2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

ASI-Seg: Audio-Driven Surgical Instrument Segmentation with Surgeon Intention Understanding

Zhen Chen[1,†], Zongming Zhang[1,†], Wenwu Guo[1], Xingjian Luo[1], Long Bai[3], Jinlin Wu[1,2,*], Hongliang Ren[3], Hongbin Liu[1,2]

Abstract— **Surgical instrument segmentation is crucial in surgical scene understanding, thereby facilitating surgical safety. Existing algorithms directly detected all instruments of pre-defined categories in the input image, lacking the capability to segment specific instruments according to the surgeon's intention. During different stages of surgery, surgeons exhibit varying preferences and focus toward different surgical instruments. Therefore, an instrument segmentation algorithm that adheres to the surgeon's intention can minimize distractions from irrelevant instruments and assist surgeons to a great extent. The recent Segment Anything Model (SAM) reveals the capability to segment objects following prompts, but the manual annotations for prompts are impractical during the surgery. To address these limitations in operating rooms, we propose an audio-driven surgical instrument segmentation framework, named ASI-Seg, to accurately segment the required surgical instruments by parsing the audio commands of surgeons. Specifically, we propose an intention-oriented multimodal fusion to interpret the segmentation intention from audio commands and retrieve relevant instrument details to facilitate segmentation. Moreover, to guide our ASI-Seg segment of the required surgical instruments, we devise a contrastive learning prompt encoder to effectively distinguish the required instruments from the irrelevant ones. Therefore, our ASI-Seg promotes the workflow in the operating rooms, thereby providing targeted support and reducing the cognitive load on surgeons. Extensive experiments are performed to validate the ASI-Seg framework, which reveals remarkable advantages over classical state-of-the-art and medical SAMs in both semantic segmentation and intention-oriented segmentation. The source code is available at https://github.com/Zonmgin-Zhang/ASI-Seg.**

I. INTRODUCTION

Developing computer-assisted surgery systems can improve the quality of interventional healthcare for patients [1], [2], [3], [4], [5]. In particular, surgical instrument segmentation stands as a cornerstone for surgical scene understanding [6], [7], [8], [9], which can benefit visual navigation, precise operation, and instrument tracking, thereby facilitating surgical safety and patient outcomes.

To achieve accurate instrument segmentation, existing works [10], [7], [11], [12], [13] have conducted a lot of research from different aspects. For instance, the ISINet [7] further improved the TernausNet [10] by identifying instrument candidates and assigning category labels. In addition, the Dual-MF [11] utilized the motion flow of surgical instruments to benefit segmentation, and the S3Net [12] focused on discriminating instrument categories. Despite great progress, these works directly segmented all instruments of pre-defined categories in the input image, lacking the capability to segment specific instruments according to the surgeon's intention. In clinical practice, surgeons exhibit varying preferences and focus toward different surgical instruments during various stages of the surgery. Therefore, surgical instrument segmentation algorithms that adhere to the surgeon's intention are highly demanded.

The recent advent of the segment anything model (SAM) [14] has revealed the superior robustness and adaptability in natural images in various scenarios. On this basis, SAM has begun to penetrate into the field of medical imaging and demonstrated its capabilities in medical image segmentation [15], [16], [17], [18]. In particular, SAM can segment specific objects based on manual prompts, showing the possibility of segmenting surgical instruments on demand in the operating room. But most existing medical SAM studies [17], [19] rely on more manual annotations, by labeling points or bounding boxes as the prompt. The extensive use of manual annotations interrupts surgical workflows, which is impractical in the operating rooms. Therefore, the ideal surgical instrument segmentation algorithm should eliminate the need for manual annotation, and automatically segment the required surgical instruments based on the intention of surgeons.

To address these limitations of surgical instrument segmentation in the operating rooms, we propose an audio-driven surgical instrument segmentation framework, named ASI-Seg, to accurately segment the required surgical instruments by parsing the audio commands of surgeons. Specifically, we propose an intention-oriented multimodal fusion to interpret the segmentation intention from audio commands, and retrieve relevant instrument details to facilitate segmentation. Moreover, to guide our ASI-Seg segment of the required surgical instruments, we devise a contrastive learning prompt encoder to effectively distinguish the required instruments from the irrelevant ones. In this way, our ASI-Seg promotes the workflow in the operating rooms, thereby providing targeted support and reducing the cognitive load on surgeons.

The contributions of this work are summarized as follows:

- We propose the ASI-Seg framework to achieve audio-driven surgical instrument segmentation based on the

This work was supported by the National Natural Science Foundation of China (Grant No.#62306313), Hong Kong RGC GRF 14211420, 14203323, NSFC/RGC Joint Research Scheme N_CUHK420/22, and InnoHK Funding.

[1]Z. Chen, Z. Zhang, W. Guo, X. Luo, J. Wu and H. Liu are with Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong SAR, China.

[2]J. Wu and H. Liu are also with State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, China.

