-Rank Adaptation). The image embedding is the sum of the image processed through both the ViT and LoRA.

Similarly, the text encoder consists of a pretrained text encoder (BioBERT) and LoRA. The text embedding is the sum of the tokenized text going through each BioBERT and LoRA respectively.

For both encoders, the pretrained model's weights remain frozen, but all other parts of the model including the LoRAs are learnable.

The feature mixer creates "fused" image and text representations by performing cross attention between the image and text embeddings, along with self attention and MLP computations.

The mask decoder generates pixel-level segmentation masks from the fused image and text embeddings. It comprises of dilated convolutions, a bi-directional transformer, and an MLP.

The loss function was the unweighted sum of the BCE (binary cross-entropy) and Dice loss as done in various medical image segmentation tasks. The FLARE, SegTHOR, and MSD datasets were used which contain medical CT images and segmentation masks of organs. The training dataset consisted of approximately 47,000 images and 100,000 high-quality masks trained for 150 epochs, taking 30 hours on 8 A100 40GB GPUs.

Med-VLM was shown to compete with SAM and nnUNET, state of the art segmentation models.

2.2 SAM

SAM (Segment Anything Model), developed by Meta AI Research and FAIR, uses a prompt to determine the segmentation on a given image (Kirillov et al., 2023). This prompt can be a set of foreground / background points, a rough box or mask, or free-form text. This model can perform zero shot transfer learning with prompt engineering, allowing it to perform on par or even superior to supervised models on various segmentation tasks. The architecture consists of a prompt encoder, image encoder, and a mask decoder to output a pixel-wise segmentation.

The image encoder is made of a pre-trained vision transformer (ViT), and the free form text goes through a CLIP text encoder.

The mask decoder maps the image and prompt embeddings to an output mask, using a modified transformer decoder block that uses prompt self-attention and cross-attention in two directions (prompt-to-image embedding and vice-versa) to update all embeddings, convolutional up-sampling to acquire the original resolution of the image, and an MLP to compute the mask foreground probability at each image location. It's worth noting, similar to BERT, a CLS [class] token (embedding) is concatenated to the prompt embedding to be used at the decoder's output.

The loss function was a linear combination of the focal loss and dice loss. And most importantly, to achieve strong generalization to new data distributions, the SAM developers created the SA-1B segmentation dataset made up of 1 billion masks and 11 million images to expose the model to a large and diverse set of masks.

2.3 nnUNET

nnUNET is an automatic training pipeline for medical imaging segmentation (Isensee et al., 2018). In the Medical Segmentation Decathlon challenge, which doesn't allow any changes to the model or pipeline between datasets, the model had the highest mean dice scores across all of the segmentation classes. The pipeline has 3 possible architectures: 2D UNET, 3D UNET, and UNET Cascade.

2D UNET is reported to be suboptimal for 3D medical image segmentation because 2D slices are processed independently, missing information across the other slices of a 3D image when being processed.

A 3D UNET is ideal for 3D images but the full volume (height x width x depth) can take up too much memory RAM for a GPU even for a batch size of 1, so the 3D images are processed instead by 3D patches which may hinder the segmentation accuracy. To resolve this shortcoming, the authors tried UNET Cascade, utilizing 2 3D UNETs in order to process the full resolution of the 3D images rather in patches. In their results, they mention that

due to the exceeding performance of 3D UNET, UNET Cascade was not necessary (for the datasets they used) since the patch size from 3D UNET contained enough of the 3D image to perform an accurate segmentation. If the median volume of a dataset (h x w x depth) was smaller than the base patch size of 128x128x128, the median volume was used as the patch size.

Preprocessing performed was cropping to the region of nonzero values to reduce image sizes (computation overhead), resampling to address voxel spacings, and normalization of the intensity scales of the CT scans.

To avoid overfitting from the use of large neural networks on limited data the following data augmentations were used: random rotations, random scaling, random elastic deformations, gamma correction augmentation and mirroring. But, it is important to note that the data augmentations were defined differently for the 2D and 3D UNET models due to 2D UNET not processing the depth (slice) dimension.

The loss function was a linear combination of dice and cross entropy loss using an Adam optimizer.

