FluoroSAM: A Language-promptable Foundation Model for Flexible X-ray Image Segmentation

Benjamin D. Killeen[*1], Liam J. Wang[1], Blanca Iñigo[1], Han Zhang[1], Mehran Armand[2], Russell H. Taylor[1], Greg Osgood[3], and Mathias Unberath[1]

[1] Johns Hopkins University, Baltimore, MD 21218, USA
[2] University of Arkansas, Fayetteville, AR 72701, USA
[3] Johns Hopkins Hospital, Baltimore, MD 21287, USA
*killeen@jhu.edu (corresponding author)

Abstract. Language promptable X-ray image segmentation would enable greater flexibility for human-in-the-loop workflows in diagnostic and interventional precision medicine. Prior efforts have contributed task-specific models capable of solving problems within a narrow scope, but expanding to broader use requires additional data, annotations, and training time. Recently, language-aligned foundation models (LFMs) – machine learning models trained on large amounts of highly variable image and text data thus enabling broad applicability – have emerged as promising tools for automated image analysis. Existing foundation models for medical image analysis focus on scenarios and modalities where large, richly annotated datasets are available. However, the X-ray imaging modality features highly variable image appearance and applications, from diagnostic chest X-rays to interventional fluoroscopy, with varying availability of data. To pave the way toward an LFM for comprehensive and language-aligned analysis of arbitrary medical X-ray images, we introduce FluoroSAM, a language-promptable variant of the Segment-Anything Model, trained from scratch on 3M synthetic X-ray images from a wide variety of human anatomies, imaging geometries, and viewing angles. These include pseudo-ground truth masks for 128 organ types and 464 tools with associated text descriptions. FluoroSAM is capable of segmenting myriad anatomical structures and tools based on natural language prompts, thanks to the novel incorporation of vector quantization (VQ) of text embeddings in the training process. We demonstrate FluoroSAM's performance quantitatively on real X-ray images and showcase on several applications how FluoroSAM is a key enabler for rich human-machine interaction in the X-ray image acquisition and analysis context. *Information on data, weights, and code is available at https://github.com/arcadelab/fluorosam.*

Keywords: radiology · fluoroscopy · medical imaging AI · multimodal foundation model · segment anything · machine learning · deep learning

1 Introduction

X-ray imaging is a workhorse imaging modality for diagnostic and interventional healthcare. There is enormous opportunity for language-promptable, automated
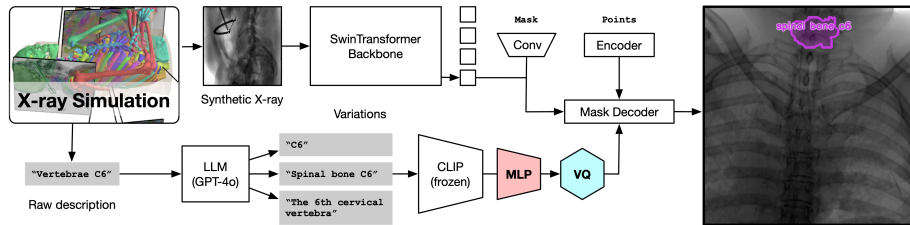
Fig. 1: **FluoroSAM** is trained fully *in silico* with mask and text prompt pairs. It features a VQ layer that enables training language-aligned training on X-ray images.

segmentation of X-ray images to enable human-in-the-loop workflows in precision medicine [2, 3, 7, 11, 12, 24, 25, 28]. Prior efforts have contributed machine learning (ML) techniques that perform well within a narrow scope of X-ray imaging, but their fixed design and limited training data limit the potential across the domain. Extending these techniques to support additional classes or more complex queries requires additional data and personnel effort for refitting and retraining ML models. Recently, foundation models (FMs) – particularly language-aligned FMs (LFMs) – have emerged as a promising direction for building more flexible models [1, 4, 15, 16, 18]. FMs are characterized by scalable training strategies, often accomplished through self-supervision, that enable learning from large, diverse data. LFMs, which incorporating text into the training process alongside images, allow for specification of tasks and classes using natural language. Our goal is to develop a language-promptable FM for X-ray image segmentation, enabling potential downstream applications ranging from interactive diagnostic systems [4, 9] to intelligent human-machine interfaces in image-guided interventions [13, 14].

