MTurn-Seg: A Large-Scale Bilingual Medical Benchmark for Multi-Turn Reasoning Segmentation

Haitao Nie[*]
School of Automation
Central South University
Changsha, China
haitaoniedr@gmail.com

Yimeng Zheng[*]
School of Automation
Central South University
Changsha, China
3188563957@qq.com

Ying Ye[*]
School of Automation
Central South University
Changsha, China
13765246401@163.com

Bin Xie[†]
School of Automation
Central South University
Changsha, China
xiebin@csu.edu.cn

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356076

***Abstract*—Multi-turn reasoning segmentation is essential for mimicking real-world clinical workflows, where anatomical structures are identified through step-by-step dialogue based on spatial, functional, or pathological descriptions. However, the lack of a dedicated benchmark in this area has limited progress. To address this gap, we introduce the first bilingual benchmark for multi-turn medical image segmentation, supporting both Chinese and English dialogues. The benchmark consists of 28,904 images, 113,963 segmentation masks, and 232,188 question–answer pairs, covering major organs and anatomical systems across CT and MRI modalities. Each dialogue requires the model to infer the segmentation target based on prior conversational turns and previously segmented regions. We evaluate several state-of-the-art models, including MedCLIP-SAM, LISA, and LISA++, and report three key findings: (1) existing models perform poorly on our benchmark, far below clinical usability standards; (2) performance degrades as dialogue turns increase, reflecting limited multi-turn reasoning capabilities; and (3) general-purpose models such as LISA can outperform medical-specific models, suggesting that further integration of domain knowledge is needed for specialized medical applications. The project and benchmark are available at https://cowboyh.github.io/MTurn-Seg/.**

***Index Terms*—Multi-Turn Medical Reasoning Segmentation, Bilingual, Benchmark**

I. Introduction

Medical image segmentation underpins computer-aided diagnosis, treatment planning, and anatomical analysis. Conventional semantic and instance segmentation methods delineate organs and lesions effectively on CT, MRI, and related modalities, but they are typically trained in closed-set label spaces and assume static, image-only inputs. These assumptions preclude language-conditioned guidance and the incorporation of rich clinical context, limiting flexibility and robustness in real-world workflows.

Foundation models such as SAM [1] and Medical SAM [2] enable class-agnostic, promptable segmentation. In parallel, referring expression segmentation and reasoning segmentation cast the task as localizing and segmenting targets from text(Fig. 1), with the latter requiring inference over implicit relations (for example, mapping "the organ responsible for gas exchange" to the lungs) [3]. Yet most medical segmentation remains single-turn, lacking mechanisms to model history or leverage prior masks [3]–[5]. In practice, successive requests often depend on earlier results and require reasoning over spatial context, anatomy, and clinical knowledge.

Motivated by this gap, we propose the multi-turn reasoning segmentation (MTRS) task in medical imaging. In MTRS, a model receives an input medical image, the current textual instruction, and the interaction history—including prior instructions and previously generated masks—and must produce the next segmentation mask. Each turn may require (i) clinical or anatomical reasoning (e.g., "segment the solid organ in the right upper abdomen involved in glucose metabolism"), (ii) spatial reasoning (e.g., "segment the elliptical structure adjacent to the right side of the abdominal aorta"), or (iii) history-based references (e.g., "segment the necrotic region surrounding the previously segmented tumor"). This formulation better reflects clinical workflows and enables targeted evaluation of a model's cross-turn memory, history-conditioned mask refinement, and language-to-image alignment across turns. However, the field currently lacks a large-scale benchmark for multi-turn reasoning segmentation in medical imaging, which hampers progress.

To address this gap, we introduce the first bilingual benchmark supporting multi-turn dialogue in both Chinese and English (Fig. 1). It comprises 28,904 images, 113,963 segmentation masks, and 232,188 question–answer pairs, covering major organs and anatomical systems across CT and MRI modalities.

We evaluated state-of-the-art models, including MedCLIP-SAM, LISA, and LISA++, and made three key observations:

- Existing models perform poorly on our benchmark, falling significantly short of clinical usability standards.
- The performance and reasoning ability of the models deteriorate as dialogue turns increase.
- General-purpose models outperform medical-specific models; further efforts are needed to integrate domain-specific medical knowledge into specialized models.