Regarding 2D vs. 3D model performance on 3D medical images, they show that the mean dice scores of 2D UNET are slightly on par with 3D UNET, indicating that 2D UNET can perform well as long as the 3D data is processed correctly.

2.4 UNETR

UNETR, developed by NVIDIA and Vanderbilt University, was designed for 3D medical image segmentation which mimics the UNET architecture but uses a transformer for down-sampling (Hatamizadeh et al., 2021).

Each 3D image is split into uniform non-overlapping 3D patches where each patch is flattened and goes through a linear layer to acquire embeddings for each patch. Learnable position embeddings are added to the linear patch projections, but in this implementation they choose not to concatenate a learnable [class] token (in contrast to SAM).

The encoder is a transformer to capture global long range dependencies that the decoder can utilize effectively which a convolution encoder cannot capture as it only extracts local information. The encoder are stacks of transformer blocks/layers that each are made of multi-head self attention (n parallel self-attention heads) and an MLP (two linear layers with GELU activations and intermediate layer normalization). There are 12 transformer layers.

The bottleneck is a 3D devconvolutional layer. The decoder then mimics UNET's "U shape" behavior by concatenating the output of different transformer blocks to a corresponding decoder layer which are 3D 3x3x3 convolutional and 2x2x2 deconvolutional layers with batch norm and ReLU activations. The final layer is a 3D 1x1x1 convolution with a softmax activation to generate voxel-wise semantic prediction masks. The loss function is a combination of soft dice loss and cross-entropy loss. No pretrained weights for the transformer were used.

More importantly, the performance was tested on the BTCV (CT) challenge (Landman et al., 2015) and MSD (MRI/CT) datasets and showed state of the art dice scores on both datasets for each segmented organ. They used data augmentation strategies such as random rotation of 90, 180 and 270 degrees, random flip in axial, sagittal and coronal views and random scale and shift intensity.

3 Method

3.1 BTCV Dataset

Experiments were conducted using the abdominal subset from the Beyond the Cranial Vault (BTCV) CT segmentation dataset. The abdominal subset contains 50 CT scans with manual annotations for 13 organs, including the spleen, right and left kidneys, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic veins, pancreas, and adrenal glands (these are labels 1-13 with 0 indicating background for the segmentation mask).

The dataset contains variable volume sizes (512 x 512 x 85 - 512 x 512 x 198) and field of views (approx. 280 x 280 x 280 mm3 - 500 x 500 x 650 mm3). The in-plane resolution varies from 0.54 x 0.54 mm2 to 0.98 x 0.98 mm2, while the slice thickness ranges from 2.5 mm to 5.0 mm.

There are 30 training that were split into 24 training and 6 validation samples since the training set had labeled segmentation masks but the 20 test samples had no ground truth masks so the test samples were excluded.

Segmentation was focused on only two organs (spleen and liver).

3.2 Baseline

The dataset used in this experiment was the BTCV dataset (Landman et al., 2015) and the focus was on the spleen and liver organs. The UNETR model achieves a 0.97 dice score on the spleen and liver organs respectively. But, due to hardware limitations, I could not acquire this result myself, so this score is cited from UNETR's findings and is kept into consideration(Hatamizadeh et al., 2021). As a result, the alternative was to use a 2D UNET model to asses performance of not only the UNET architecture compared to Med-VLM but using only an image as input versus image and text as done in Med-VLM.

3.3 Models

Three models were tested: Med-VLM Large, Med-VLM Small, and 2D UNET. There was no existing repository for Med-VLM, so the Med-VLM model implementations were followed as closely to the methodology as described in their paper with some grey details that I had to interpret myself. Larger and smaller pretrained encoders differ the Med-VLM models and the 2D UNET served as an interpretation of an image only segmentation model performance compared to Med-VLM using both text and an image.

All three models work on 2D slices of the 3D CT images and had the same preprocessing. From HuggingFace, MedVLM Large uses the pretrained vision transformer (image encoder) "vit_base_patch16_224", and a BioBert text encoder "dmis-lab/biobert-base-cased-v1.1", and MedVLM Small used smaller pretrained encoders: "vit_tiny_patch16_224" image encoder and "sentence-transformers/all-MiniLM-L6-v2" text encoder.