[3]L. Bai and H. Ren are with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong SAR, China.

†Equal contribution, ∗Corresponding author.

surgeon's intention.

- We devise an intention-oriented multimodal fusion to interpret the intention and retrieve details for ASI-Seg.
- We devise a contrastive learning prompt encoder to distinguish the required instruments from irrelevant ones.
- Extensive experiments on the EndoVis2018 and EndoVis2017 datasets confirm the superior performance of ASI-Seg in both semantic segmentation and intention-oriented segmentation.

II. Related work

Surgical Instrument Segmentation. Existing works [10], [7], [11], [12], [13] conducted surgical instrument segmentation from different aspects. In particular, the TernausNet [10] improved the network structure to achieve accurate instrument segmentation. The ISINet [7] achieved the segmentation by identifying instrument candidates and assigning category labels. The Dual-MF [11] utilized the motion flow of surgical instruments for more accurate segmentation decoding. The S3Net [12] further addressed the difficulty in discriminating instrument categories. In addition, Wang *et al.* [13] blended the irrelevant tissues with required instruments to facilitate segmentation with augmented samples. Different from these works that directly segmented all instruments of pre-defined categories, our ASI-Seg can segment specific instruments according to the surgeon's intention.

The SAM for Medical Imaging. By leveraging both sparse (*e.g.*, point, box, and text) and dense (*e.g.*, mask) prompts, the segment anything model (SAM) [14] has well revealed the advantage in image segmentation across a variety of scenarios. To transfer SAM to downstream scenarios, existing works adopted different fine-tuning strategies, including directly fine-tuning the image encoder [20] or mask decoder [17], and using the parameter efficient fine-tuning (*e.g.*, the low-rank adaptation (LoRA) [21] and adapter [22]), considering the huge amount of SAM parameters. Many medical SAM works [17], [18], [19], [23], [24], [25] have been investigated to customize segmentation capability to medical imaging. Huang *et al.* [18] explored the impact of different prompts on medical image segmentation, and the MedSAM [17] further fine-tuned the SAM with bounding box prompts on large-scale medical image datasets. For the surgical images, the SurgicalSAM [26] introduced the class prototypes and designated target class to guide the segmentation with the category information. In general, most medical SAMs either demand huge computational resources in fine-tuning [20] or rely on manual annotations for prompt during inference [17], [18], [19], [23], [24], which is impractical for clinical usage.

III. Methodology

A. Overview of ASI-Seg Framework

In this work, we propose the ASI-Seg framework to segment required surgical instruments by following the audio commands of surgeons. As elaborated in Fig. 1, we propose the ASI-Seg framework to segment required surgical instruments by following the audio commands of surgeons. Given a surgical image, the ASI-Seg pareses the audio command for segmentation intention and generates the required instrument masks to meet the demand of surgeons.

B. Intention-Oriented Multimodal Fusion

To obtain the features of the surgeon's specified instruments, we propose an Intention-Oriented Multimodal Fusion module in this section. Firstly, we propose an audio intention recognition module to predict the surgeon's segment intention. Then, we propose a text fusion module and a visual fusion module to inject detailed language description information and richer visual information into a group of learnable queries. Lastly, we utilize the recognized audio intention to select the intention-oriented features.

Audio Intention Recognition. We sample the discretion audio signals a' from raw audio signals a with 16K Hz. Then, we transfer the discretion audio signals to the Mel spectrogram as follows:

$$A_{\text{mel}} = \pi(a, a', C_s, W_s, s), \tag{1}$$

where π is the Mel spectrogram transformation [27], C_s is the channel size, W_s is the window size and s is the stride size. For better numerical calculations, we further normalize the scale of A_{mel} to the range of $[-1, 1]$, as follows:

$$A_{\text{norm}} = \frac{2 * (A_{\text{mel}} - \mu)}{\max(A_{\text{mel}}) - \min(A_{\text{mel}})} - 1, \tag{2}$$

where μ is the mean of Mel spectrogram A_{mel} among the training data. To predict the intention of surgeons, we feed A_{norm} to an Audio Encoder E_A and an audio classifier ϕ:

$$\mathcal{C} = \phi(E_A(A_{\text{norm}})), \tag{3}$$

where $\mathcal{C}$ is the audio intention recognition result.

Text Fusion. The surgeon's audio commands may only include the names of instruments. These high-level commands make it challenging for the model to capture the necessary features of the required instruments from visual information. Therefore, we incorporate detailed textual descriptions of each instrument into the learnable query as a complement to the high-level audio commands.

Specifically, we first use a Text Encoder E_T to extract textual features f_t from a pre-prepared Instrument Description Bank $\{B_k\}_{k=1}^{K}$, which stores detailed descriptions of K instruments as follows:

$$f_t = \text{concat}(E_T(\{B_i\}_{k=1}^{K})), \tag{4}$$

where B_k refers to the specific instrument description of k-th instrument, and $f_t \in \mathbb{R}^{K \times d}$ is the concatenated textual feature of all K instruments with feature dimension d.