II. Benchmark Construction

A. Public Dataset Selection and Inclusion Criteria

We present a multi-turn medical reasoning segmentation benchmark built from multiple public medical imaging

[*]The first three authors contributed equally. [†] Corresponding author.

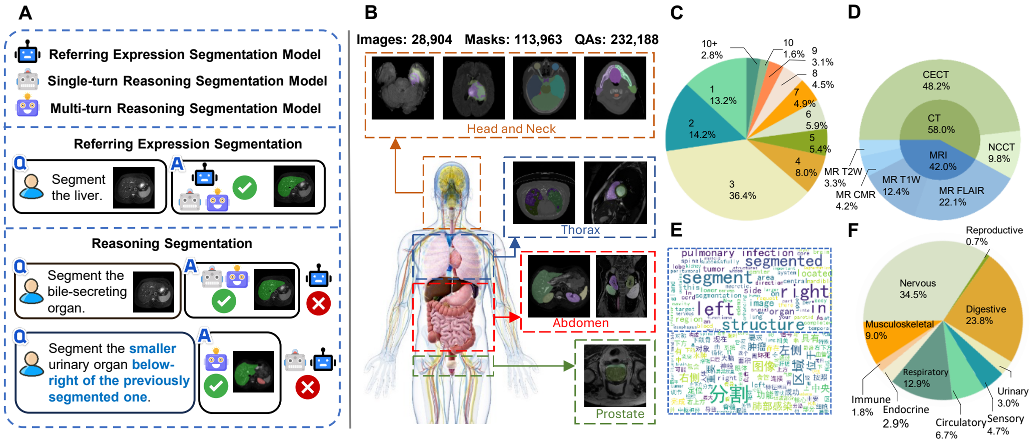

Fig. 1. Research Motivation and Overview of the MTurn-Seg Benchmark. A: Research Motivation; B: Major organs covered by the benchmark; C: Distribution of dialogue turns; D: Distribution of imaging modalities; E: Bilingual word clouds; F: Distribution of QA pairs across human organ systems.

datasets, with rigorous inclusion criteria to ensure comprehensive coverage, clinical representativeness, and reliable annotations. The key principles are:

- **Organ System Coverage**: Covers the major human organ systems—such as the circulatory, respiratory, and digestive—with at least one representative organ from each system.
- **Organ and Lesion Size Coverage**: Both large and small anatomical structures and lesion regions were included to increase diversity in area distribution.
- **High-Quality Segmentation Annotations**: Expert-drawn segmentation masks as reliable ground truth.

In total, we incorporated seven publicly available medical imaging datasets into our benchmark: BraTS-TCGA-GBM [6], SegRap2023 [7], COVID19CTscans [8], CMRxMotions [9], CHAOS Task 4 (T2 modality) [10], AMOS2022 (MR modality) [11], and MSD Prostate [12], [13].

B. Multi-Turn Reasoning Dialogue Generation

Following the clinical workflow, we categorize reasoning into three types:

- **Spatial reasoning**, which involves understanding directions, positions, and the geometric properties of anatomical structures;
- **Functional reasoning**, which identifies target regions based on the physiological functions of organs;
- **Pathological reasoning**, which locates relevant structures according to disease manifestations or abnormal conditions.

To support these reasoning capabilities, we extracted the centroid and pixel area of each segmented object. The centroids were used to compute relative positions and distances between objects, as well as the absolute position of each object within the image. Functional and pathological language descriptions were curated by medical experts based on both Chinese and English medical textbooks, thus eliminating the need for translation. For each organ or lesion, at least 20 diverse descriptions were provided to enhance linguistic variety. Based on this, we constructed a dataset comprising image–mask–text triplets.

We build multi-turn samples by generating each round's instruction from prior segmentation results. Two patterns are used: (1) medical QA grounded in earlier segmentations and (2) new segmentations conditioned on prior outputs. Cross-turn references to any of up to 18 previous rounds (not just the last) are allowed, reflecting real use where users skip rounds or cite earlier content.

Data construction follows a three-stage, human-in-the-loop pipeline: template authoring, LLM-based rewriting, and human review (Fig. 2). We first author templates for two interaction modes and populate them with spatial, functional, and pathological descriptors to create QA pairs. These pairs are paraphrased via the ChatGPT API to increase linguistic diversity while minimizing ambiguity and preserving suitability for medical pretraining. Finally, a 10% random sample undergoes human audit to ensure quality and to correct identified errors.