For Med-VLM, each organ has a respective fixed text prompt, that was manually created, describing the general spatial context relative to the human anatomy. For example, liver was: "Liver: It is located in the right upper quadrant of the abdominal cavity, resting just below the diaphragm. The liver lies to the right of the stomach and overlies the gall bladder."

There were 2 fundamentally different setups between Med-VLM and the 2D UNET.

Med-VLM used BCE plus dice as the loss, rather CE, and involves a text prompt for a specific organ describing spatial location, which implies that *single* organ segmentations are predicted per slice (not multi organ segmentation). As a result, the same CT slices are re-used for each organ (spleen and liver) but with a binary mask, 1 indicating that the pixel is for that organ and 0 otherwise, along with the respective text prompt. Thus, at inference, the text prompt would dictate which organ to segment.

On the other hand, 2D UNET only takes the 2D CT image slice as input so having duplicate CT slices with different ground truth masks (for each organ) would confuse the model. So, it was setup as a multi-organ segmentation task where each prediction was a mask for all of the organs (both spleen and liver) on each 2D CT slice. The segmentation mask for 2D UNET were relabeled as: 1 spleen, 2 liver, and all other mask values are relabeled as 0 background, thus "Non-binary" mask as indicated in 4. Lastly, CE + dice was the loss (rather BCE).

3.4 Loss Function, Metrics, & Optimizer

Med-VLM model's loss function was the unweighted linear combination of BCE (Binary Cross Entropy) and dice, and the 2D UNET used CE (Cross Entropy) plus dice. The main performance metric used was dice and trained using an AdamW optimizer with a learning rate of 1e-4. Training was done on a collab high-RAM T4 GPU with the Med-VLM models trained for about 10 epochs and 2D UNET for 50 epochs.

3.5 Data Augmentation

The training pipeline loads each CT volume, standardizes orientation, resamples voxel spacing, applies CT window normalization, removes empty background, and performs 2D augmentations including random left–right and anterior–posterior flips, random 90° rotations, intensity shifting, scaling, and Gaussian noise, before resizing slices to 224×224. Then the train and validation sets are made using the 2D slices.

4 Results

All three models exhibit substantial overfitting, reflected in large gaps between training and validation dice scores 3. Validation performance plateaus early, suggesting that limited slice-level supervision (approximately 5k slices) is insufficient for robust generalization, or that errors in the data pipeline may have reduced training diversity 1. The best achievable mean dice score on the spleen and liver segmentations was 0.58 using Med-VLM

Organ	Text Prompt
Spleen	The spleen's long axis corresponds most closely to the eleventh or tenth rib and passes anterior to the mid-axillary line.
Right Kidney	The right kidney lies in the right retroperitoneum, inferior to the liver and lateral to the psoas muscle.
Left Kidney	The left kidney is positioned in the left retroperitoneum, inferior to the spleen and lateral to the psoas muscle.
Gallbladder	The gallbladder is located in the right upper quadrant, beneath the liver at the gallbladder fossa.
Esophagus	The esophagus courses vertically in the posterior mediastinum, posterior to the trachea and anterior to the spine.
Liver	Located in the right upper quadrant just below the diaphragm. The liver lies to the right of the stomach and overlies the gallbladder.
Stomach	The stomach is located in the left upper quadrant, posterior to the liver and anterior to the pancreas.
Aorta	The aorta descends along the midline in the retroperitoneum, anterior to the spine and left of the inferior vena cava.
Inferior Vena Cava	The inferior vena cava ascends along the right side of the aorta in the retroperitoneum.
Portal/Splenic Veins	The portal and splenic veins form posterior to the pancreas and course toward the liver hilum within the hepatoduodenal ligament.
Pancreas	The pancreas lies transversely in the upper abdomen, posterior to the stomach and anterior to the splenic vein.
Right Adrenal Gland	Superior and medial to the right kidney, adjacent to the upper pole and near the inferior vena cava.
Left Adrenal Gland	Superior and medial to the left kidney, adjacent to the upper pole and near the aorta.

Table 1: Organ-specific anatomical text prompts used for Med-VLM text-conditioned segmentation.

Model	Trainable	Frozen	Total Params
Med-VLM Large	53M	194M	247M
Med-VLM Small	2.9M	28M	31M
2D U-Net	7.7M	0	7.7M

Table 2: Parameter counts for different model variants used in experiments. MedVLM models include frozen pretrained encoders; 2D U-Net uses no pretrained components.

