GSAM+Cutie: Text-Promptable Tool Mask Annotation for Endoscopic Video

Roger D. Soberanis-Mukul*†
Johns Hopkins University
Baltimore, 21211, MD, USA
rsobera1@jhu.edu

Jiahuan Cheng*
Johns Hopkins University
Baltimore, 21211, MD, USA
jcheng65@jhu.edu

Jan Emily Mangulabnan*†
Johns Hopkins University
Baltimore, 21211, MD, USA
jmangul1@jh.edu

S. Swaroop Vedula
Johns Hopkins University
Baltimore, 21211, MD, USA
swaroop@jhu.edu

Masaru Ishii
Johns Hopkins Medical Institutions
Baltimore, 21287, MD, USA
mishii3@jhmi.edu

Gregory Hager
Johns Hopkins University
Baltimore, 21211, MD, USA
hager@cs.jhu.edu

Russell H. Taylor
Johns Hopkins University
Baltimore, 21211, MD, USA
rht@jhu.edu

Mathias Unberath†
Johns Hopkins University
Baltimore, 21211, MD, USA
unberath@jhu.edu

Abstract

Machine learning approaches for multi-view geometric scene understanding in endoscopic surgery often assume temporal consistency across the frames to limit challenges that algorithms contend with. However, in monocular scenarios where multiple views are acquired sequentially rather than simultaneously, the static scene assumption is too strong because surgical tools move during the procedure. To enable multi-view models despite tool motion, masking these temporally inconsistent tool regions is a feasible solution. However, manual tool-masking requires a prohibitive effort, given that endoscopic video can contain thousands of frames. This underscores the need for (semi-)automated techniques to 1) automatically mask the tools and/or 2) semi-automatically annotate large datasets such that algorithms for 1) may be developed. To facilitate semi-automated annotation, any solution must be both generalizable, such that it can be used out-of-the-box on diverse datasets, and easy to use. Recent methods for surgical tool annotation require either fine-tuning on domain-specific data or excessive user interaction, limiting their application to new data. Our work introduces GSAM+Cutie, a surgical tool annotation process relying on a combination of two recent foundation models for text-based image segmentation and video object segmentation. We show that a combination of Grounded-SAM and Cutie models provides good generalization for robust text-prompt-based video-level binary segmentation on endoscopic video, streamlining the mask annotation task. Through quantitative evaluation on two datasets, including a proprietary in-house dataset and EndoVis, we show that GSAM+Cutie outperforms similar ensembles, like SAM-PT, for video object segmentation. We also discuss the limitations and future research directions that GSAM+Cutie can motivate. Our code is available at https://github.com/arcadelab/cutie_plus_gsam

1. Introduction

Surgical tool masking in endoscopic video is an important data curation and annotation step in computer-aided medical procedures, as tools are the main point of interaction with the anatomical tissue. Tool segmentation masks delineate the static anatomical environment from the dynamic movement of the surgical tools, enabling advanced visualization and navigation for vision-based tool-tracking and surgical scene understanding in minimally invasive surgeries. Recent approaches for multi-view geometric scene understanding [12, 15], however, rely on a static scene assumption. This limits the availability of data that can be employed in these algorithms as tool movement propagates inconsistencies in the scene. The generation of surgical tool masks can mitigate this issue by only extracting relevant anatomical information, allowing such models to be employed on a wider variety of endoscopic videos.

Despite the benefits of surgical tool masking, manual an-

*Share first author
†Corresponding authors

notation represents a challenging and time-consuming bottleneck, considering that the primary source of information in endoscopy is a video stream. This has motivated the development of deep learning models and datasets aiming to automate or semi-automate this process [1–3, 7, 11, 22]. However, these models face challenges from the lack of training annotations in different endoscopic domains, and hence, the resulting networks to date are largely domain-specific. For example, a tool segmentation model trained for laparoscopy can have reduced performance in sinus endoscopy due to the domain shift derived from the different procedures.

Considering that tool mask annotation would be employed in the early stages of the data processing pipeline, we consider that this tool masking process should be generalizable out of the box (ideally, without requiring fine-tuning or retraining). Additionally, since users of different backgrounds can perform the annotation and curation task, this process must be accessible and easy to use. Recently, the introduction of novel large vision models (LVM) like Segment Anything [10] (SAM) and CoTracker [8] have set new standards on model generalizability. These foundation models trained on large datasets have shown strong zero- or few-shot capabilities to different input domains in tasks like segmentation (SAM) and video point-tracking (CoTracker), representing an attractive property for the tool masking process.