III. Quantitative Analysis of the Benchmark

A. Benchmark Overview

MTurn-Seg comprises 232,188 bilingual (Chinese–English) reasoning segmentation QA pairs derived from 28,904 medical images with 113,963 expert-annotated masks(Fig. 1). Of these, 228,374 (98.4%) are multi-turn and 3,814 (1.6%) are single-turn. English questions average 27.58 words (median, 27), and Chinese questions average 40.54 characters (median, 40).

The dataset spans 10 major body systems (e.g., immune, endocrine, respiratory) and two imaging modalities (CT and MRI), including 16,764 CT images (non-contrast and contrast-enhanced) and 12,150 MRI images (T1-weighted, T2-weighted, FLAIR, and cardiac), yielding 147,308 and 84,880 QA pairs, respectively.

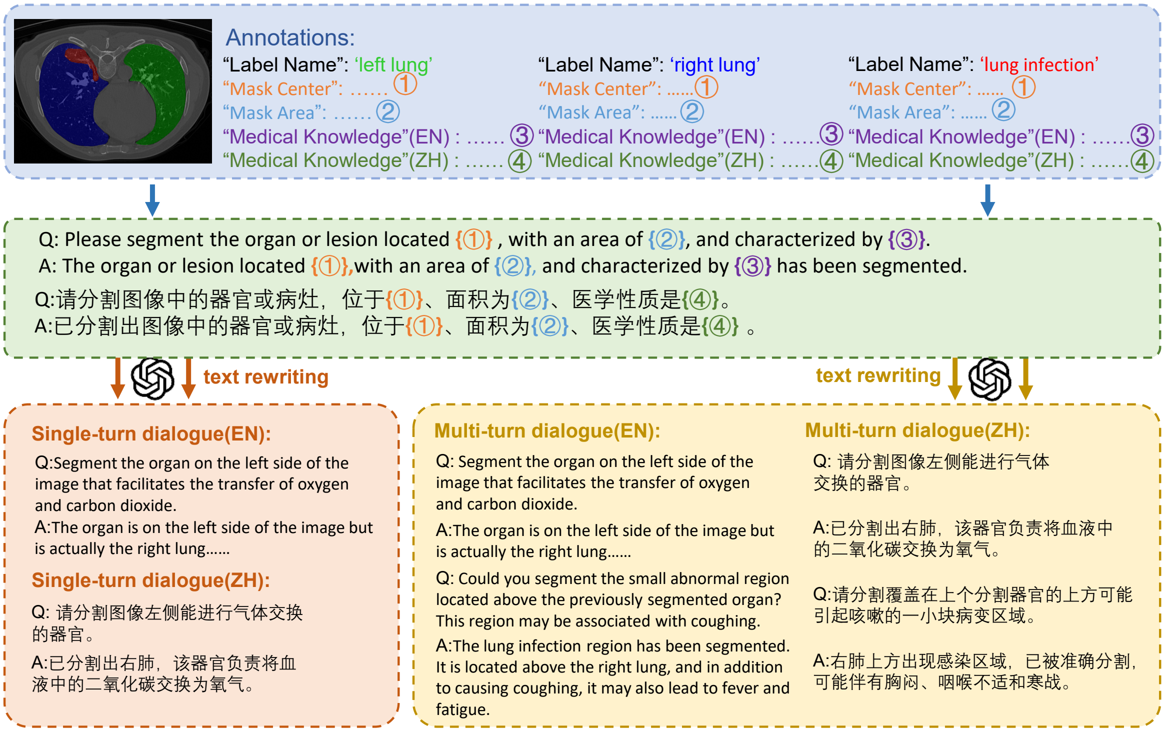

Fig. 2. Dialogue Data Generation

B. Organ and System Coverage

The MTurn-Seg benchmark offers broad coverage of organ systems, making it well-suited as a comprehensive benchmark for whole-body anatomical evaluation. Since an organ may belong to multiple anatomical systems, the total number of masks here does not match the count shown in Fig. 1. The included organs for each system are as follows:

- **Immune System**: Includes the spleen, with a total of 2,121 masks and 4,266 QA pairs.
- **Endocrine System**: Includes the thyroid, pituitary gland, and adrenal glands, with 3,452 masks and 6,904 QA pairs.
- **Respiratory System**: Includes the lungs, trachea, larynx (and subregions), and pharynx, with 15,317 masks and 30,702 QA pairs.
- **Circulatory System**: Includes the heart (left ventricle, right ventricle, myocardium), aorta, and inferior vena cava, with 7,931 masks and 15,864 QA pairs.
- **Sensory System**: Includes the eyes and lenses, cochlea, middle ear, tympanic cavity, auditory tube, and vestibular semicircular canals, with 5,623 masks and 11,274 QA pairs.
- **Urinary System**: Includes the kidneys and bladder, with 3,556 masks and 7,128 QA pairs.
- **Digestive System**: Includes the esophagus, stomach, duodenum, liver, pancreas, gallbladder, oral cavity, pharynx, larynx, and salivary glands (parotid and submandibular), with 28,079 masks and 56,386 QA pairs.
- **Reproductive System**: Includes the peripheral and transitional zones of the prostate, with 748 masks and 1,636 QA pairs.
- **Nervous System**: Includes the brain, brainstem, spinal cord, hippocampus, temporal lobe, optic chiasm, optic nerve, and internal auditory canal, with 38,961 masks and 81,926 QA pairs.
- **Musculoskeletal System**: Includes the mandible, mastoid, and temporomandibular joint, with 10,635 masks and 21,300 QA pairs.

C. Multi-Turn Dialogue Analysis

The Fig. 3 presents illustrative examples of multi-turn medical dialogues, covering both English and Chinese interactions.

Each dialogue contains linguistic features that are essential for reasoning-based segmentation:

- Spatial cues (e.g., "left," "right," "smaller") are highlighted in orange to direct the model to locate structures relative to previously segmented organs or key image landmarks.
- Functional descriptions, shown in blue, provide physiological context (e.g., "metabolizes drugs and detoxifies harmful substances"), helping link textual input to anatomical knowledge.
- Pathological cues, marked in red, simulate clinical prompts that require the model to infer segmentation targets under disease conditions.

IV. Experiments

We focus on the following three questions:

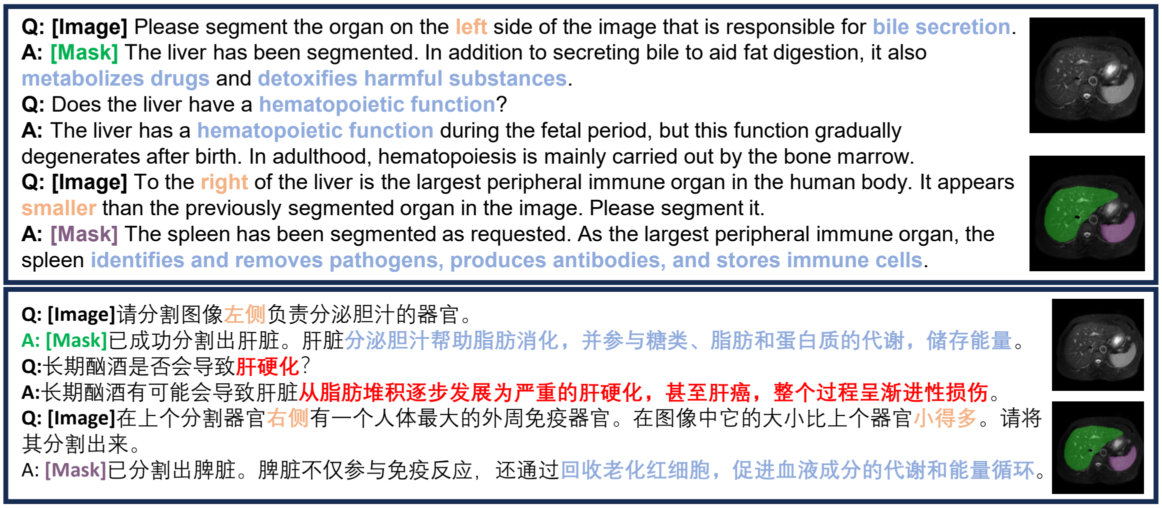

Fig. 3. Multi-Turn Dialogue Example

- **Question 1:** How do existing segmentation models perform on our benchmark?
- **Question 2:** Does model performance degrade when referencing previous segmentation results?
- **Question 3:** Can state-of-the-art models from the natural image domain generalize to medical images?

A. *Experimental Setup*

To address the above questions, we evaluate two categories of models: (1) medical segmentation models, consisting of MEDCLIP-SAM1 (VIT-H) [14], MEDCLIP-SAM1 (VIT-B), and MEDCLIP-SAM1 (VIT-I); (2) general-purpose segmentation models, consisting of LISA-7B [15], LISA-13B, and LISA++ [16].