Promptable segmentation models have become increasingly prominent in recent years, as they are adaptable to both automated and human-in-the-loop workflows. Rather than predict segmentation masks for a fixed number of classes, the Segment-Anything Model (SAM) [15] predicts a semantically meaningful mask for any object in an image based on a prompt, which can be a mask, bounding box, or point. Trained on a large scale dataset of over 1B masks, the original SAM is a powerful tool for automated and interactive image segmentation, and it has been successfully fine-tuned on a variety of medical imaging modalities [18]. Its successor, SAM 2, further lends itself to 3D medical image segmentation by incorporating a memory bank, increasing the efficiency and consistency of segmentation over multiple frames [22]. However, to the extent that SAM and SAM 2 support text prompts, they rely on a previously trained LFM, CLIP [21], to convert image patches into language-aligned prompts. It is not immediately clear how to transfer this approach outside of natural images, on which CLIP was trained. Whereas natural images – and many medical imaging modalities – feature objects with clear boundaries and a nested structure, X-ray images are transmissive by nature, with many overlapping masks belong-
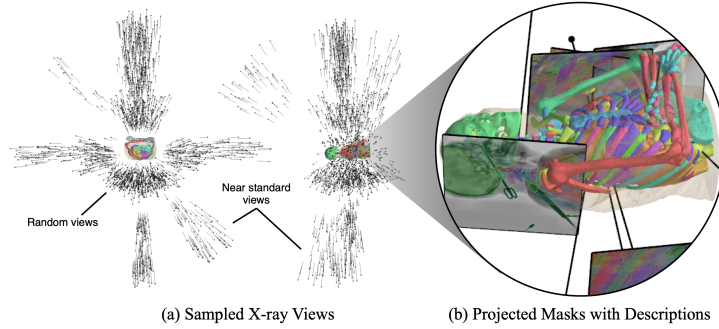
(a) Sampled X-ray Views   (b) Projected Masks with Descriptions

Fig. 2: **Simulation of the FluroSeg dataset.** (a) Virtual C-arm views, shown here as principle rays, are sampled from a range of both random and near-standard views, which are determined automatically. (b) The environment outputs masks and descriptions for organs and tools.

ing to wholly different objects. Even if a CLIP-like model were available for the X-ray imaging modality, a single image patch would contain visual features from multiple objects, and it would be unclear how to distinguish among them. Further, the data available for training LFMs for X-ray imaging has so far been limited to diagnostic chest X-ray, where the imaging geometry is relatively consistent and detailed radiology reports can be retrospectively sourced for training data [2, 4, 8]. By contrast, interventional X-ray features a wide variety of imaging geometries, anatomies, and objects, with little data available for training.

Recent advances in simulation [25], automated CT image segmentation [27], and sim-to-real transfer [7, 12] introduce the possibility of training LFMs for X-ray image segmentation with paired masks and text descriptions. To enable flexible natural language prompting, we use a vector quantization (VQ) module on top of a frozen CLIP encoder to provide a more consistent signal to the mask decoder. In this context, we make two major contributions:

1. The FluoroSeg dataset, a large-scale dataset of 3M synthetic X-ray images, with mask and text pairs for organs and tools. The dataset is generated from a wide variety of human anatomies, imaging geometries, and viewing angles, and it is designed to support training of a language-promptable FM for X-ray image segmentation.

2. FluoroSAM, a language-promptable variant of SAM, trained from scratch on FluoroSeg. FluoroSAM features a novel text encoder that enables generalizable language alignment during training. In this paper, we focus our evaluation on real fluoroscopic images and chest X-rays, for which ground truth segmentations are available.

2   Methods

FluoroSeg Dataset.   FluoroSeg is a large-scale dataset of 3 million simulated X-ray images of the full body, for both diagnostic and interventional exams. Building on prior work [7, 12, 25, 26], the simulation environment takes as input a patient model derived from a real CT scan, along with associated segmentations and descriptions. Computer modeled surgical tools are positioned relative to the patient anatomy. Where FluoroSeg stands out in this space is in the scalability of image and ground truth simulation as well as the number of radiological exams which are replicated in a given environment. By combining volumetric and mesh-based rendering methods for CT, tool, and ground truth projection, the FluoroSeg simulation environment is able to generate numerous images with myriad ground truth segmentations and descriptions in an efficient, scalable manner. To replicate the variability of real-world X-ray imaging, the simulation environment samples from imaging geometries and viewing angles that are both random and near-standard, as determined automatically based on organ meshes.