Then, we initialize the K learnable queries f_c corresponding to each surgical instrument. These queries f_c are then fused with textual feature f_t through a mutual cross-attention

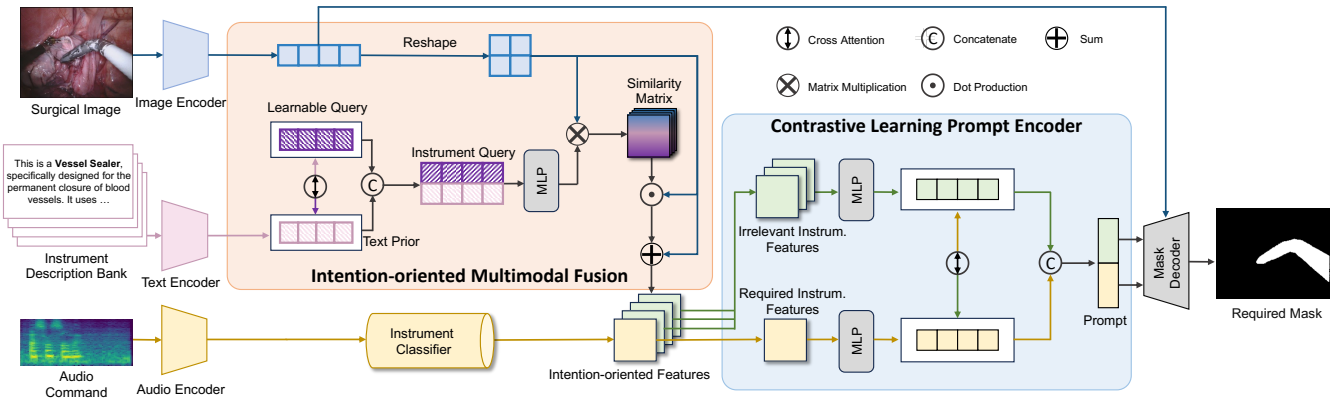

Fig. 1: **Overview of ASI-Seg.** The ASI-Seg mainly consists of the Intention Multimodal Fusion and the Contrastive Learning Prompt Encoder. By parsing the segmentation intention of surgeons, the ASI-Seg first exploits the multimodal knowledge to generate the intention-oriented features and then performs the contrastive learning between required instrument features and the irrelevant ones to produce prompt for segmenting the required instruments.

module to form an Instrument Query q, as follows:

$$\begin{aligned} q_t &= \text{softmax}(\frac{Q_t K_c^T}{\sqrt{D}})V_c, \\ q_c &= \text{softmax}(\frac{Q_c K_t^T}{\sqrt{D}})V_t, \\ q &= \text{MLP}(\text{concat}(q_t, q_c)), \end{aligned} \tag{5}$$

where Q_t, Q_c, K_t, K_c, V_t, V_c are attention queries, keys and values from textual feature f_t and learnable query f_c correspondingly. D is the dimension of the keys and values. The instrument query $q \in \mathbb{R}^{K\times d}$ is a fused result of $q_t \in \mathbb{R}^{K\times d}$ and $q_c \in \mathbb{R}^{K\times d}$. We concatenate them and use a MLP to reduce the concatenated dimension $2d$ to d. These instrument queries fused with detailed textual knowledge can provide more distinguishable information between surgical instruments.

Visual Fusion. We leverage the instrument queries to extract cross-modality visual information from the input image. First, we extract and reshape the image features $f_i \in \mathbb{R}^{h\times w\times d}$ from the image $I \in \mathbb{R}^{H\times W\times 3}$ by an Image Encoder E_I as follows:

$$f_i = E_I(I), \tag{6}$$

where H, W are the original image size and h, w are the reshaped size, d is the feature dimension. Then, we compute the similarity between Instrument Query q and image feature f_i to get a sequence of Similarity Matrix $\{S^k | S^k = q^k \cdot f_i\}_{k=1}^K$, where q_k and S^k are the corresponding query and Similarity Matrix of instrument n. Finally, we add the image feature f_i to the Similarity Matrix sequence $\{S^k\}_{k=1}^K$ to get the Multimodal Features $F = \{f_{i\text{-}t}^k\}_{k=1}^K$ that contains both image and textual information as follows:

$$\{f_{i\text{-}t}^n\}_{k=1}^K = \{f_i \cdot S^k + f_i\}_{k=1}^K, \tag{7}$$

where f_i represents the image feature and S^k refers to the Similarity Matrix of instrument k.

Feature Assignment with Audio Intention. We use $\mathcal{C}$ to divide the Multimodal Feature F into required feature sequence F^+ and irrelevant feature sequence F^-, as follows:

$$\begin{aligned} F^+ &= \{f_{i\text{-}t}^{\mathcal{C}}\}, \\ F^- &= \{f_{i\text{-}t}^{k,k\neq\mathcal{C}}\}_{k=1}^K, \end{aligned} \tag{8}$$

where F^+ refers to feature of the current segment target $\mathcal{C}$, F^- refers to the rest of the features in F, and $f_{i\text{-}t}^k$ is the multimodal feature of surgical instrument k. Following the surgeon's intention, this approach divides the multimodal features F into two groups of Intention-oriented Features.

