Organ-Seg: A Vision-Language and LLM-Enhanced Framework for User-Guided Abdominal Organ Image Segmentation

Md. Akil Raihan Iftee[1], Tasfia Faija[1], Partho Choudhury Shoumya[1] and Sunanda Das[1, 2]
[1] *Department of CSE, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh.*
[2] *Department of EECS, University of Arkansas, Fayetteville, AR 72701, USA.*
{iftee1807002, faija1807042, shoumya1807021}@stud.kuet.ac.bd, and sunandad@uark.edu

***Abstract*—Medical image segmentation plays a vital role in diagnostic and treatment planning, where precision is crucial for accurate outcomes. Traditional segmentation methods, while effective in many areas, often fail to incorporate user-driven guidance, leading to errors in region identification, especially when irrelevant regions are segmented. In this study, we present a new, instruction-based medical image segmentation framework that enhances user interaction while delivering precise and context-aware results. Our approach addresses the limitations of previous works, such as vision-large language models (LLM) like LLaVA, which provide context but do not perform segmentation, and the Segment Anything Model, which performs segmentation but does not incorporate user's text-guided instruction. We propose a segmentation model framework that combines vision-language embeddings from LLava with SAM to perform accurate, query-based segmentation of medical images. A key innovation of our model framework is its ability to handle false premises—situations where a user queries for an organ not present in the image—by employing a similarity-based mechanism that prevents incorrect segmentation. Tested on MRI datasets, FLARE22, our system achieves the highest segmentation dice coefficient 63.9%, with significantly improved relevance and reliability. The results demonstrate the effectiveness of our approach in refining segmentation quality and enhancing user-guided interaction, thus offering an advanced tool for medical imaging applications.**

***Index Terms*—Vision Language Model, LLava, MED, User-Guided Interaction**

I. Introduction

In the medical domain, image segmentation is critical for identifying specific regions in diagnostic scans, such as organs or tumors. The field of computer vision has advanced significantly, particularly in image segmentation, where the goal is to accurately separate objects from their background. For example, a segmentation model tasked with identifying a tumor in an MRI scan would isolate the tumor from surrounding tissues. This technique is crucial for precise image interpretation in applications like medical diagnostics and autonomous systems. In many real-world applications, the ability to guide and refine the segmentation process is crucial. However, traditional segmentation models have limitations, especially regarding user interaction. They often lack the flexibility to incorporate user text guidance as input, leading to less accurate or irrelevant results.

To address these shortcomings, integrating large language models into the segmentation process opens up new possibilities for user-driven segmentation. By allowing users to provide specific instructions on what they want segmented [1], models can deliver more tailored and accurate outcomes. Language-assisted models like LLaVA (Language-Assisted Vision Architecture) [2] are designed to generate context based on both the image and the user's query. However, LLaVA on its own does not perform the segmentation—it only provides the context.

This is where the Segment Anything Model (SAM) [3] comes into play. SAM excels at extracting features from different regions of an image, allowing for precise segmentation of specific objects or areas based on user prompts. Its flexibility makes it particularly useful for applications that require interactive and adaptable segmentation, such as medical imaging or complex multi-object scenes.

Making sure that the segments match the user's query closely is important for ensuring that the model's output aligns with what the user wants. However, challenges remain, particularly in avoiding false positives. When asked to segment an object that isn't present in the image, models sometimes improperly segment irrelevant regions. Ensuring that the system can recognize when an object is absent is crucial for improving the accuracy and reliability of segmentation tasks.

In this work, we propose an instruction-based segmentation framework that addresses these challenges. Our approach combines the strengths of LLaVA-MED [4] and MED-SAM [5] to create an interactive medical image segmentation system capable of responding to user queries with reliable performance. LLaVA-MED is equipped with comprehensive knowledge of various medical terms and resources, providing it with a foundational understanding of medical contexts. Similarly, MED-SAM has been trained to segment medical images from diverse modalities, including MRI and CT scans, across various body regions such as neuroimaging, chest X-rays, and abdominal MRIs. The system leverages joint vision-language embeddings, a context cache of medical features, and advanced feature extraction mechanisms to ensure precise organ segmentation from MRI images. The framework extracts features from the MRI using a Vision Language