Model	Best Train Dice	Best Val Dice
MedVLM Large	0.56	0.37
MedVLM Small	0.58	0.34
2D U-Net	0.38	0.24

Table 3: Best mean Dice scores achieved during training and validation for spleen and liver segmentation.

Large on about 5k training slices versus UNETR which achieved a dice score of 0.97 on both organs individually(Hatamizadeh et al., 2021).

Despite these issues, Med-VLM Large and Med-VLM Small achieve higher validation Dice scores than the 2D U-Net, likely due to their pretrained encoders and possibly from the text guidance. Visual inspection shows that all models learn coarse organ locations but frequently undersegment or oversmooth boundaries 5. Lastly, Med-VLM predictions generally correspond to the organ named in the prompt, supporting the notion that text conditioning influences spatial reasoning, although the extent to which descriptive prompts improve precision, as suggested in the Med-VLM paper (Zhao et al., 2025), remains inconclusive.

5 Conclusion

This exploratory study evaluated Med-VLM and 2D U-Net models for spleen and liver segmentation on the BTCV abdominal CT dataset. Despite careful preprocessing and augmentation, all models exhibited substantial overfitting, with validation Dice scores plateauing far below training performance. This indicates that slice-based supervision, limited data volume, and potential information loss from treating 3D CT volumes as independent 2D slices remain major bottlenecks. My implementation worked on 2D slices of 3D volumes, thus mask predictions are missing information across slices of the full 3D volume, possibly hindering segmentation performance.

Nonetheless, both Med-VLM variants outperformed the 2D U-Net on validation Dice, suggesting that pretrained encoders, specifically vision transformer image encoders and BERT like text encoders, and text-conditioned representations provide meaningful advantages even under constrained training settings. Visual inspection further showed that Med-VLM generally predicts masks consistent with the prompted organ, supporting claims from (Zhao et al., 2025) that language cues can guide

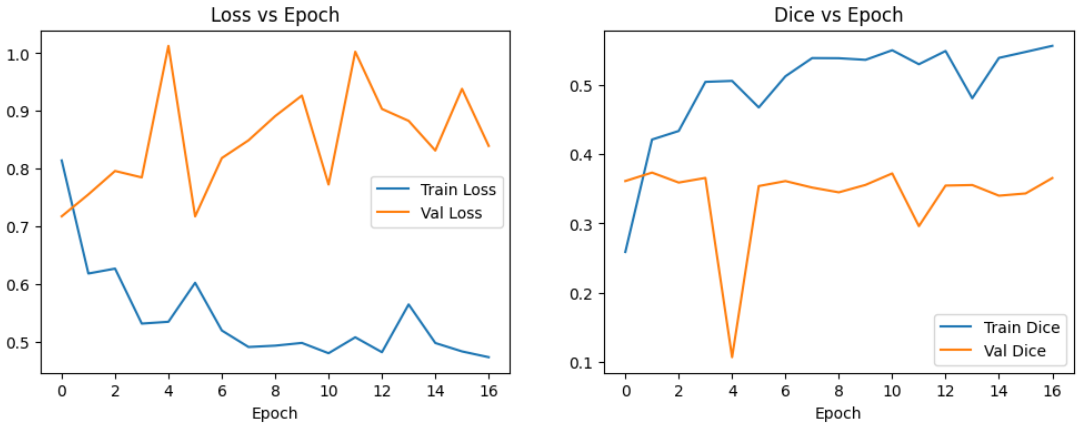

Figure 1: MedVLM Large training and validation losses and dice scores across epochs.