The simulation pipeline is as follows. We source 1621 high resolution ($0.96 \times 0.96 \times 0.5$ mm) CT scans from the New Mexico Decedent Image Database [5], including scans of the head & neck region, the torso, and the lower extremity. Each scan is segmented into 128 organ classes using TotalSegmentator [27], from which we obtain surface meshes. Depending on the scan type, the simulation environment selects a range of standard views from which to render synthetic X-ray images, including the chest, abdominal, shoulder, clavicle, humerus, elbow, forearm, hand, pelvis, femur, sacroiliac joint, knee, tibia/fibula, ankle, foot, skull, and spine series as defined by [19]. Each view is approximated with minor random variations based on the meshes present, so as to capture a variety of imaging geometries for each exam, suited to both diagnostic and interventional applications. In addition, fully random C-arm views are sampled by selecting a primary organ to focus on and then randomly selecting a viewing angle to within 60°of the anterior direction. This strategy is reflected in Fig. 2, in which the sampled source positions and principle ray directions can be seen clustered around near-standard views as well as distributed randomly. For each image, we select a random subset out of 464 tools to include in the image. These include 111 models from [17], 296 from GrabCAD, and 57 modeled internally. Each tool is associated with a comprehensive text description, *e.g.* "cannulated 110mm screw." They are placed along the field of view with random location and orientation. The number of images sampled per CT varies based on the number of standard views which are possible to determine for the given anatomy. Images are rendered at a resolution of 512^2 pixels alongside masks and text descriptions. Using an NVIDIA A6000 GPU, the simulation environment is able to generate 6.5 ± 15.7 images per second, depending on the exam, with a total of 2.95M images generated in about 6 GPU days. The final images are split into training and validation sets based on a 90/10% split of the base CT scans.

FluoroSAM: An LFM for X-ray Image Segmentation.   FluoroSAM is a language-promptable variant of SAM [15] with support for text, point, and mask prompts, as shown in Fig. 1. When using only a single prompt, SAM predicts
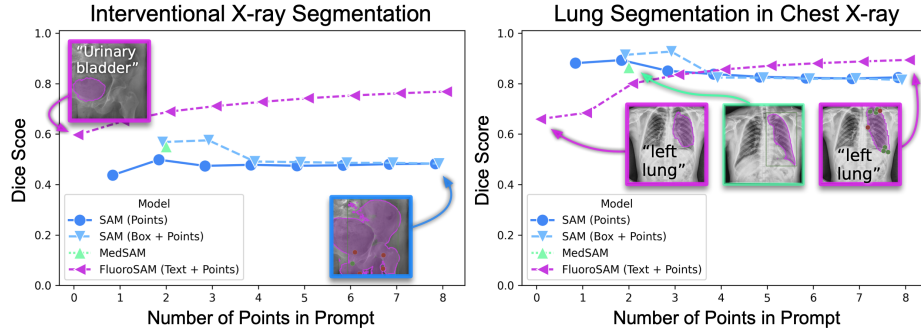
Fig. 3: **Quantitative results**. (a) On interventional X-ray images, FluoroSAM outperforms its peers even with text-only prompting. (b) On CXRs, FluoroSAM adapts to hand-annotated lung segmentations despite being trained on synthetic data. MedSAM [18] includes this task in its training data.

three segmentation maps, often corresponding to the whole, part, or sub-part of an object, backpropagating only through the branch with the lowest loss. X-ray images, on the other hand, often contain overlapping projections of many anatomical and non-anatomical objects. MedSAM [18] mitigates this ambiguity for other medical imaging domains by only allowing for bounding box prompts. For X-ray imaging, even if MedSAM were trained from scratch using the FluoroSeg dataset, this approach is undesirable because it (a) still features significant ambiguity, (b) makes automatic or even non-expert prompting impractical, and (c) does not lend greater flexibility for human-in-the-loop systems. Language prompts, on the other hand, enable greater flexibility for both automated and human-in-the-loop systems, while also reducing ambiguity in prompts.