979-8-3315-1909-4/24/$31.00 ©2024 IEEE

Projection Embedding, which splits into two streams: one directed towards a prompt encoder that extracts segmentation information features of a particular organ based on user instruction, and the features are sent to the MED-SAM decoder for prediction mask generation, and another towards a Context Cache which is the extracted language features of external knowledge resource of each organ for feature comparison and query validation. This dual-path approach allows the system to intelligently determine query relevance and refine segmentation outcomes. A key focus of our methodology is dealing with false premises, where the user might query for an organ not present in the image. To tackle this, we implement a context-matching between the features extracted by MED-SAM and a cache of pre-computed organ features. This allows the system to intelligently determine whether the query is relevant or not. If the system detects that the requested organ is absent, it avoids producing an incorrect segmentation, ensuring higher reliability. This method not only refines segmentation outcomes but also significantly reduces the risk of false positives in real-world applications. The main contributions of our segmentation framework are as follows:

- Provides a user-guided tool for precise medical organ segmentation.
- Intelligently assesses the relevance of user queries, avoiding incorrect segmentations when the requested organ in the query is absent, thereby enhancing reliability.
- Generates queries from various external resources and introduces an image-query pair dataset for effective model fine-tuning.

II. Related Work

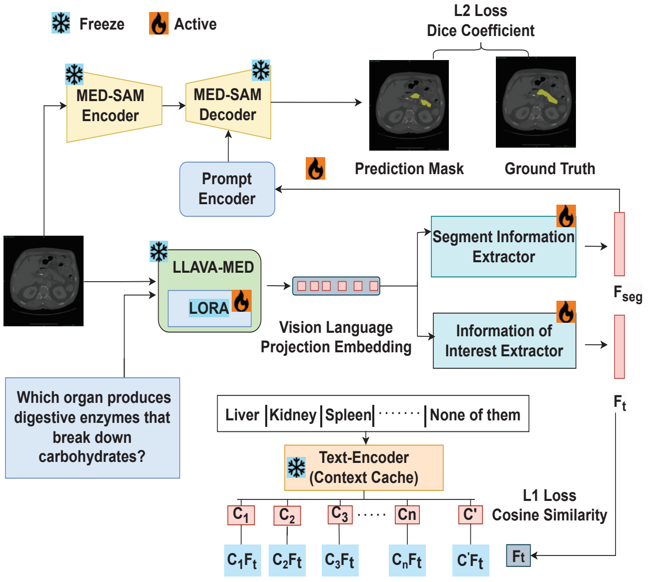
Medical image segmentation aids in identifying organs in scans. Weiwei Tian et al. [6] proposed MOSMOS, excelling on BTCV and AMOS but needing large datasets. Zhang et al. [7] addressed catastrophic forgetting in segmentation but faced pseudo-label and computational issues. Jun Ma et al. [5] developed MedSAM, outperforming traditional models but struggling with boundary accuracy.

These studies are relevant to our research on organ-relevant segmentation. In addition, several papers that utilize vision models or vision-language model techniques for overall segmentation are presented in Table I.

III. Methodology

A. Problem Statement

This work addresses the challenge of *instruction-based medical image segmentation*, where a user query guides the system to identify and segment a specific organ from an abdominal MRI image. The method combines both visual data from the MRI and semantic data from the user's query, aiming to segment the correct organ even when multiple organs are present in the image. The system is designed to respond to queries like *"Which organ is responsible for bile production?"* and generate a segmentation mask for the corresponding organ. The problem is framed using various notations: Let I represent the abdominal MRI image and Q represent the user's query. The joint vision-language embedding F_{vl} is generated by *LLaVA-MED*, encoding both image and query features. The segmentation-related visual feature is denoted as F_{seg}, and the extracted textual feature representing semantic details is denoted as F_t. The system also leverages C', a feature representing irrelevant organs. The system generates a prediction mask $\mathcal{M}_{\text{organ}}$ through the SAM Decoder based on query relevance. The binary indicator I_{out} determines whether the final mask should be valid or blank based on query relevance.