However, early works on applying SAM to the medical domain show a set of limitations derived from the differences in the appearance of the medical data compared with the real-world scenes used to train these LVMs [6, 9, 17, 18]. Another limitation is the requirement of a visual prompt and the sensitivity of the model's results to the prompt selection [21, 23]. While different works have been proposed to overcome these limitations, they require fine-tuning the models. The use of ground truth for proper fine-tuning brings us to the initial annotation availability problem. Additionally, given the number of parameters in LVMs, the risk of overfitting imposes additional challenges to the translation of these models for automatic medical data processing.

Some works in the literature have suggested that combining LVM at different levels can help to overcome some of their individual limitations. For example, employing the robust point-tracking capabilities of CoTracker and the strong prompt-guided segmentation performance of SAM can allow SAM to perform video-level segmentation in an iterative form [19]. Similarly, the integration of the text-prompt object detection of grounding DINO [13] with SAM, allows for text-based object segmentation, resulting in a simpler prompt strategy, compared with visual prompts [20]. In this regard, we show that Cutie [4], a recent model for video object segmentation, is able to propagate an initial tool mask across the endoscopic video but lacks mechanisms for proper mask initialization. At the same time, the text-prompt capabilities of grounded SAM allow for binary tool segmentation in single endoscopic frames, which can provide the initial mask required by Cutie to segment the whole video. Furthermore, the use of text represents a simplified mechanism to prompt the model that only requires describing the scene with natural language, avoiding the direct use of visual prompts.

In this work, we evaluate the performance of the combination of Grounded-SAM (GSAM) with Cutie (GSAM+Cutie) as a binary endoscopy tool mask annotator. We compare the performance of GSAM+Cutie with similar foundation models assembles like SAM+Cutie, SAM PT [19], and MedSAM [16]+ CoTracker, showing favorable performance for the GSAM+Cutie combination. Finally, we discuss the limitations of the models, and future work that these combinations of foundation models can inspire.

2. Grounded-SAM Meets Cutie Video Object Segmentation

Our proposed binary surgical tool masking process employs two models to generate a video-level all-tools segmentation. First, we generate an initial segmentation mask employing a text-based object segmentation model and an input frame. After this process, this initial frame segmentation feeds a video object segmentation (VOS) model that propagates the mask temporally across the video. Figure 1 presents a graphical overview of the process.

Text-based frame segmentation. While the initial mask can be generated employing image editors or visual-prompt models like SAM, we consider that the tool mask generation process should allow for easy use by users with different backgrounds. In this sense, we found a text-based object segmentation model an ideal solution since it avoids complex image manipulation and allows the user to obtain an initial segmentation just by describing the target object. We evaluated Grounded-SAM [20] for this objective. GSAM combines the text-based prompt object detection capabilities of grounding DINO [13] with the generalizable performance of SAM. The model employs the object detection of Grounding DINO to prompt SAM, leading to a general-purpose foundation model assembled for object segmentation. We employ GSAM to generate an initial frame-level mask. GSAM generates the initial mask considering a text prompt and a reference frame (indicated by the frame index number) from the set of frames of the video. After obtaining the segmentation, we save an image of the reference frame overlapped with the obtained segmentation for visual validation. If the initial segmentation is not appropri-