With access to the text descriptions in the FluoroSeg dataset, it is possible to train FluoroSAM from scratch with text prompting, using a novel text encoder. The text encoder consists of a frozen CLIP encoder [21], followed by a multi-layer perceptron (MLP) with 2 hidden layers and a VQ bottleneck [20], which outputs the prompt token. While VQ theoretically limits the generalizability of our framework to new segmentation classes, it expands the generalizability to new language prompts by reducing the variability from disparate descriptions for the same object. During training, we use `gpt-4o` [1] to perform text augmentation [6] on the original descriptions, generating up to 30 comprehensive and non-comprehensive prompts for each mask. In addition, we procedurally combine masks from related organs in 38 groups, such as the "left ribs" or "cervical vertebrae," which are separate classes in the base TotalSegmentator, and likewise augment their descriptions. Finally, with small probability, we sample text prompts which are not present in the image, so that FluoroSAM ignores bad prompts. Because text prompts themselves may be ambiguous, we leverage the multi-output capabilities of SAM to predict multiple masks for each prompt,

Table 1: **Real X-ray results.** (*) indicates that a few (< 10) false negatives were removed to compute HDD. HDD is in mm for cadaver, pixels for CXRs.

Dataset	Model	Text Only			2 Points or Box			8 Points		
		IoU↑	Dice↑	HDD↓	IoU↑	Dice↑	HDD↓	IoU↑	Dice↑	HDD↓
Cadaver	SAM			—	0.36	0.50	139.1*	0.35	0.48	152.6*
	SAM (Box)			—	<u>0.42</u>	<u>0.57</u>	86.6*	<u>0.35</u>	<u>0.48</u>	<u>143.9</u>*
	MedSAM			—	0.41	0.55	**76.0**			—
	FluoroSAM	**0.47**	**0.60**	102.8*	**0.56**	**0.69**	<u>81.0</u>	**0.64**	**0.77**	**60.8**
CXR	SAM			—	<u>0.83</u>	<u>0.89</u>	153.9	<u>0.73</u>	<u>0.82</u>	406.8
	SAM (Box)			—	**0.85**	**0.91**	87.6	0.73	0.81	<u>303.9</u>
	MedSAM			—	0.34	0.48	<u>81.9</u>			—
	FluoroSAM	**0.50**	**0.66**	**90.1**	0.67	0.80	**53.9**	**0.81**	**0.89**	**31.9**

and we select the mask with the lowest loss during training. During inference, the IOU prediction head enables selection of the best mask for each prompt.

The image encoder consists of a SwinTransformer backbone, which is better suited to the simulated image size than the original ViT backbone, pre-trained on ImageNet-22k, with additional pre-training for instance segmentation on a reduced set of FluoroSeg classes. We use the Swin-S variant for ablation studies and Swin-L for the final model, with an input image size of 448^2. Following [7, 12], we apply strong domain randomization of the image during training to facilitate sim-to-real transfer, including coarse dropout, inversion, blurring, Gaussian contrast adjustment [10], random windowing, and CLAHE histogram equalization. To accommodate the wide variety of intensity values in X-ray images, we used a 3-channel input, where the second and third channel window and level are determined by a K-Means clustering of the pixel values.

During training, for each mask, we run a total of 9 prompts, with 1 text prompt followed by 8 point prompts randomly selected from the incorrect mask region. As in [15], we use each previous predicted mask as an additional mask prompt. We use Dice and focal loss for masks, re-weighting the contributions from text-only prompting to equal that of all point prompts. Only the losses from text, point 1, point 8, and an additional random point are backpropagated. We train FluoroSAM for 10 epochs with a base learning rate of 8×10^{-4} after a linear warm-up of 20k iterations from 8×10^{-6}, reduced by a factor of 10 at step 200k and 400k. We use a batch size of 16 across 2 NVIDIA H100 GPUs, with a total training time of 6 days.