Freeze
Active
L2 Loss
Dice Coefficient
MED-SAM Encoder
MED-SAM Decoder
Prediction Mask
Ground Truth
Prompt Encoder
LLAVA-MED
LORA
Vision Language Projection Embedding
Segment Information Extractor
Information of Interest Extractor
F_{seg}
F_t
Which organ produces digestive enzymes that break down carbohydrates?
Liver | Kidney | Spleen | .... | None of them
Text-Encoder (Context Cache)
C_1
C_2
C_3
Cn
C'
L1 Loss
Cosine Similarity
C_1F_t
C_2F_t
C_3F_t
C_nF_t
C'Ft
Ft

Fig. 1: Illustrates the user-guided segmentation framework. In this framework, a query and abdominal image are input into LLaVA-MED, which generates a vision-language context embedding. The Segment Information Extractor then retrieves segmentation features, which are sent to the pre-trained Prompt Encoder of MED-SAM to generate the prediction mask for the specified organ. The Information of Interest Extractor measures the similarity between features and external text resources, serving as a context cache. L1 Loss optimizes LoRA fine-tuning in both LLaVA-MED and the Information of Interest Extractor, while L2 Loss fine-tunes the Segment Information Extractor and the LoRA in LLaVA-MED.

B. LORA Fine-Tuning in LLAVA-MED

In this system, the *LLaVA-MED Vision-Language Encoder* plays a key role in extracting features containing the context of a segmented organ from both the image and the query. The model undergoes fine-tuning where only the *lora adapter* is updated, while the *vision and text encoder* remains frozen to preserve its pre-trained language understanding. The fine-tuning goal is to adapt the model to medical images, specifically improving its segmentation capabilities.

TABLE I: shows the Relevant of Vision-Language Models for Segmentation

Papers	Datasets	Ideas	Limitations
Lai et al. (2023) [8]	OpenImages, ScanNetv2, ReasonSeg	Introduces reasoning-based segmentation where the model produces segmentation masks from complex text queries.	High computational costs for training the model on large-scale data; challenges in handling extremely long instructions.
Xin Lai et al. (2021) [9]	Cityscapes, Pascal VOC, ADE20K	Proposes Directional Context-Aware Consistency and Directional Contrastive Loss (DC Loss) to improve semi-supervised segmentation by maintaining consistency between features of varying context.	Relies on limited labeled data; computational complexity increases when scaling up to large datasets or higher resolutions.
Zhuotao Tian et al. (2023) [10]	ADE20K	Proposes a context-aware classifier for semantic segmentation that dynamically adjusts decision boundaries based on contextual features in the input. It enhances model performance with minimal extra computational overhead.	Slightly increases inference time and parameters. The method may struggle with noisy or highly varied environments.
Badrinarayanan et al. (2017) [11]	CamVid (road scenes), SUN RGB-D (indoor scenes)	Proposed SegNet, an encoder-decoder architecture for image segmentation using pooling indices for upsampling, reducing memory requirements at inference.	Struggles with fine-grained boundary details in some complex segmentation tasks; limited performance on larger datasets due to memory constraints.
Xueyan Zou et al. (2023) [12]	9 datasets covering interactive segmentation, referring segmentation, video object segmentation, etc.	SEEM introduces a versatile model for multi-task segmentation using visual and textual prompts in a joint visual-semantic space. It supports dynamic composition and interactive segmentation through memory prompts.	The model requires multiple rounds of interactions for refinement and may have limitations in handling complex or overlapping prompts. Some prompt types may not align perfectly in certain cases.
Deyao Zhu et al. (2023) [13]	Conceptual Caption, SBU, LAION (Approx. 5M image-text pairs)	Uses a two-stage training approach with a ViT backbone, Q-Former for visual features, and a linear projection layer to connect vision and LLMs.	Limited in complex reasoning tasks and real-world scenario generalization.
Chenfei Wu et al. (2023) [14]	N/A (Collaborates with existing models)	Integrates ChatGPT with various visual models like BLIP and Stable Diffusion to handle multimodal tasks.	Struggles with fine-grained image editing and often requires detailed prompts to work effectively.
Wenhai Wang et al. (2023) [15]	Vision-language datasets like COCO, Visual Genome	Adopts an open-ended decoding mechanism to extend LLMs' capabilities to vision-related tasks.	Requires substantial computational resources and might face challenges in handling diverse visual tasks.
Henghui Ding et al. (2021) [16]	RefCOCO, RefCOCO+, RefCOCOg	Combines transformers with query generation to enhance performance in referring segmentation.	Lacks efficiency in processing large amounts of natural language descriptions in real-time environments.
Zhaoyang Liu et al. (2023) [17]	LaViT	Combines chatbots with vision models to enhance interaction in visual tasks.	Limited in handling nuanced vision-language interaction, especially when tasks require intricate image details or contextual understanding.
Haotian Liu et al. (2023) [18]	Instruction-based datasets and image-caption pairs	Employs visual instruction tuning and multimodal pretraining to boost model understanding.	Might not generalize well to tasks outside the scope of pre-trained instructions; requires additional fine-tuning on domain-specific datasets for specialized tasks.