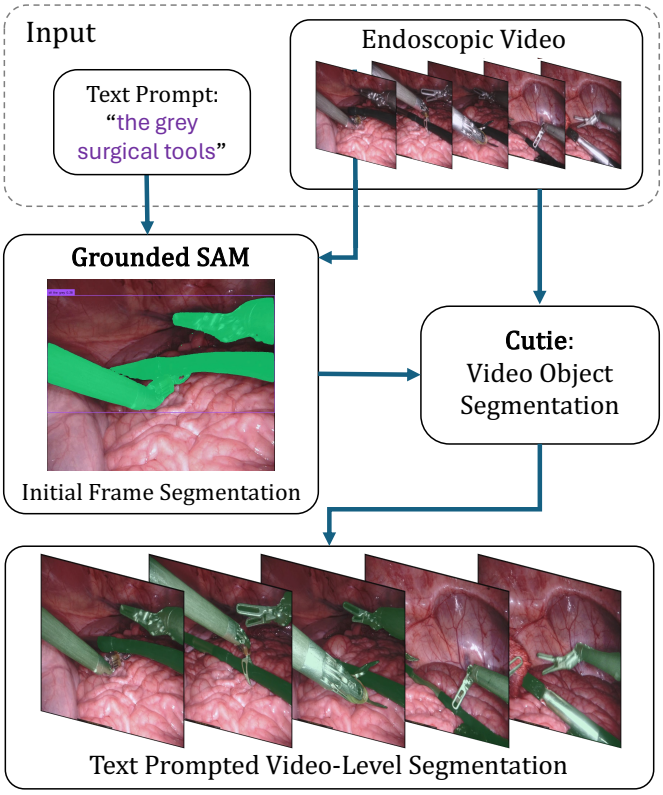

Figure 1. Overview of the text-prompt video tool mask annotator. The model employs Grounding SAM to generate an initial segmentation mask from an endoscopic frame and a text prompt. Then, Cutie VOS uses this mask to propagate the annotation across the video.

ate, the process can be repeated with a different text prompt or frame index. For some cases, a precise description of tool location might be necessary (e.g., " the grey surgical instrument on the skin on the left part of the image.").

Video Object Segmentation. Cutie [4] is a video object segmentation network that tracks and segments objects defined by an initial frame annotation. Cutie incorporates an object transformer and an object-level memory in addition to the pixel-level memory employed by previous VOS models. This allows the model to perform object segmentation even under heavy occlusions. When tested on endoscopy data, Cutie demonstrated strong tool segmentation capabilities in the endoscopic video, with the main disadvantage of requiring an initial segmentation mask. We overcome this limitation employing the results of GSAM. Hence, we added Cutie to the assembly, employing the segmentation masks generated by GSAM to initialize the Cutie VOS model. Then, the model propagates the segmentation across the entire video, leading to a text-based tool segmentation process.

In the remaining sections of this paper, we describe the experimental setup and evaluate the model GSAM+Cutie assembly to assess its performance and generalizability as an out-of-the-box, ready-to-use solution for tool masking in endoscopy video sequences.

3. Experiments and Results

We compare GSAM+Cutie against similar assemblies of foundation models. Particularly Grounded-SAM [20], SAM-PT [19], and SurgicalSAM [21]. First, we compare the quantitative performance of the models in the EndoVis17 and EndoVis18 datasets. Then, we perform a qualitative evaluation of the generalization properties of the models in an in-house sinus endoscopy dataset.

3.1. Datasets

3.1.1 EndoVis Datasets

We quantitatively evaluate our algorithm on the robotic surgical tool segmentation datasets of EndoVis17 [1] and EndoVis18 [2]. These datasets were collected using the da Vinci surgical system in a porcine procedure. We consider the task of binary image segmentation, where the objective is to generate tool masks for each frame of the image to differentiate between all tools and tissue. In our experiment, we employ the validation examples (eight video sequences) of the EndoVis17 dataset, and (four video sequences) of the EndoVis18 dataset to perform the experiments. For EndoVis18, we use the ground-truth annotations and validation set proposed in [5].

3.1.2 Sinus Endoscopy Dataset

We also employed an in-house sinus dataset obtained through simulated sinus surgery on cadaveric specimens performed by an experienced surgeon. We employed two video sequences with 71 and 242 frames, respectively. We qualitatively evaluate these results inspired by the advancements in sinus monocular depth estimation [14] and 3D reconstruction [15] since a tool masking procedure would enable the use of intraoperative data with tools in these algorithms.

3.2. Implementation Details

To analyze the models from a generalizability perspective, we compare all the models employing an out-of-the-box policy with the default weights (no fine-tuning), and only adjusting the prompts, and the inference hyper-parameters. All models run in an Ubuntu 20.04 computer with an NVIDIA Quadro 6000 graphic card.

surgical instruments
surgical instrument on the left. surgical instrument on the top right corner
the grey surgical instrument on the skin on the left part of the image. the surgical instrument on the right part of the image
the grey surgical instrument on the top left. the grey surgical instrument on the top right
the grey surgical instrument on the left. the grey surgical instrument on the right bottom
the grey surgical instrument on the right. the grey surgical instrument on the right bottom

Table 1. List of text prompts used during the experiments.