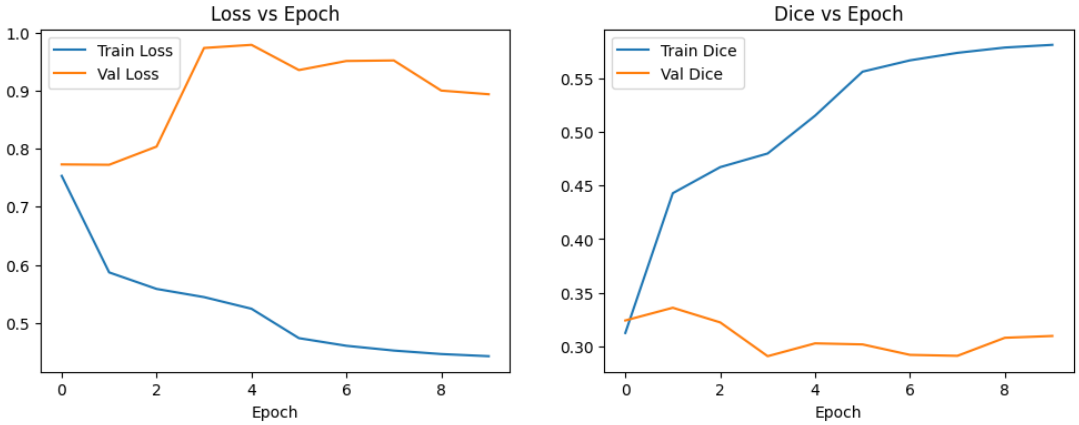

Figure 2: MedVLM Small training and validation losses and dice scores across epochs.

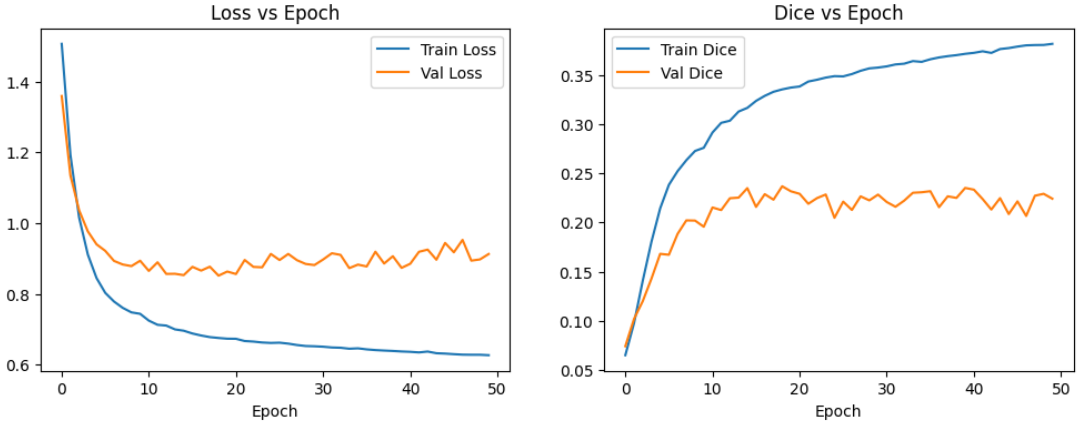

Figure 3: 2D UNET training and validation losses and dice scores across epochs.

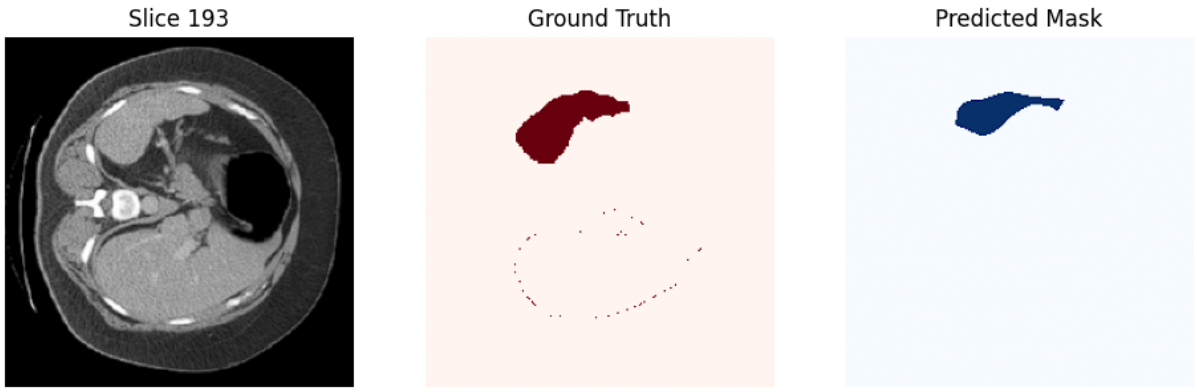

Figure 4: MedVLM Large spleen mask on a slice of a CT image from the validation set.