C. *Main Workflow*

The core workflow of the system can be described in three stages: processing the user query and image through *LLaVA-MED*, utilizing the *Context Cache from Text Encoder*, and performing segmentation via the *SAM Decoder processed the segmentation prompt* .

1) LLaVA-MED Processing of Question and Image: The first stage involves feeding the input pair consisting of the MRI image I and the prompt generated from user's question Q into *LLaVA-MED*, which produces a joint *vision-language projection embedding* F_{vl}. This embedding encodes both the visual content of the image and the semantic meaning of the query. Two main components use this feature: the *Segment Information Extractor*, which extracts a visual feature of F_{seg} relevant to the organ asked to be segmented in the query which is the segmentation prompt actually, and the *Information of Interest Extractor*, which produces a textual feature F_t representing the queried organ's information and context.

2) Context Cache: The *Context Cache* is a pre-computed feature bank containing text feature representations of various organs. Each organ, C_i, is represented as F_{C_i}, which is generated by passing publicly available medical descriptions through *Text Enocoder*. Additionally, an irrelevant feature representation C', denoted as $F_{C'}$, is generated using irrelevant or

TABLE II: shows the dice coefficient of segmentation for different SAM-based pre-trained models in our dataset in multi-organ segmentation task. It follows only segmentation where user instruction is not given and only segmentation is performed by the models

Methods	**Aorta**	**Duodenum**	**Esophagus**	**Gallbladder**	**IVC**	**Kidney_L**	**Kidney_R**	**LAG**	**Liver**	**Pancreas**	**RAG**	**Spleen**	**Stomach**	**Dice Coefficient**
SAM [3]	80.5	42.7	34.4	50.2	46.5	84.8	89.9	35.4	89.1	51.5	20.6	83.3	69.8	60.5
SAM-Med2D [19]	58.6	46.9	30.3	31.3	19.3	88.9	85.9	28.9	93.8	60.2	12.2	82.0	70.3	51.7
MA-SAM [20]	88.2	92.6	68.4	79.0	78.0	79.4	82.5	48.4	95.1	76.9	46.8	85.2	77.3	73.1
MedSAM [5]	80.9	52.2	66.8	58.7	74.2	87.4	91.8	50.2	91.1	71.2	38.1	87.0	81.5	**74.5**

TABLE III: shows the dice coefficient of segmentation for different SAM-based models after training with our dataset where the user instruction is given and only a particular organ will be segmented based on the instruction or query.

Segmentation Model	**Aorta**	**Duodenum**	**Esophagus**	**Gallbladder**	**IVC**	**Kidney_L**	**Kidney_R**	**LAG**	**Liver**	**Pancreas**	**RAG**	**Spleen**	**Stomach**	**Dice Coefficient**
nn-Unet [21]	44.3	25.6	17.2	25.1	23.2	46.2	49.5	18.6	44.6	25.8	10.3	41.7	35.7	31.2
SAM [3]	50.7	28.1	22.4	31.1	29.7	54.3	58.9	22.5	57.2	33.9	13.2	50.5	43.1	39.9
Swin-Unetr [22]	52.3	27.8	22.4	33.1	30.2	55.2	58.4	23.0	58.5	34.5	13.5	53.1	44.8	40.2
SAM-Med2D [19]	46.9	36.6	23.1	24.5	14.9	68.1	66.3	22.0	74.2	46.2	9.5	61.5	53.2	45.7
MA-SAM [20]	66.2	70.2	52.1	60.2	59.4	61.6	66.0	36.8	71.3	57.9	35.1	63.9	59.0	58.5
MedSAM [5]	71.4	44.4	58.2	49.9	64.3	74.3	78.0	42.7	78.5	62.2	32.4	73.9	69.3	**63.9**

unrelated organ data. To match the query with a corresponding organ, the cosine similarity $\text{sim}(F_t, F_{C_i})$ is computed between the text feature F_t from the Information of Interest Extractor and each organ feature F_{C_i} in the cache:

$$\text{sim}(F_t, F_{C_i}) = \frac{F_t \cdot F_{C_i}}{\|F_t\| \|F_{C_i}\|}$$

If the highest similarity score corresponds to $F_{C'}$, the system identifies the query as irrelevant.