3.3. Binary Tool Segmentation

The first experiment evaluates the ability of the models to perform video-level surgical tool annotations in endoscopic sequences. In this sense, the tasks is defined as a binary segmentation problem where the segmentation mask created indicated the areas covered by any of the instruments present in the scene. The evaluation employs the open-source EndoVis17 [1] and EndoVis18 [2] datasets. For the models that require a text prompt, we selected the prompt from Table 1 that gives a good initialization considering a visual assessment. We run GSAM in a frame-wise manner, using as much as possible the same text prompt for all the frames. However, if no tools were found by GSAM in a given frame during the process, we selected a new prompt from the list, and/or relax the bounding box threshold (re-initialization) and continue from that frame. We visually verify the selected prompt can segment the instruments before continuing with the remaining frames. Otherwise, a different prompt is selected. We performed a similar re-initialization for the GSAM+Cutie combination. For SAM-PT, we manually prompted four positive instruments points, and two negative tissue/background points to generate the segmentation. Results are presented in Table 2 and Table 3.

The SAM-PT model uses CoTracker, relying on the quality of the tracked points in order to prompt SAM and generate tool segmentation masks. Table 2 and Table 3 show this model presents difficulties to generate the surgical tool masks. We observe difficulties in maintaining point tracking on the surgical tools likely due to challenges in endoscopic video where the reflective nature of the tools cause variable appearance. Additionally, quick tool movements cause a motion blur in certain frames which also changes the tool appearance and cause the points to lose track. When the point tracker fails, either no segmentation is generated or a segmentation of the background tissue is propagated instead. GSAM also presents degraded performance when run for all frames. This can suggest that a single prompt is not enough to generate all the masks of the video. However, using multiple prompts for the different frames of the same video can become inefficient. The use of GSAM with the Cutie VOS model (GSAM+Cutie) is observed to be significantly more robust to these appearance variations as it considers the entire object for tracking. The combination of GSAM+Cutie even outperforms Surgical-SAM, a model fine tuned to the EndoVis datasets. Qualitative examples of the segmentation results are shown in Figure 2.

Method	EndoVis2017	EndoVis2018
SAM-PT [19]	0.21 ± 0.3	0.25 ± 0.3
Surgical-SAM [21]	0.90 ± 0.1	0.86 ± 0.2
GSAM [20]	0.87 ± 0.2	0.69 ± 0.4
GSAM+Cutie (Ours)	0.93 ± 0.1	0.88 ± 0.2

Table 2. Average DICE score $\pm$ std comparison of methods on EndoVis datasets.

Method	EndoVis2017	EndoVis2018
SAM-PT [19]	0.17 ± 0.3	0.19 ± 0.3
Surgical-SAM [21]	0.84 ± 0.2	0.78 ± 0.2
GSAM [20]	0.81 ± 0.2	0.61 ± 0.3
GSAM+Cutie (Ours)	0.88 ± 0.1	0.81 ± 0.2

Table 3. Average IoU $\pm$ std comparison of methods on EndoVis datasets.

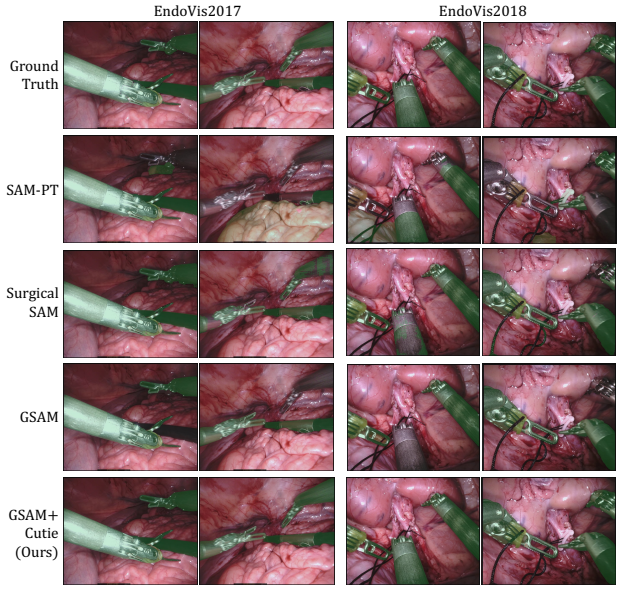

Figure 2. Qualitative segmentation results on EndoVis datasets, where binary segmentations are overlayed on the image in green.