C. Contrastive Learning Prompt Encoder

To effectively distinguish between required and irrelevant instrument features, we design the Contrastive Learning Prompt Encoder to provide the mask decoder with the specific prompt of the instrument to be segmented.

Distinguishing Cross-Attention. We employ a mutual cross-focusing mechanism between the required instrument feature F^+ and the irrelevant instrument feature F^-, which aims to enhance the focus on the unique properties of the surgical instruments to be segmented. Firstly, we compute the attention similarity to obtain easily confounded regions as follows:

$$\text{Attention}(F^+, F^-) = \text{softmax}\left(\frac{Q_{F^+} K_{F^-}^T}{\sqrt{D}}\right) V_{F^-}, \tag{9}$$

where Q_{F^+}, K_{F^-}, V_{F^-} are attention query, key, and value from the required instrument feature F^+ and the irrelevant instrument feature F^- correspondingly. In addition, $\text{Attention}(F^-, F^+)$ is the same.

Then, we adopt an inverse residual mechanism as follows:

$$P^* = P - \text{Attention}(F^+, F^-), \tag{10}$$

where P^* is the output required instrument feature. P^* eliminates information similar to the irrelevant instrument feature and maintains its unique attributes and characteristics, which is essential for accurate segmentation.

Contrastive Learning. To further push relevant instrument features and irrelevant instrument features to be separable, we design a contrast learning between the required

TABLE I: Semantic Segmentation Comparison on the EndoVis2018 Dataset.

Method	Challenge IoU	IoU
TernausNet [10]	46.22	39.87
MF-TAPNet [8]	67.87	39.14
Dual-MF [11]	70.40	-
ISINet [7]	73.03	70.94
S3Net [12]	75.81	74.02
MaskTrack-RCNN [28] + SAM [14]	78.49	78.49
Mask2Former [29] + SAM [14]	78.72	78.72
TrackAnything (1 Point) [30]	40.36	38.38
TrackAnything (5 Points) [30]	65.72	60.88
PerSAM [31]	49.21	49.21
PerSAM (Fine-Tune) [31]	52.21	52.21
SurgicalSAM [26]	80.33	80.33
ASI-Seg (Ours)	**82.37**	**82.37**

TABLE II: Semantic Segmentation Comparison on the EndoVis2017 Dataset.

Method	Challenge IoU	IoU
TernausNet [10]	35.27	12.67
MF-TAPNet [8]	37.25	13.49
Dual-MF [11]	45.80	-
ISINet [7]	55.62	52.20
TraSeTr [35]	60.40	-
S3Net [12]	72.54	71.99
Mask2Former [29] + SAM [14]	66.21	66.21
TrackAnything (1 Point) [30]	54.90	52.46
TrackAnything (5 Points) [30]	67.41	64.50
PerSAM [31]	42.47	42.47
PerSAM (Fine-Tune) [31]	41.90	41.90
SurgicalSAM [26]	69.94	69.94
ASI-Seg (Ours)	**71.64**	**71.64**

instrument features and the irrelevant ones. Specifically, the contrastive learning loss $\mathcal{L}_{\text{CL}}$ is defined using three parameters, including the required instrument features P, the irrelevant instrument features N, as well as features v from image embeddings filtered through ground truth masks. The formula is as follows:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{K}\sum_{n=1}^{K}\log\frac{\exp(P^{(\mathcal{C})}\cdot v^{(\mathcal{C})}/\tau)}{\sum_{n=1}^{K}\exp(P^{(\mathcal{C})}\cdot v^{(n)}/\tau)}, \tag{11}$$

where τ refers to the temperature factor, and $P^{(\mathcal{C})}$ represents the required instrument features of class $\mathcal{C}$. This contrastive loss pushes the required instrument away from the irrelevant instrument features, enhancing the feature discrimination capability of our ASI-Seg.

Mask Decoder. We adapt the SAM [14] mask decoder to generate the mask of the required instruments. Our ASI-Seg exhibits enhanced differentiation between required instruments and irrelevant instruments using tailored contrastive learning. This distinction significantly augments the segmentation capability of the SAM mask decoder. We regard the features derived from required instruments as foreground prompts and the features obtained from irrelevant instruments as background prompts for the SAM mask decoder. Therefore, the ASI-Seg can generate the accurate mask of the required surgical instruments with comprehensive prompts.