3 Evaluation

Synthetic X-ray images. Evaluated on the FluoroSeg dataset, FluoroSAM achieves a Dice score of 0.59 based solely on text prompting, increasing to 0.70
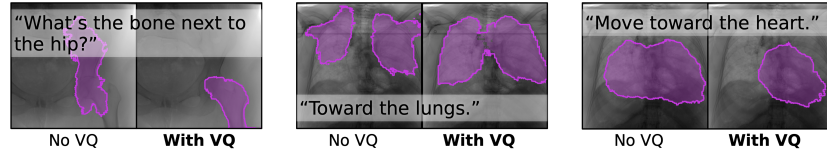
Fig. 4: In a limited user study, we observe qualitative results consistent with the hypothesis that VQ improves segmentation robustness to variable text prompts. For example, FluoroSAM with VQ was able to correctly segment the femur, answering the question "What's the bone next to the hip?"

with 2 points and 0.79 with 8 points. This is a promising result given the challenge of full body X-ray segmentation based on text prompts.

Real Interventional X-ray images.   To evaluate FluoroSAM's performance on real interventional X-ray iamges, we collect a dataset of registered X-ray images with a full torso specimen using a Brainlab Loop-X device. The specimen consisted of a torso section from mid-femur to T2, excluding the arms, obtained from a 60-year-old female donor with a living height of 157 cm and weight of 50 kg. Before the study, the specimen was thawed at 4°C for six days. All fluoroscopic images were acquired with navigation relative to a fixed patient array. To generate complete ground truth masks, we stitched together four navigated cone-beam CT images acquired with the Loop-X immediately after the study and projected organ segmentations [27] onto each image. We observe some spurious masks in the ground truth, possibly due to decomposition, so we limit our evaluation to masks which are at least 2.5% of the image size. In total, the dataset has 1,741 masks suitable for evaluation.

We find that FluoroSAM effectively segments structures in real X-ray images from a variety of views. With text prompts, FluoroSAM achieves a mean Intersection over Union (IoU) of 0.47 and a Dice score of 0.60, as shown in Table 1. This includes hard prompts like "fifth right rib" and is sufficient for many downstream applications in interventional imaging [14]. Additionally, although VQ result in marginal improvement for text prompts similar to the training set, we observe qualitative benefits to mask quality for unusual prompts, as seen in Fig. 4. Fig. 5a shows additional examples with text-only prompting. With 2 points, FluoroSAM outperforms SAM [15] (prompted with points or boxes) and MedSAM [18] in terms of IoU and Dice score, achieving 0.56 and 0.69, respectively. In terms of HDD, MedSAM holds a slight advantage over FluoroSAM, possibly because bounding box prompts tightly constrain the mask, whereas point prompts are more ambiguous. As indicated by the lower IoU and Dice, MedSAM generally fails to reflect the underlying anatomy and simply fills the provided box. Moreover, as can be seen in Fig. 3a, point prompting improves the performance of FluoroSAM, whereas SAM predicts increasingly erroneous masks.

Zero-shot evaluation on CXR.   To show FluoroSAM's potential for general diagnostic X-ray image segmentation, we evaluate it on a whole lung segmenta-

(a) Additional Results with Text-only Prompting

(b) Efficient X-ray Annotation

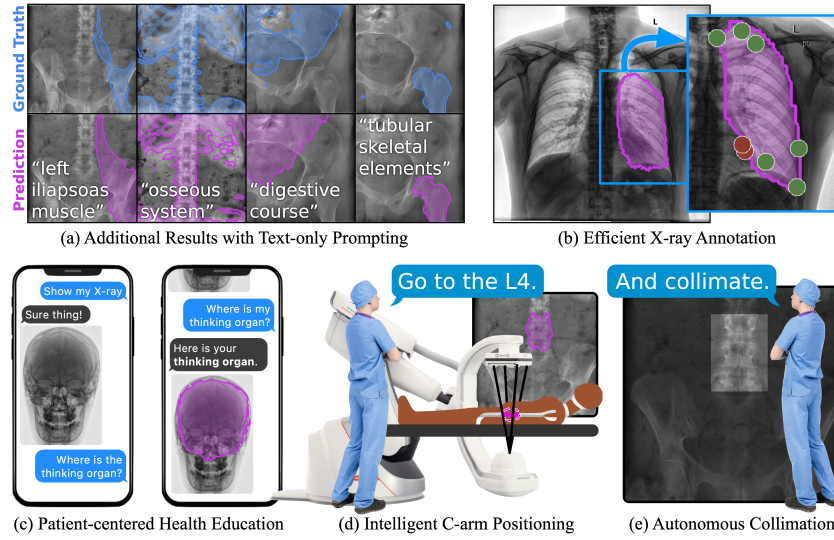(c) Patient-centered Health Education   (d) Intelligent C-arm Positioning   (e) Autonomous Collimation