3.4. How Well The Models Can Generalize?

Considering the diversity of endoscopic video sequences, the ability of a tool annotation model to generalize to un-

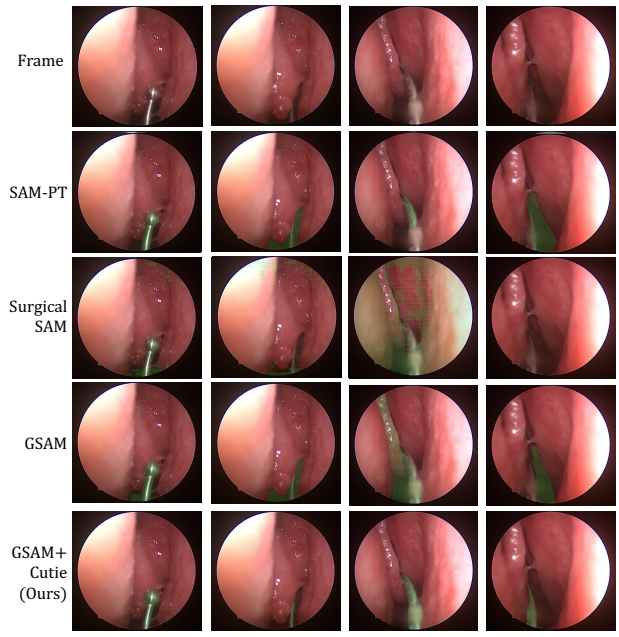

Figure 3. Qualitative segmentation results on in-house sinus endoscopy dataset with two different tools (two examples each), where binary segmentations are overlayed in green.

seen domains without fine-tuning is an important factor to consider. In this experiment, we evaluated the video segmentation abilities of the models in sinus endoscopy. This represents an entirely different scenario compared with the EndoVis sequences. Note that in all our experiments, we avoid fine-tuning the models, for a true zero-shot ready-to-use evaluation. The qualitative results of this comparison are presented in Figure 3. We employed the same setup and configuration as in subsection 3.3 with the main difference on the text prompt employed, which changed to "the grey long surgical tool in the circular endoscopy on the skin". Note that we employed the Surgical-SAM model fine-tuned on EndoVis2017 on our sinus dataset to examine if the model could be applicable to a different endoscopic domain. As an input class prompt is also necessary to identify the tool for segmentation in this model, we used the "Others" prompt as these tools were not directly classified in the original dataset. The qualitative results show that this model struggles to generalize, as some frames produced segmentations on the majority of the scene, while other frames detected no tool present at all when that was not the case.

SAM-PT presents similar problems as with the EndoVis sequences, with the point-tracker diverging when advancing in the video. Finally, similar to our initial experiment, the use of GSAM alone generates inconsistencies in the segmentation across the frames, reinforcing the benefits of integrating a VOS in the masking procedure.

4. Limitations and Future Research Directions

While GSAM+Cutie offers significant advantages in binary surgical tool mask generation, we also acknowledge its limitations for instruments that leave and enter the scene and class-level or instance segmentation. Overall, Cutie presents a level of robustness to occlusions and instruments that leave and re-enter the scene. The performance mainly depends on the time the instrument (or a portion of the instrument) is not visible. In general, new instruments entering the scene will disagree with the initial segmentation mask, requiring re-initialization. While it is possible to address this problem by defining an iterative masking process, the ideal solution should be able to identify when new structures enter the scene. It is worth noticing that this limitation is not exclusive to GSAM+Cutie but to most of the models that require an initial prompt/segmentation to operate. Also, even though a similar iterative re-initialization can be applied to models like SAM-PT, the leakage of the segmentation to the tissue is still a challenge that needs to be addressed (and that is not present in GSAM+Cutie). In any case, we consider that dealing with dynamic surgical videos with objects entering and leaving the scene is an interesting future research direction.

