AffordTissue: Dense Affordance Prediction for Tool-Action Specific Tissue Interaction

Aiza Maksutova[1,⋆], Lalithkumar Seenivasan[1,⋆], Hao Ding[1,⋆], Jiru Xu[1], Chenhao Yu[1], Chenyan Jing[1], Yiqing Shen[1], Mathias Unberath[1]

Johns Hopkins University, Baltimore MD, USA
{unberath}@jhu.edu

Abstract. Surgical action automation has progressed rapidly toward achieving surgeon-like dexterous control, driven primarily by advances in learning from demonstration and vision-language-action models. While these have demonstrated success in table-top experiments, translating them to clinical deployment remains challenging: current methods offer limited predictability on where instruments will interact on tissue surfaces and lack explicit conditioning inputs to enforce tool-action specific safe interaction regions. Addressing this gap, we introduce AffordTissue, a multimodal framework for predicting tool-action specific tissue affordance regions as dense heatmaps during cholecystectomy. Our approach combines a temporal vision encoder capturing tool motion and tissue dynamics across multiple viewpoints, language conditioning enabling generalization across diverse instrument-action pairs, and a DiT-style decoder for dense affordance prediction. We establish the first tissue affordance benchmark by curating and annotating 15,638 video clips across 103 cholecystectomy procedures, covering six unique tool-action pairs involving four instruments (hook, grasper, clipper, scissors) and their associated tasks: dissection, grasping, clipping, and cutting. Quantitatively, our task-specific architecture achieves an average symmetric surface distance (ASSD) of 20.6 px, substantially outperforming vision/vision-language baseline models (U-Net: 44.3 px and Molmo-VLM: 60.2 px) in dense surgical affordance prediction. By predicting tool-action specific tissue affordance regions, AffordTissue provides explicit spatial reasoning for safe surgical automation, potentially unlocking explicit policy guidance toward appropriate tissue regions and early safe stop when instruments deviate outside predicted safe zones.

Keywords: Affordance Prediction · Tissue Affordance · Surgical Automation · Surgical Scene Understanding · Multimodal Models

1 Introduction

Recent advances in learning from demonstration [6], imitation policies [33, 24], and vision-language-action models (VLAs) [12, 5] have significantly accelerated

⋆ Equal contribution.

surgical task automation, with pioneering works demonstrating surgeon-like dexterous control in controlled table-top experiments [11, 23]. However, with safety taking precedence over efficiency in clinical practice, truly translating these advances to clinical deployment hinges on ensuring safe automation [4, 9]. A fundamental limitation of current learning-based approaches lies in their "black box" nature: while these models learn complex dexterous manipulation from expert demonstrations, they offer limited predictability regarding *where* and *how* the robot will interact with tissue, and *whether* the action will succeed [4, 9]. Furthermore, they provide no explicit mechanism to condition or verify safe interaction. This poses significant safety concerns – there is limited controllability in enforcing action-specific tissue interaction zones or intervening before potentially harmful contact occurs.

Current surgical scene understanding approaches rely predominantly on semantic segmentation [17, 2], which can identify anatomical structures but lacks *interaction-aware* perception-reasoning about *where within* those structures a specific tool can safely engage for a given action. Affordance prediction has emerged as a powerful paradigm for interaction-aware perception in robotic manipulation, enabling systems to identify and enforce *where* and *how* objects can be acted upon [29, 19]. However, these methods predominantly target rigid-object interaction, with limited exploration of deformable tissue dynamics, context-dependent surgical constraints, and safety-critical medical requirements.

To this end, we introduce **AffordTissue**, a multimodal framework for predicting tissue affordance regions as dense heatmaps conditioned on tool-action specifications. Similar to costmaps in robot navigation [16], our affordance representation encodes interaction suitability with maximum affordance at the region center that decreases toward boundaries. This potentially unlocks two complementary modes for safe automation: (i) a conditioning signal, guiding learned policies toward appropriate tissue regions, and (ii) a safety layer, triggering automated early robot stopping when instrument trajectories deviate toward tissue outside the predicted affordance region – *before* potentially harmful contact occurs. **Our key contributions are**: (i) we introduce dense tissue affordance prediction as a novel task that provides explicit spatial reasoning about tool-action specific affordance regions on tissue surfaces; (ii) we propose a multimodal architecture producing tool-action conditioned affordance heatmaps toward action guidance and safety verification; and (iii) we curate and annotate 103 cholecystectomy cases, establishing the first tissue affordance benchmark.