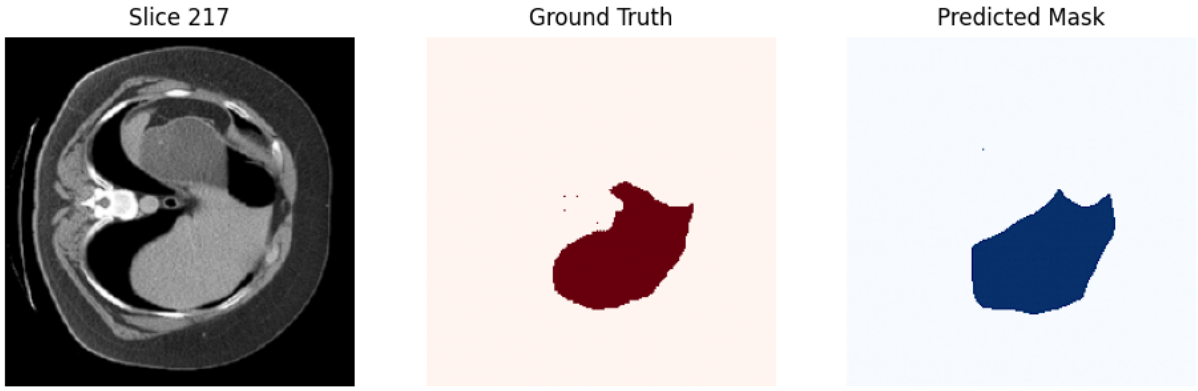

Figure 5: MedVLM large liver mask on a slice of a CT image from the validation set.

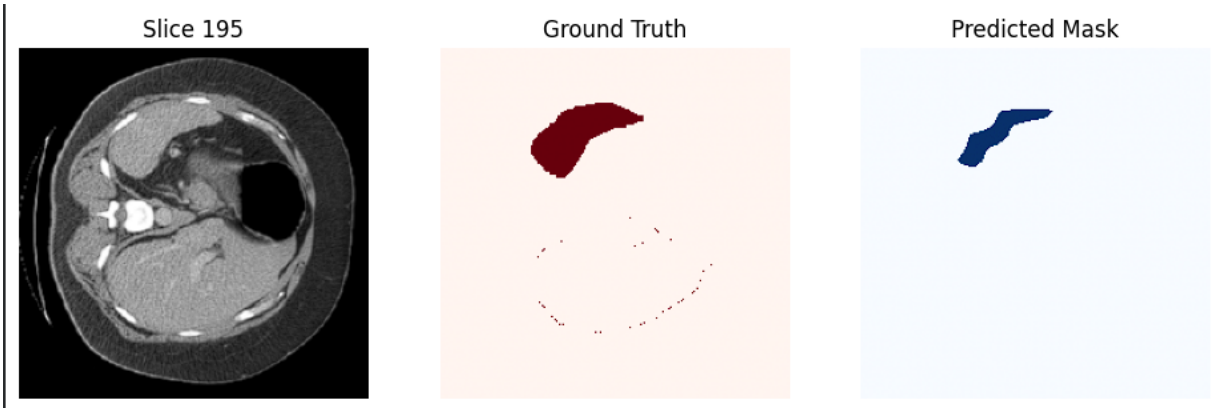

Figure 6: MedVLM small spleen mask on a slice of a CT image from the validation set.