Fig. 5: **Qualitative results and downstream applications.** (a) Additional results on real X-ray images, with text prompts not used during training. The flexibility of text prompting supports a wide variety of downstream applications. (b) Efficient annotation of X-rays can reduce clinical burden and accelerate real data annotation. (c) Flexible text-based prompting may lend itself to patient-facing education, empowering patients to better understand their own anatomy. In the OR, FluoroSAM can be integrated with robotic C-arms to deliver intelligent positioning (d) and autonomous collimation (e), reducing radiation [14]. These figures use real FluoroSAM predictions on real radiographs.

tion dataset with 1,131 CXRs [2]. Using only text prompts, FluoroSAM provides a reasonable mask of each lung, achieving a mean IoU of 0.50 and a Dice score of 0.66 (Table 1). This is despite training on a dataset of synthetic images with projected masks, which differ systematically from hand-annotated masks [2].

4   Discussion and Conclusion

Thanks to their flexibility, LFMs potentially support a wide variety of autonomous and human-in-the-loop downstream applications. As shown in Fig. 5, FluoroSAM's text and point-prompting capabilities lend themselves to multiple areas within a patient's healthcare journey, from efficient annotation of diagnostic images to accessible tele-health tools, equipped to interpret natural language while interacting with exams. In fact, [14] demonstrates how to use FluoroSAM to equip robotic C-arms with autonomous positioning and collimation capabilities, demonstrating the utility of FluoroSAM's current performance level. A mature FluoroSAM may enable chain-of-thought-based X-ray image analysis related to anatomy in the image [9]. Here, we have demonstrated FluoroSAM's

ability to segment a wide range of anatomical and non-anatomical objects in synthetic, interventional, and chest X-rays. This is a challenging task, as many organs are not plainly visible in X-ray images, which often contain overlapping projections of multiple objects. Yet despite relying solely on synthetic data for training, our model achieves comparable performance with foundation models trained on large, annotated datasets including real CXR images.

Although the approach here has already yielded promising results, there are some notable limitations and immediate opportunities for future work. First, while the FluoroSeg dataset is large, it is still limited in terms of the number of CTs and the variety of anatomical structures represented. Using the FluoroSeg simulation pipeline, incorporating additional CT images is computationally intensive but highly scalable using a fully automated pipeline. Additionally, greater variety could be achieved by incorporating masks from recently released TotalSegmentator tasks or more specialized datasets, *e.g.* fracture fragments [23]. Expanding support in this way may require changes in the FluoroSAM architecture. Although vector quantization of text prompt embeddings qualitatively improves the robustness of FluoroSAM to variable text prompts (see Fig. 4), the VQ codebook size constrains the representational power of text embeddings. This is an acceptable tradeoff in the current context, but other strategies may be required to support a wider range of anatomical structures while maintaining robustness to variable text prompts. Ultimately, with the ability to interpret a wide range of prompts, we envision FluoroSAM will become a key enabler for human-machine interaction in the X-ray domain, opening up new avenues in both diagnostic and interventional imaging.

Acknowledgments

Disclosure of Interests

The authors have no competing interests.

Bibliography

[1] Brown, T.B., et al.: Language Models are Few-Shot Learners. arXiv (May 2020). https://doi.org/10.48550/arXiv.2005.14165

[2] Candemir, S., et al.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans. Med. Imaging **33**(2), 577–590 (Feb 2014). https://doi.org/10.1109/TMI.2013.2290491

[3] Çallı, E., et al.: Deep learning for chest X-ray analysis: A survey. Med. Image Anal. **72**, 102125 (Aug 2021). https://doi.org/10.1016/j.media.2021.102125

[4] Chen, Z., et al.: CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv (Jan 2024). https://doi.org/10.48550/arXiv.2401.12208

[5] Edgar, H., et al.: New mexico decedent image database. Office of the Medical Investigator, University of New Mexico (2020). https://doi.org/10.25827/5s8c-n515

[6] Fan, L., et al.: Improving CLIP Training with Language Rewrites. arXiv (May 2023). https://doi.org/10.48550/arXiv.2305.20088

[7] Gao, C., et al.: Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. Nat. Mach. Intell. **5**(3), 294–308 (Mar 2023). https://doi.org/10.1038/s42256-023-00629-1