D. Optimization

In the training of ASI-Seg, we freeze the image encoder, the audio encoder and the text encoder with massive parameters, and merely optimize the lightweight instrument classifier and mask decoder, as well as the proposed intention-oriented multimodal fusion and contrastive learning prompt encoder, which makes the end-to-end training efficient. The ASI-Seg is optimized by two loss terms, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{DICE}} + \mathcal{L}_{\text{CL}}, \tag{12}$$

where the dice loss $\mathcal{L}_{\text{DICE}}$ [32] is for segmentation and the contrastive learning loss $\mathcal{L}_{\text{CL}}$ is used to dynamically update the learnable query in the ASI-Seg. In this way, ASI-Seg is capable of segmenting the required instruments according to the intention of surgeons.

IV. Experiment

A. Implementation and Datasets

Dataset. We perform the comprehensive evaluation on the EndoVis2018 [33] and EndoVis2017 datasets [34]. To guarantee fair comparisons, we follow the standard protocol [7], [10]. Specifically, the EndoVis2017 dataset, comprising eight videos, is subjected to a 4-fold cross validation [10]. The video sequences with a high resolution of $1,280 \times 1,024$ are acquired from *da Vinci Xi* surgical system during different porcine procedures. Meanwhile, the EndoVis2018 dataset encompasses 11 training videos alongside four validation videos, thereby presenting a comprehensive platform for benchmarking. Both datasets feature seven unique categories of surgical instruments, enabling an in-depth assessment of our segmentation effectiveness.

Implementation Details. We implement our ASI-Seg in PyTorch on a single NVIDIA A800 GPU. In our ASI-Seg, we use the pre-trained ViT [36] as the image encoder, and use the text encoder of CLIP [37] as the text encoder, and the pre-trained audio encoder [38] as our audio encoder. Additionally, we randomly initialize audio embeddings as the category query. To enhance architectural stability and concentrate on novel components, we maintain static image encoders while dynamically updating the learnable query and mask decoder weights. We set the temperature factor τ of contrastive loss as 0.07, Adam as the optimizer with a learning rate of 0.0001 across both datasets to accommodate their distinct complexities. The training leverages pre-computed image embeddings and a batch size of 16 for EndoVis2017 and 64 for EndoVis2018 datasets.

Evaluation Metrics. We perform the evaluation using three critical segmentation metrics following [26], including the Challenge IoU [34], IoU, and mean class IoU (mc IoU) [7], [12]. These metrics ensure our ASI-Seg is rigorously measured and validated against these benchmarks.

B. Comparisons with State-of-the-arts

We conduct the comprehensive comparison between our ASI-Seg framework and state-of-the-art surgical instrument segmentation methods and advanced SAM approaches on the EndoVis2018 [33] and EndoVis2017 [34] datasets.