We conduct single- and multi-turn experiments. Single-turn establishes a history-free baseline by rewriting prompts to replace referential phrases (e.g., "the previous segmentation result") with explicit targets (e.g., "the lung"). Multi-turn conditions on dialogue history to simulate multi-round interactions and assess the performance trend in Question 2.

B. *Single-turn Reasoning Segmentation*

Based on the results presented in Tab. I, we make the following observations:

- **Overall performance is poor, with general-purpose models outperforming medical-specific ones.** The best Dice—0.3099 in EN-Abdomen—remains far below clinical usability. Notably, the generalist LISA (7B) surpasses all domain-specific models, achieving the highest IoU and Dice in both CN-Overall and EN-Overall and leading across multiple anatomical regions.
- **Performance varies by anatomical region.** Abdomen-EN achieves relatively higher scores (IoU 0.2201; Dice 0.3099), whereas head regions are consistently lowest in both CN and EN (IoU <0.12), likely due to fine nasal structures that complicate segmentation.
- **Language Differences Exist, but Are Not Dominant**: While English-language QA pairs tend to yield slightly higher performance than their Chinese counterparts, the gap is neither consistent nor substantial enough to suggest a fundamental language bias. This implies that the primary difficulty likely lies in understanding and reasoning over domain-specific instructions, rather than language itself.

C. *Multi-turn Reasoning Segmentation*

Fig. 4 present the multi-turn segmentation performance across five reasoning rounds, evaluated using mean Dice (Fig. 4A: English, Fig. 4B: Chinese) and mean IoU (Fig. 4C: English, Fig. 4D: Chinese).

The LISA series consistently outperforms the MedCLIP-SAM variants across both languages and metrics. In particular,

TABLE I
SINGLE-TURN REASONING SEGMENTATION. CN AND EN INDICATE DATASETS CONSTRUCTED FROM CHINESE AND ENGLISH CORPORA, RESPECTIVELY. THE BEST VALUE FOR EACH METRIC IS HIGHLIGHTED IN BOLD.

Model	2D Medical Images																			
	Overall				Head				Chest				Abdomen				Prostate			
	CN		EN		CN		EN		CN		EN		CN		EN		CN		EN	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
LISA(13B)	0.0947	0.1484	0.1080	0.1640	0.0485	0.0797	0.0753	0.1210	0.1305	0.1982	0.0465	0.0736	0.0880	0.1351	**0.2201**	**0.3099**	**0.1116**	**0.1805**	0.0900	0.1513
LISA(7B)	**0.1014**	**0.1565**	**0.1423**	**0.2096**	0.0508	0.0840	0.0641	0.1038	**0.1438**	**0.2175**	**0.1361**	**0.1950**	**0.1038**	**0.1527**	0.2025	0.2869	0.1074	0.1719	**0.1666**	**0.2528**
LISA++(7B)	0.0891	0.1395	0.0630	0.1021	0.0513	0.0823	0.0510	0.0877	0.1410	0.2132	0.0617	0.0957	0.0901	0.1336	0.0740	0.1117	0.0741	0.1288	0.0653	0.1135
MedCLIPSAM(VIT-H)	0.0576	0.0912	0.0541	0.0838	0.1004	0.1637	**0.0942**	**0.1489**	0.0545	0.0855	0.0301	0.0496	0.0506	0.0763	0.0626	0.0884	0.0251	0.0393	0.0295	0.0481
MedCLIPSAM(VIT-L)	0.0605	0.0968	0.0524	0.0818	**0.1117**	**0.1817**	**0.0942**	**0.1489**	0.0515	0.0816	0.0288	0.0475	0.0515	0.0816	0.0575	0.0830	0.0274	0.0422	0.0295	0.0479
MedCLIPSAM(VIT-B)	0.0553	0.0880	0.0523	0.0812	0.1026	0.1683	0.0890	0.1419	0.0488	0.0776	0.0282	0.0453	0.0470	0.0714	0.0643	0.0927	0.0227	0.0347	0.0276	0.0451