3) Segmentation via MED-SAM Mechanism: The *SAM Decoder* performs the actual segmentation of the organ. The visual feature F_{seg} from the *Segment Information Extractor* is passed to the prompt encoder and extracts the segmentation context feature for *SAM Decoder*, which produces a segmentation mask $\mathcal{M}_{\text{organ}}$.

D. Training and Inference

During training, the model optimizes the *Segment Information Extractor*, the *Info of Interest Extractor*, and the LORA adapter layers in LLAVA-MED. The SAM Encoder, Decoder and the rest of the model operate in a *zero-shot* manner, leveraging pre-trained weights. Two losses are computed during training: The first loss, L1 Loss, is designed to optimize the classification task, where *LLaVA-Med* predicts an organ class label and compares it to the ground truth label from the question-answer pair. This is done using cross-entropy loss, which measures the divergence between the predicted and true labels. The cross-entropy loss is expressed as:

$$\text{L1 Loss} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability for class i, and C is the number of classes, including the abdominal organs and ”none of them.” This L1 Loss will optimize the parameters of both Information of Interest Extractor layers which are nothing but simple neural networks and LORA layers inside LLAVA-MED.

The second loss, L2 Loss, is used to optimize the segmentation task. Once LLaVA-Med predicts the organ class, the decoder of Med-SAM uses this prediction to segment the corresponding organ in the abdominal image after passing through the Segment Information Extractor to reveal the segmentation context. This segmentation later on plays as a prompt in the Prompt encoder which extracts segmentation information for the decoder of MED-SAM and predicts the mask. The *Dice Loss* used to supervise the segmentation mask:

$$\text{L2 Loss} = 1 - \frac{2\,|\mathcal{M}_{\text{organ}} \cap \mathcal{M}_{\text{gt}}|}{|\mathcal{M}_{\text{organ}}| + |\mathcal{M}_{\text{gt}}|}$$

which is used to compare the predicted segmentation mask $\mathcal{M}_{\text{organ}}$ and the ground truth mask $\mathcal{M}_{\text{gt}}$.

IV. Result and Discussion

In this study, we utilized the FLARE22 dataset to assess the performance of our model on external test cases. FLARE22, introduced at MICCAI 2022 as part of a semi-supervised challenge, includes MRI scans of patients with conditions affecting organs such as the liver, kidneys, spleen, and pancreas, providing annotations for 13 organs. For our experiments, we used a subset consisting of 50 MRI (volume) cases. From these cases, we manually selected 500 images for each organ along with their ground truth masks to conduct external evaluations. The dataset was split into an 80% training set and a 20% test set for model training and evaluation. We evaluated the dataset using a SAM-based model with its pre-trained weights. The results are presented in Table II where Med-SAM achieved the highest dice coefficient of the dataset. The methods mentioned in Table 2 are exclusively segmentation models; they are not user-guided segmentation approaches.

The processing of the dataset begins by gathering a dataset of abdominal images along with publicly available health blogs that contain detailed information about 13 abdominal organs, such as the liver, spleen, and pancreas. Using GPT-4, questions are generated from these health blog documents, each designed to ask about a specific organ. For example, a question could be ”Which organ is responsible for filtering blood in the abdomen?” with the answer being ”Liver.” These generated question-answer pairs are then manually reviewed to ensure that only relevant and appropriate questions are retained.

SEGMENTATION TASK RESPONSE PROMPT

USER: `<image_file>`
Please answer the following multiple-choice question based on the image analysis and generate a segmentation mask for the correct organ.

Question: Which organ is responsible for breaking down food using gastric acid?