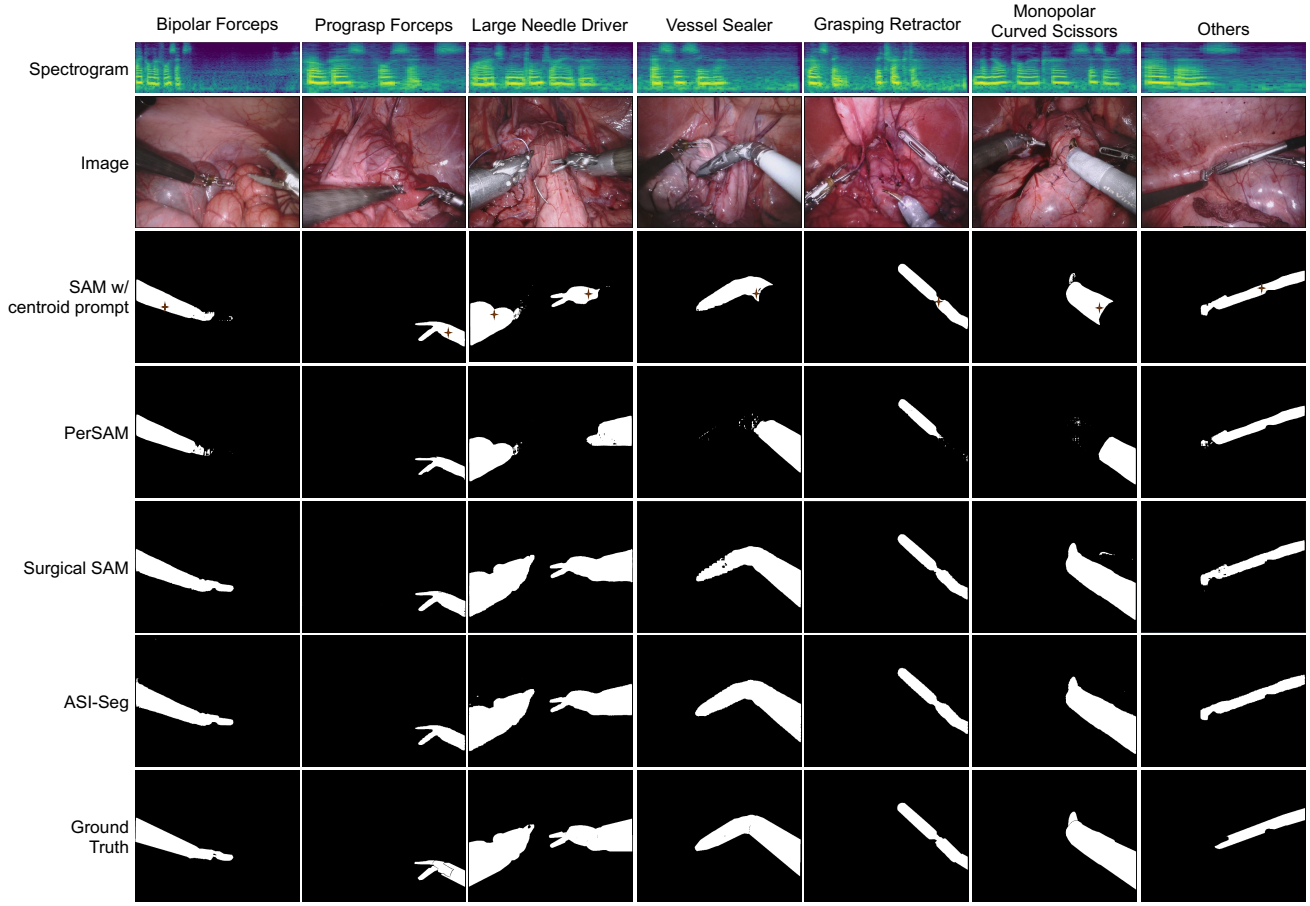

Fig. 2: Qualitative comparison of intention-oriented segmentation on the EndoVis2017 dataset.

TABLE III: Intention-oriented Segmentation Comparison on the EndoVis2018 Dataset.

Method	mc IoU	BF	PF	LND	SI	CA	MCS	UP
MaskTrack-RCNN [28] + SAM [14]	56.07	79.83	74.86	43.12	62.88	16.74	91.62	23.45
Mask2Former [29] + SAM [14]	52.50	**85.95**	**82.31**	44.08	0.00	**49.80**	**92.17**	13.18
TrackAnything (1 Point) [30]	20.62	30.20	12.87	24.46	9.17	0.19	55.03	12.41
TrackAnything (5 Points) [30]	38.60	72.90	31.07	**64.73**	10.24	12.28	61.05	17.93
PerSAM [31]	34.55	51.26	34.40	46.75	16.45	15.07	52.28	25.62
PerSAM (Fine-Tune) [31]	37.24	57.19	36.13	53.86	14.34	25.94	54.66	18.57
SurgicalSAM [26]	58.87	83.66	65.63	58.75	54.48	39.78	88.56	21.23
ASI-Seg (Ours)	**64.18**	83.12	65.87	59.24	**90.43**	34.90	60.10	**55.62**

Semantic Segmentation Analysis. As shown in Table I and Table II, we first perform the comparison of semantic segmentation by segmenting all the instruments in the input image on the EndoVis2018 and EndoVis2017 datasets, respectively. In general, our ASI-Seg achieves the best performance in the semantic segmentation landscape, with IoU of 82.37% and 71.64% on the EndoVis2018 and EndoVis2017 datasets, respectively. Note that our ASI-Seg outperforms the second-best method [26] with an IoU advantage of 2.04% and 1.70% on these two datasets. These improvements indicate a profound improvement of our ASI-Seg in the model capacity to distinguish surgical instruments from irrelevant ones and complex backgrounds.

Intention-oriented Segmentation Analysis. To evaluate the capability of the ASI-Seg, we perform the comparison of intention-oriented segmentation, as shown in Table III and Table IV for EndoVis2018 and EndoVis2017 datasets, respectively. This comparison encompasses a broad spectrum of surgical instruments, including Bipolar Forceps (BF), Prograsp Forceps (PF), Large Needle Driver (LND), Suction Instrument (SI), Vessel Sealer (VS), Clip Applier (CA), Grasping Retractor (GR), Monopolar Curved Scissors (MCS), and Ultrasound Probe (UP). Specifically, we calculate the IoU of each category with segmentation intention and average them for the mean class IoU (mc IoU). Our ASI-Seg achieves the superior mc IoU of 64.18% and 68.17% in the EndoVis2018 and EndoVis2017 datasets, with overwhelming improvement over advanced SAM approaches. In particular, our ASI-Seg reveals the advantage of 5.31% in mc IoU over the second-best SurgicalSAM with category prompt

TABLE IV: Intention-oriented Segmentation Comparison on the EndoVis2017 Dataset.