The text prompts allow for adjusting the initial mask using natural language. While this can allow for easy use, different prompts (prompt engineering) might be required for complex sequences. Similarly, our process is primarily proposed for tool masking, and we consider a binary segmentation problem where only a single "tool" class is available. Particularly a second limitation is the tools' instance-level segmentation. Even though our primary objective is binary tool masking, exploring instance-level segmentation is of general interest for surgical applications. In tests with the EndoVis dataset, GSAM+Cutie can generate and track instance-level masks to some extent. It requires a more complex description of the surgical scene, and in some cases, not all instances will be individually detected. In the case of instance-level surgical tool segmentation, this can suggest that a fine-tuned approach is required. However, another approach that could be generalizable to different domains is the direct integration of text prompts to Cutie, which could help propagate the segmentation mask. At the same time, it attends to the semantic information introduced by the text prompts. We consider that such model can lead to a second possible research direction in surgical tool masking.

5. Conclusions

Our work shows that using GSAM+Cutie allows for robust video-level binary tool segmentation without fine-tuning or retraining on domain-specific data. This represents a promising avenue for models that assist in data curation and

data preprocessing in general endoscopic video sequences. Similar combinations of foundation models that integrate SAM and CoTracker are sensitive to the quality of point tracking, as point-tracking inconsistencies cause leakages when prompting SAM at the moment of propagate the segmentations. In contrast, by employing GSAM, we are further able to streamline data annotation by incorporating text-prompt inputs for mask initialization, easing annotation efforts as SAM is sensitive to visual prompting. The initial mask combined with Cutie allows for video-level segmentation, providing more stable results by maintaining object-level representations. Our proposed ensemble of GSAM+Cutie leverages the capacity of foundation models for generalizable and consistent out-of-the box performance on endoscopic data, facilitating reliable data annotation for surgical tool segmentation.

Acknowledgements

This work was funded in part by Johns Hopkins University internal funds and in part by NIH R01EB030511. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019. 2, 3, 4

[2] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020. 3, 4

[3] Nicolás Ayobi, Alejandra Pérez-Rondón, Santiago Rodríguez, and Pablo Arbeláez. Matis: Masked-attention transformers for surgical instrument segmentation. *arXiv preprint arXiv:2303.09514*, 2023. 2

[4] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. *arXiv preprint arXiv:2310.12982*, 2023. 2, 3

[5] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: an instance-based approach for surgical instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–605. Springer, 2020. 3

[6] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 2

[7] Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In *MICCAI, 2019*, 2019. 2

[8] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 2

[9] Benjamin D Killeen, Liam J Wang, Han Zhang, Mehran Armand, Russell H Taylor, Greg Osgood, and Mathias Unberath. Fluorosam: A language-aligned foundation model for x-ray image segmentation. *arXiv preprint arXiv:2403.08059*, 2024. 2

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[11] Eung-Joo Lee, William Plishker, Xinyang Liu, Shuvra S. Bhattacharyya, and Raj Shekhar. Weakly supervised segmentation for real-time surgical tool tracking. *Healthc Technol Lett.*, 2019. 2

[12] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 1

[13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[14] Xingtong Liu, Ayushi Sinha, Masaru Ishii, Gregory D Hager, Austin Reiter, Russell H Taylor, and Mathias Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE transactions on medical imaging*, 39(5):1438–1447, 2019. 3

[15] Xingtong Liu, Maia Stiber, Jindan Huang, Masaru Ishii, Gregory D Hager, Russell H Taylor, and Mathias Unberath. Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 3–13. Springer, 2020. 1, 3

[16] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2

[17] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 2

[18] Kanyifeechukwu J Oguine, Roger D Soberanis-Mukul, Nathan Drenkow, and Mathias Unberath. From generalization to precision: Exploring sam for tool segmentation in surgical environments. *arXiv preprint arXiv:2402.17972*, 2024. 2

[19] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 2, 3, 4

[20] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2, 3, 4

[21] Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalsam: Efficient class promptable surgical instrument segmentation. *arXiv preprint arXiv:2308.08746*, 2023. 2, 3, 4

[22] Zixu Zhao, Yueming Jin, and Pheng-Ann Heng. Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In *ICRA*, 2022. 2

[23] Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. Text promptable surgical instrument segmentation with vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2