[8] Irvin, J., et al.: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv (Jan 2019). https://doi.org/10.48550/arXiv.1901.07031

[9] Jiang, Y., et al.: CoMT: Chain-of-Medical-Thought Reduces Hallucination in Medical Report Generation. arXiv (Jun 2024). https://doi.org/10.48550/arXiv.2406.11451

[10] Killeen, B.D., et al.: An autonomous X-ray image acquisition and interpretation system for assisting percutaneous pelvic fracture fixation. Int. J. CARS pp. 1–8 (May 2023). https://doi.org/10.1007/s11548-023-02941-y

[11] Killeen, B.D., et al.: In silico simulation: a key enabling technology for next-generation intelligent surgical systems. Prog. Biomed. Eng. **5**(3), 032001 (May 2023). https://doi.org/10.1088/2516-1091/acd28b

[12] Killeen, B.D., et al.: Pelphix: Surgical Phase Recognition from X-Ray Images in Percutaneous Pelvic Fixation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, pp. 133–143. Springer, Cham, Switzerland (Oct 2023). https://doi.org/10.1007/978-3-031-43996-4_13

[13] Killeen, B.D., et al.: Take a shot! Natural language control of intelligent robotic X-ray systems in surgery. Int. J. CARS **19**(6), 1165–1173 (Jun 2024). https://doi.org/10.1007/s11548-024-03120-3

[14] Killeen, B.D., et al.: Intelligent control of robotic X-ray devices using a language-promptable digital twin. Int. J. CARS **20**(6), 1125–1134 (Jun 2025). https://doi.org/10.1007/s11548-025-03351-y

[15] Kirillov, A., et al.: Segment Anything. arXiv (Apr 2023). https://doi.org/10.48550/arXiv.2304.02643

[16] Liu, X., et al.: Efficientvit: Memory efficient vision transformer with cascaded group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14420–14430 (June 2023)

[17] Luijten, G., et al.: 3D surgical instrument collection for computer vision and extended reality. Sci. Data **10**(796), 1–12 (Nov 2023). https://doi.org/10.1038/s41597-023-02684-0

[18] Ma, J., et al.: Segment anything in medical images. Nat. Commun. **15**(654), 1–9 (Jan 2024). https://doi.org/10.1038/s41467-024-44824-z

[19] Murphy, A., et al.: General radiography curriculum. Radiopaedia.org (2020). https://doi.org/10.53347/rID-48992, https://doi.org/10.53347/rID-48992, reference article, Accessed on 22 Feb 2025

[20] Oord, A.v.d., et al.: Neural Discrete Representation Learning. arXiv (Nov 2017). https://doi.org/10.48550/arXiv.1711.00937

[21] Radford, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. arXiv (Feb 2021). https://doi.org/10.48550/arXiv.2103.00020

[22] Ravi, N., et al.: SAM 2: Segment Anything in Images and Videos. arXiv (Aug 2024). https://doi.org/10.48550/arXiv.2408.00714

[23] Sang, Y., et al.: Benchmark of Segmentation Techniques for Pelvic Fracture in CT and X-ray: Summary of the PENGWIN 2024 Challenge. arXiv (Apr 2025). https://doi.org/10.48550/arXiv.2504.02382

[24] Silva, G., et al.: Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives. Expert Syst. Appl. **107**, 15–31 (Oct 2018). https://doi.org/10.1016/j.eswa.2018.04.001

[25] Unberath, M., et al.: DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pp. 98–106. Springer, Cham, Switzerland (Sep 2018). https://doi.org/10.1007/978-3-030-00937-3_12

[26] Unberath, M., et al.: Enabling machine learning in X-ray-based procedures via realistic simulation of image formation. Int. J. CARS **14**(9), 1517–1528 (Sep 2019). https://doi.org/10.1007/s11548-019-02011-2

[27] Wasserthal, J., et al.: TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence (Jul 2023), https://pubs.rsna.org/doi/10.1148/ryai.230024

[28] Zhao, W., et al.: Artificial intelligence in image-guided radiotherapy: a review of treatment target localization. Quantitative Imaging in Medicine and Surgery **11**(12), 4881 (Dec 2021). https://doi.org/10.21037/qims-21-199