Method	mc IoU	BF	PF	LND	VS	GR	MCS	UP
Mask2Former [29] + SAM [14]	55.26	66.84	**55.36**	**83.29**	73.52	26.24	36.26	45.34
TrackAnything (1 Point) [30]	55.35	47.59	28.71	43.27	82.75	**63.10**	66.46	55.54
TrackAnything (5 Points) [30]	62.97	55.42	44.46	62.43	**83.68**	62.59	67.03	**65.17**
PerSAM [31]	41.80	53.99	25.89	50.17	52.87	24.24	47.33	38.16
PerSAM (Fine-Tune) [31]	39.78	46.21	28.22	53.12	57.98	12.76	41.19	38.99
SurgicalSAM [26]	67.03	68.30	51.77	75.52	68.24	57.63	86.95	60.80
ASI-Seg (Ours)	**68.37**	**73.92**	47.61	80.33	75.44	52.60	**89.78**	58.90

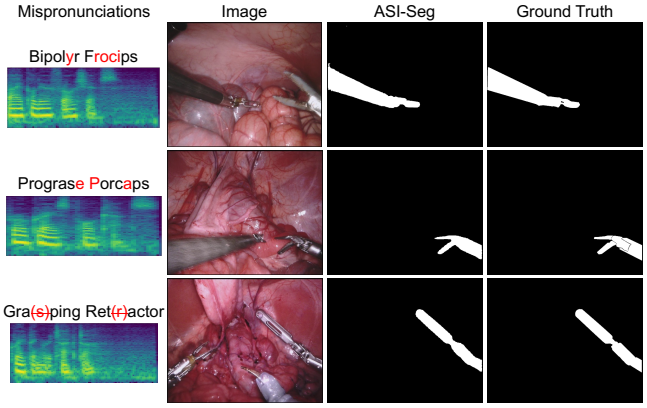

Fig. 3: Impact of mispronunciations on our ASI-Seg.

[26] on the EndoVis2018 dataset, and also achieves the balanced segmentation performance in different categories of surgical instruments, as shown in Table III. As such, these comparisons confirm the versatility of our ASI-Seg across a wide range of surgical scenarios to meet the requirements of surgeons.

Furthermore, we qualitatively compare the segmentation masks of the ASI-Seg and SAM-based approaches with the intention of segmenting instruments of each category, as illustrated in Fig. 2. It is worth noting that our ASI-Seg does not require manual annotations [14] or assigned category [26] for prompt. In the comparison, our ASI-Seg correctly understands the segmentation intention and generates the most accurate masks of the required instruments. These enhancements highlight the proficiency of ASI-Seg in identifying and classifying diverse surgical instruments. In general, the performance advantage of ASI-Seg substantiates the superiority in surgical instrument segmentation.

C. Robustness Study

We further investigate the robustness of our ASI-Seg against defective audio commands, *e.g.*, the mispronunciation of instrument names. As illustrated in Fig. 3, when there are obvious mispronunciations in the input audio, *e.g.*, surgeons may mistakenly articulate the Bipolar Forceps as *Bipolyr Frocips*, our ASI-Seg is still capable to recognize the intention into the correct instrument category and complete accurate segmentation. These results confirm the robustness of our ASI-Seg to identify instruments that surgeons intend to use despite verbal errors, which is also an advantage compared to text instructions.

TABLE V: Ablation Study on our ASI-Seg.

Instrument Description Bank	Contrastive Learning	Challenge IoU	IoU	mc IoU
		76.14	76.14	51.00
✓		80.17	80.17	59.42
	✓	78.63	78.63	55.98
✓	✓	**82.37**	**82.37**	**64.18**

D. Ablation Study

To validate the effectiveness of the proposed modules, we perform the ablation study on the EndoVis2018 [33] dataset, as shown in Table V. Compared with the vanilla baseline, our framework with the instrument description bank gains a 8.42% increase in mc IoU. This confirms that the integration of textual knowledge is a pivotal component in cultivating distinct learnable queries for different categories, thereby ameliorating the precision of instrument segmentation. On the other hand, our framework obtains a 4.98% increase in mc IoU when adding the contrastive learning in ASI-Seg. In our ASI-Seg, the advantage provided by contrastive learning is mainly attributed to its ability to dynamically emphasize the required instrument features while attenuating irrelevant ones. Therefore, the contrastive learning enables the ASI-Seg to become more proficient in differentiating instruments by focusing on the attributes necessary for differentiation. In this way, the proposed ASI-Seg benefits from these tailored designs, resulting in the performance advantage in surgical instrument segmentation at the operating rooms.

V. CONCLUSIONS

In this work, we propose the ASI-Seg framework to accurately segment the required surgical instruments by parsing the audio commands of surgeons. In our ASI-Seg framework, the intention-oriented multimodal fusion can interpret the segmentation intention and retrieve relevant instrument details to facilitate segmentation. Moreover, the contrastive learning prompt encoder can distinguish the required instruments from the irrelevant ones to guide our ASI-Seg segment of the required surgical instruments. Therefore, our ASI-Seg can minimize distractions from irrelevant instruments assist surgeons to a great extent, and promote the workflow in the operating rooms. Extensive experiments are performed to validate the ASI-Seg framework, which reveals remarkable advantages over classical state-of-the-art and advanced SAMs in both semantic segmentation and intention-oriented segmentation.