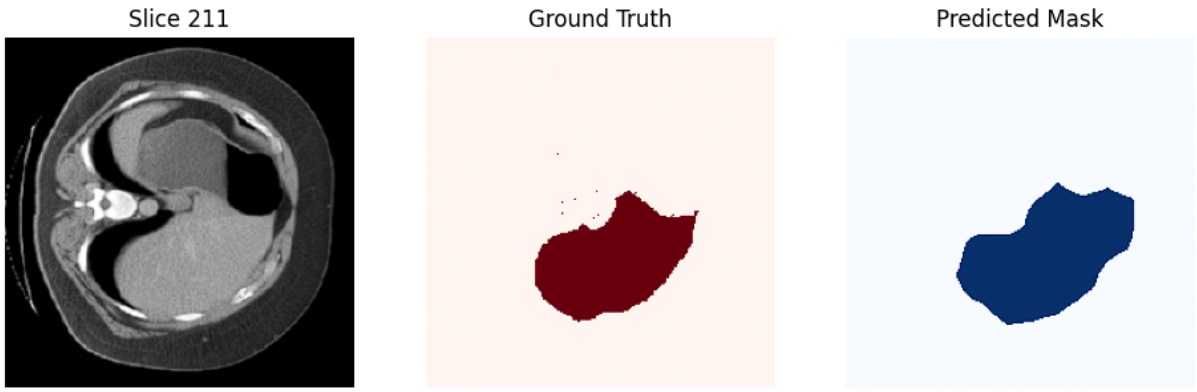

Figure 7: MedVLM small liver mask on a slice of a CT image from the validation set.

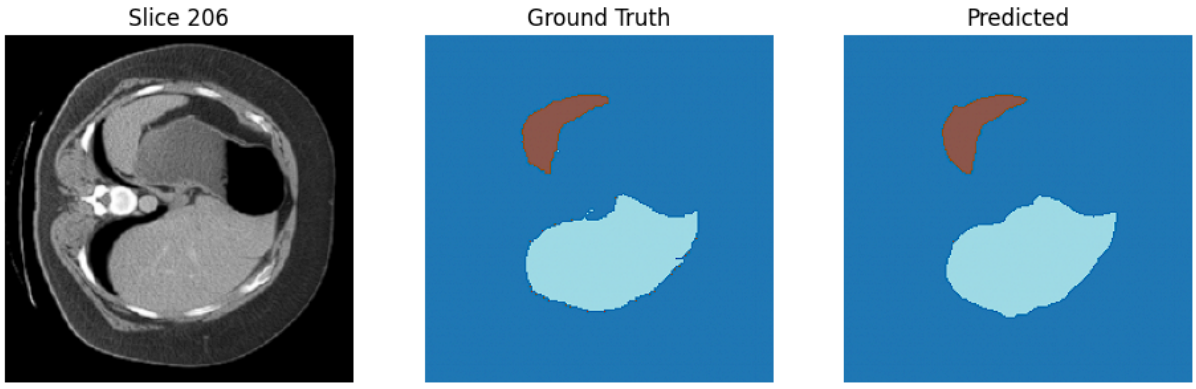

Figure 8: 2D UNET Mask on a slice of a CT image from the validation set.

Model	Input	Pretrained	Task	Loss	Mask Type	Duplicate Slices
MedVLM Large/Small	Text + CT slice	Yes	Single-organ	BCE + Dice	Binary	Yes
2D U-Net	CT slice only	No	Multi-organ	CE + Dice	Non-binary	No

Table 4: Comparison of training setups for MedVLM models and the 2D U-Net baseline. MedVLM performs text-guided single-organ segmentation using binary masks and duplicated slices paired with different prompts. The 2D U-Net predicts multiple organs jointly from a single image using a non-binary mask, requiring CE loss.

spatial reasoning. However, the degree to which detailed anatomical descriptions improve accuracy, as proposed in the original Med-VLM paper, could not be conclusively demonstrated.

Several limitations impacted performance, including the absence of full 3D context, limited training data, hardware constraints, and ambiguities in re-implementing Med-VLM without an official repository as a guide. Future work would involve exploring 3D VLM architectures, proper data augmentations for 3D medical data, richer prompting strategies such as SAM, and leverage larger and more diverse datasets.

References

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. 2021. Unetr: Transformers for 3d medical image segmentation. *Preprint*, arXiv:2103.10504.

Fabian Isensee, Jens Petersen, André Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Köhler, Tobias Norajitra, Sebastian J. Wirkert, and Klaus H. Maier-Hein. 2018. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *CoRR*, abs/1809.10486.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *Preprint*, arXiv:2304.02643.

Bennett A. Landman, Zhoubing Xu, Juan E. Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. 2015. Miccai multi-atlas labeling beyond the cranial vault – workshop and challenge. In *Proceedings of MICCAI Multi-Atlas Labeling Beyond the Cranial Vault Challenge*. Springer.

Yihao Zhao, Enhao Zhong, Cuiyun Yuan, Yang Li, Man Zhao, Chunxia Li, Jun Hu, Wei Liu, and Chenbin Liu. 2025. Med-vlm: Enhancing medical image segmentation accuracy through vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 7283–7293.