Options: A. Liver, B. Right kidney, C. Spleen, D. Pancreas, E. Aorta, F. Inferior Vena Cava (IVC), G. Right Adrenal Gland (RAG), H. Left Adrenal Gland (LAG), I. Gallbladder, J. Esophagus, K. Stomach, L. Duodenum, M. Left kidney, N. None of them

Instructions for Assistant:

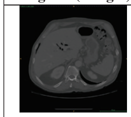
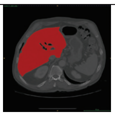
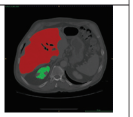
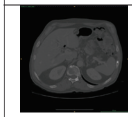
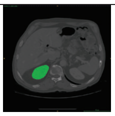
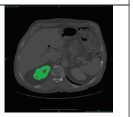
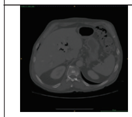
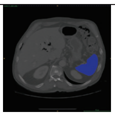
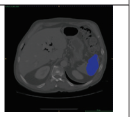
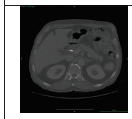
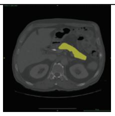
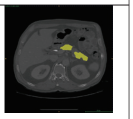
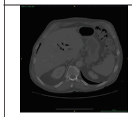
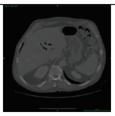
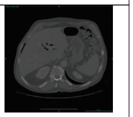
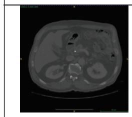
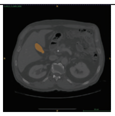
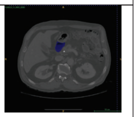
1) Analyze the image carefully and identify the relevant organ.
2) Choose the correct answer from the given options.
3) Provide a short explanation for your answer, clarifying the function of the chosen organ.
4) If none of the options are correct based on the image, select ”None of them.”
5) Generate the segmentation mask for the selected organ and provide segmentation information (bounding box coordinates, organ label).

ASSISTANT:
Answer: K
Explanation: The stomach is responsible for breaking down food using gastric acid, which helps in digestion. The image shows the stomach, confirming it as the correct answer.
Segmentation Information: Organ: Stomach, Bounding Box Coordinates: (x1, y1), (x2, y2), Segmentation Mask: Binary mask with 1 for stomach region, Organ Label: K (Stomach)

Each data sample in the dataset includes an abdominal image, a generated question, the correct organ label (from the set of 13 organs or ”none of them” for irrelevant cases), and a ground truth segmented mask of the organ in the image. This combination of data is used to train the classification and segmentation models, where the classification task identifies the organ based on the question, and the segmentation task focuses on accurately identifying the location of the organ within the image.

Additionally, we leverage prompt engineering to enable the LLaVA-Med model to generate detailed contextual information about the specific organ and incorporate this understanding into the vision-language projection embedding. This method enhances the precision of organ identification and segmentation in medical imaging. The following prompt IV initiates a segmentation task where an assistant analyzes an image, identifies the correct organ based on a multiple-choice question, and generates a segmentation mask.

TABLE IV: Comparison Table of Original, Question, Ground Truth, and Prediction

Original (Images)	Query	Ground Truth	Predictions
	**Question:** Which organ is involved in fat metabolism? **Answer:** Liver		
	**Question:** Which organ filters waste from the blood to produce urine? **Answer:** Kidney		
	**Question:** Which organ filters and removes old red blood cells from circulation? **Answer:** Spleen		
	**Question:** Which organ produces digestive enzymes that break down carbohydrates, fats, and proteins? **Answer:** Pancreas		
	**Question:** Which organ is responsible for breathing? **Answer:** None of them		
	**Question:** Which organ stores bile produced by the liver? **Answer:** Gallbladder **Predicted:** Duodenum Portion **(Misclassified)**		

Then we train our model framework, finetune the specific layers, and test our framework after 1000 training epochs. Table III presents the Dice coefficients for various SAM-based models after being trained on our dataset. The results demonstrate the segmentation performance for organs such as the aorta, liver, and kidneys, among others. MedSAM achieved the highest overall Dice coefficient of 63.9, indicating superior accuracy in organ segmentation tasks compared to other methods.

A series of experiments are conducted to evaluate the framework's performance for our task. For each query, the model was prompted to identify the relevant organ, generate a segmentation mask, and compare its prediction with the ground truth. The results provide insight into the model's performance and areas where misclassifications occurred in Table IV.

We observed instances of misclassification in our model framework, where it incorrectly segmented other organs due to a lack of contextual understanding of user instructions. To further evaluate our model's performance, we assessed the frequency of correct organ classifications based on user queries. The results are illustrated in the confusion matrix shown in Figure 2, highlighting the model's accuracy and areas for improvement.