REFERENCES

[1] Z. Chen, Q. Guo, L. K. Yeung, D. T. Chan, Z. Lei, H. Liu, and J. Wang, "Surgical video captioning with mutual-modal concept alignment," in *MICCAI*. Springer, 2023, pp. 24–34.

[2] X. Luo, Y. Pang, Z. Chen, J. Wu, Z. Zhang, Z. Lei, and H. Liu, "Surgplan: Surgical phase localization network for phase recognition," in *ISBI*. IEEE, 2024.

[3] Z. Chen, Y. Zhai, J. Zhang, and J. Wang, "Surgical temporal action-aware network with sequence regularization for phase recognition," in *BIBM*. IEEE, 2023, pp. 1836–1841.

[4] L. Bai, Q. Tan, T. Chen, W. J. Nah, Y. Li, Z. He, S. Yuan, Z. Chen, J. Wu, M. Islam, *et al.*, "Endouic: Promptable diffusion transformer for unified illumination correction in capsule endoscopy," in *MICCAI*. Springer, 2024.

[5] H. Xu, J. Wu, G. Cao, Z. Chen, Z. Lei, and H. Liu, "Transforming surgical interventions with embodied intelligence for ultrasound robotics," in *MICCAI*. Springer, 2024.

[6] L. Sestini, B. Rosa, E. De Momi, G. Ferrigno, and N. Padoy, "Fun-sis: A fully unsupervised approach for surgical instrument segmentation," *Medical Image Analysis*, vol. 85, p. 102751, 2023.

[7] C. González, L. Bravo-Sánchez, and P. Arbelaez, "Isinet: An instance-based approach for surgical instrument segmentation," in *MICCAI*, 2020, p. 595–605.

[8] Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *MICCAI*. Springer, 2019, pp. 440–448.

[9] Z. Sun, H. Xu, J. Wu, Z. Chen, Z. Lei, and H. Liu, "Pwiseg: Weakly-supervised surgical instrument instance segmentation," in *ICIP*, 2024.

[10] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *ICMLA*, 2018, pp. 624–628.

[11] Z. Zhao, Y. Jin, X. Gao, Q. Dou, and P.-A. Heng, "Learning motion flows for semi-supervised instrument segmentation from robotic surgical video," in *MICCAI*. Springer, 2020, pp. 679–689.

[12] B. Baby, D. Thapar, M. Chasmai, T. Banerjee, K. Dargan, A. Suri, S. Banerjee, and C. Arora, "From forks to forceps: A new framework for instance segmentation of surgical instruments," *WACV*, pp. 6180–6190, 2022.

[13] A. Wang, M. Islam, M. Xu, and H. Ren, "Rethinking surgical instrument segmentation: A background image can be all you need," in *MICCAI*, 2022, pp. 355–364.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *ICCV*, October 2023, pp. 4015–4026.

[15] 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.

[16] Y. Zhang, Z. Shen, and R. Jiao, "Segment anything model for medical image segmentation: Current applications and future directions," *arXiv preprint arXiv:2401.03495*, 2024.

[17] 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.

[18] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, *et al.*, "Segment anything model for medical images?" *Med. Image Anal.*, vol. 92, p. 103061, 2024.

[19] X. Lin, Y. Xiang, L. Zhang, X. Yang, Z. Yan, and L. Yu, "Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation," *arXiv preprint arXiv:2309.06824*, 2023.

[20] Z. Shui, Y. Zhang, K. Yao, C. Zhu, Y. Sun, and L. Yang, "Unleashing the power of prompt-driven nucleus instance segmentation," *arXiv preprint arXiv:2311.15939*, 2023.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *ICLR*, 2022.

[22] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *ICML*. PMLR, 2019, pp. 2790–2799.

[23] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.13785*, 2023.

[24] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[25] Z. Chen, Q. Xu, X. Liu, and Y. Yuan, "Un-sam: Universal prompt-free segmentation for generalized nuclei images," *arXiv preprint arXiv:2402.16663*, 2024.

[26] W. Yue, J. Zhang, K. Hu, Y. Xia, J. Luo, and Z. Wang, "Surgicalsam: Efficient class promptable surgical instrument segmentation," in *AAAI*, 2024.

[27] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[28] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *ICCV*, 2019, pp. 5188–5197.

[29] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.

[30] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," *arXiv preprint arXiv:2304.11968*, 2023.

[31] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, Y. Qiao, P. Gao, and H. Li, "Personalize segment anything model with one shot," in *ICLR*, 2024.

[32] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *IEEE International Conference on 3D Vision*, 2016.

[33] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, *et al.*, "2018 robotic scene segmentation challenge," *arXiv preprint arXiv:2001.11190*, 2020.

[34] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[35] Z. Zhao, Y. Jin, and P.-A. Heng, "Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery," in *ICRA*. IEEE, 2022, pp. 11 186–11 193.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[38] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. PMLR, 2023, pp. 28 492–28 518.