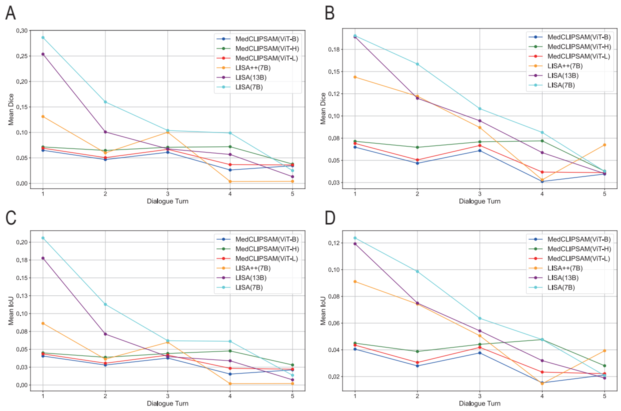

Fig. 4. Multi-turn Reasoning Segmentation. A: mean Dice (English benchmark); B: mean Dice (Chinese benchmark); C: mean IoU (English benchmark); D: mean IoU (Chinese benchmark).

LISA(7B, 13B) and LISA++(7B) achieve higher segmentation accuracy and maintain more stable results in most rounds, indicating stronger reasoning and segmentation capability.

All models exhibit performance degradation as the number of reasoning turns increases, reflecting the growing complexity and accumulated difficulty in multi-turn segmentation tasks. This pattern indicates that errors made in earlier turns may accumulate and propagate, leading to degraded segmentation quality in later stages.

These results highlight the need for models that not only perform well in early rounds but also maintain consistency and robustness throughout extended reasoning processes.

V. Conclusion

We introduce a novel task: multi-turn medical reasoning segmentation. To support this task, we present the first bilingual medical segmentation benchmark, consisting of 28,904 images, 113,963 segmentation masks, and 232,188 question–answer pairs. Experimental results reveal that existing models struggle with multi-turn reasoning and fall significantly short of clinical standards. This benchmark establishes a foundation for advancing research in dialogue-driven, context-aware medical image segmentation.

Acknowledgment

This work was supported by the Hunan Provincial Key Research and Development Program (2025JK2115) and the National High Level Hospital Clinical Research Funding (BJ-2025-133).

References

[1] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma *et al.*, "Sam 2: Segment anything in images and videos," in *International Conference on Representation Learning*, 2025, pp. 28 085–28 128.

[2] J. Zhu, A. Hamdi, Y. Qi, Y. Jin, and J. Wu, "Medical sam 2: Segment medical images as video via segment anything model 2," arXiv preprint arXiv:2408.00874, 2024.

[3] Q. Tong, Z. Lu, J. Liu, Y. Zheng, and Z.-M. Lu, "Medisee: Reasoning-based pixel-level perception in medical images," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 2742–2751.

[4] Q.-H. Trinh, M.-V. Nguyen, J. Peng, U. Bagci, and D. Jha, "Prs-med: Position reasoning segmentation with vision-language model in medical imaging," arXiv preprint arXiv:2505.11872, 2025.

[5] Y. Huang, Z. Peng, Y. Zhao, P. Yang, X. Yang, and W. Shen, "Medseg-r: Reasoning segmentation in medical images with multimodal large language models," arXiv preprint arXiv:2506.10465, 2025.

[6] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby *et al.*, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[7] X. Luo, J. Fu, Y. Zhong, S. Liu, B. Han, M. Astaraki *et al.*, "Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma," *Medical Image Analysis*, vol. 101, p. 103447, 2025.

[8] J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen *et al.*, "Towards data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation," *Medical Physics*, vol. 48, no. 3, pp. 1197–1210, 2021.

[9] S. Wang, C. Qin, C. Wang, K. Wang, H. Wang, C. Chen *et al.*, "The extreme cardiac mri analysis challenge under respiratory motion (cmrxmotion)," arXiv preprint arXiv:2210.06385, 2022.

[10] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza *et al.*, "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[11] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang *et al.*, "Amos: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, pp. 36 722–36 732.

[12] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman *et al.*, "The medical segmentation decathlon," *Nature Communications*, vol. 13, no. 1, p. 4128, 2022.

[13] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," arXiv preprint arXiv:1902.09063, 2019.

[14] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "Medclip-sam: Bridging text and image towards universal medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2024, pp. 643–653.

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

[16] S. Yang, T. Qu, X. Lai, Z. Tian, B. Peng, S. Liu *et al.*, "Lisa++: An improved baseline for reasoning segmentation with large language model," arXiv preprint arXiv:2312.17240, 2024.