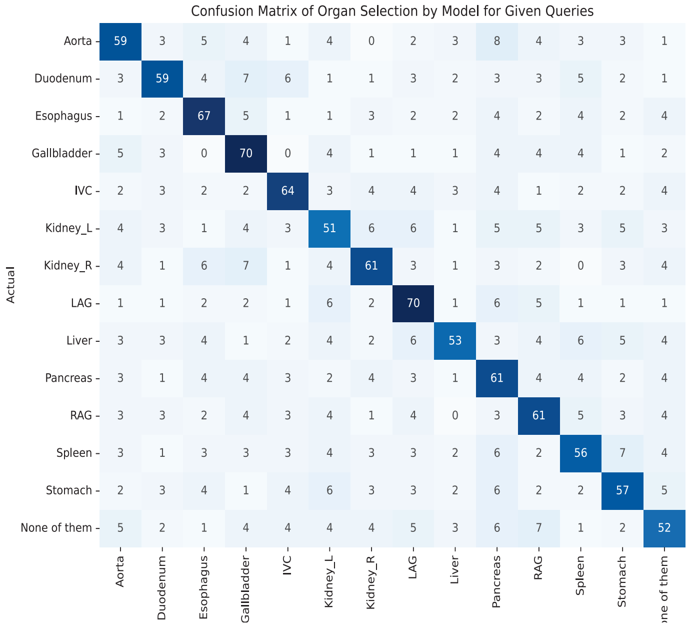


Fig. 2: Confusion matrix illustrating the frequency of correct and incorrect organ classifications by the model based on user queries

V. Conclusion

In this study, we presented an instruction-based medical image segmentation framework that addresses key challenges in user-guided segmentation, particularly in handling false premises. Our model was tested on well-known abdominal organ segmentation datasets, including FLARE22, and demonstrated good accuracy and reliability. The use of a pre-computed context cache from external resources and similarity-based feature comparison further enhanced the system's ability to deliver refined, user-specific segmentation outcomes. However, a current limitation of our system is that it only supports the segmentation of a single organ based on user queries. If a user provides an instruction involving multiple organs (e.g., asking to segment both the stomach and liver in one query), the model is not yet capable of handling this request. Addressing this limitation is a key focus for future work, where we plan to extend the system's functionality to support multi-organ segmentation in response to more complex user queries. Looking ahead, our framework holds great potential for more sophisticated medical imaging tasks and could be integrated into clinical workflows, enhancing the accuracy of diagnostic and treatment planning.

References

[1] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola, *et al.*, "Efficientsam: Leveraged masked image pretraining for efficient segment anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16111–16121, 2024.

[2] G. Sun, C. Qin, J. Wang, Z. Chen, R. Xu, and Z. Tao, "Sq-llava: Self-questioning for large vision-language assistant," *arXiv preprint arXiv:2403.11299*, 2024.

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

[4] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[6] W. Tian, X. Huang, J. Hou, C. Ren, L. Jiang, R.-W. Zhao, G. Jin, Y. Zhang, and D. Geng, "Mosmos: Multi-organ segmentation facilitated by medical report supervision," *arXiv preprint arXiv:2409.02418*, 2024.

[7] Y. Zhang, X. Li, H. Chen, A. L. Yuille, Y. Liu, and Z. Zhou, "Continual learning for abdominal multi-organ and tumor segmentation," in *International conference on medical image computing and computer-assisted intervention*, pp. 35–45, Springer, 2023.

[8] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

[9] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia, "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1205–1214, 2021.

[10] Z. Tian, J. Cui, L. Jiang, X. Qi, X. Lai, Y. Chen, S. Liu, and J. Jia, "Learning context-aware classifier for semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 2438–2446, 2023.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[12] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[14] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023.

[15] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, *et al.*, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16321–16330, 2021.

[17] Z. Liu, Y. He, W. Wang, W. Wang, Y. Wang, S. Chen, Q. Zhang, Z. Lai, Y. Yang, Q. Li, *et al.*, "Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language," *arXiv preprint arXiv:2305.05662*, 2023.

[18] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

[19] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, *et al.*, "Sam-med2d," *arXiv preprint arXiv:2308.16184*, 2023.

[20] C. Chen, J. Miao, D. Wu, A. Zhong, Z. Yan, S. Kim, J. Hu, Z. Liu, L. Sun, X. Li, *et al.*, "Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation," *Medical Image Analysis*, vol. 98, p. 103310, 2024.

[21] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128*, 2019.

[22] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI brainlesion workshop*, pp. 272–284, Springer, 2021.