2 AffordTissue

We define tool-tissue affordance as the optimal and safe contact area on a tissue surface for a specific tool-action pair. Our proposed framework for predicting affordance reformulates image diffusion into a novel dense heatmap prediction task (Fig. 1). It takes in two inputs: (i) a language prompt specifying the tool–action pair, the surgical context, and the prediction objective, and (ii) a video sequence consisting of the target (t_0) and past frames ($t_{-256} \rightarrow t_{-1}$). The architecture

includes (a) a language encoder that embeds the text prompt, (b) a temporal video transformer that encodes spatiotemporal visual information, and (c) a multimodal diffusion decoder that fuses both to produce a task-aware dense tool-tissue interaction heatmap for the target frame. We begin by detailing the pre-trained backbones and architectural components integrated into our pipeline. Afterwards, we formalize the input space and the end-to-end workflow.

2.1 Preliminaries

(a) SigLIP2 [25]: Built on SigLIP [31], a CLIP-style model, this architecture replaces CLIP's conventional softmax with a **pairwise sigmoid loss**, which eliminates batch-normalization dependencies and boosts training efficiency. SigLIP2 introduces captioning-based pretraining and self-supervised losses, significantly enhancing dense feature representation for segmentation and localization tasks.
(b) Video Swin Transformer [14]: This hierarchical backbone extends the 2D Swin Transformer [13] architecture to the spatiotemporal domain. The model processes an input video by partitioning it into 3D patch tokens. It employs a *3D Shifted Window* mechanism to compute self-attention locally within non-overlapping spatiotemporal windows. To enable cross-window and cross-frame information exchange, the window partitions are shifted along the T, H, and W axes in successive layers, allowing the model to capture multi-frame temporal dependencies.
(c) Adaptive Layer Normalization (AdaLN) Decoder: Adapted from DiT [21], this decoder injects conditional information directly into the transformer's normalization layers. Unlike standard cross-attention, which computes token-wise pairwise similarities, AdaLN performs global conditioning by regressing Layer Normalization scale (γ) and shift (β) parameters directly from the input condition. By dynamically scaling and shifting activations based on the task, AdaLN enables efficient embedding fusion that highlights target spatial regions while maintaining low computational overhead.

2.2 Pipeline

(a) Text input: With multiple instruments and actions present in same surgical procedure, text conditioning is crucial for model scalability, enabling the network to differentiate between them. Following VLM best practices, we structure the text input using a template prompt that specifies surgery type, instrument type, and the prediction task (Fig. 1). Specifically, the prompt encodes a standardized surgical triplet – *surgery type, instrument type, action type* – into a direct command: *'You are given a surgical video frame from a {surgery type}. Predict the tissue affordance region for the given {action type} using the {instrument type}.'*
(b) Visual input: We input a temporal window of $N = 256$ frames (stride of 8) to capture tool motion and tissue dynamics from multiple viewpoints, enabling implicit modeling of deformation patterns and surgical intent beyond single-frame approaches.

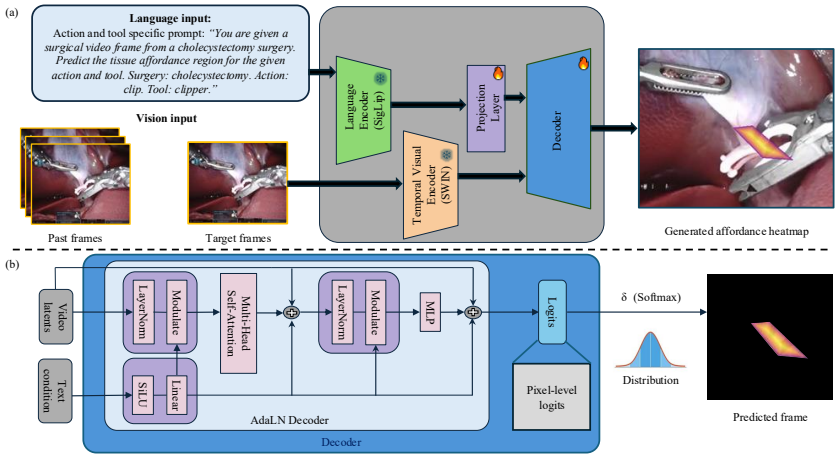

Fig. 1. Architecture of AffordTissue: a) Pipeline overview: An AdaLN-based decoder fuses tool-action prompts and temporal vision context from SigLIP2 and Swin encoders to predict an affordance heatmap. b) Decoder details: The decoder applies text-conditioned adaptive layer normalization (AdaLN) to refine temporal vision latents for a specific tool-action pair, generating pixel-level probabilities that map safe tissue affordance regions.

(c) Decoder adaptation: We adapt the AdaLN decoder for dense heatmap prediction, as detailed in Fig. 1(b). While the original DiT uses AdaLN to predict diffusion noise from latents conditioned on class labels, our architecture processes temporal vision embeddings conditioned on language inputs to predict per-pixel logits. These logits estimate the probability that each pixel belongs to the tissue affordance region. This demonstrates that AdaLN conditioning can be effectively repurposed from generative diffusion modeling to spatially grounded dense prediction.

(d) Workflow: Our end-to-end pipeline is shown in Fig. 1(a). Given a text prompt, the SigLIP 2 encoder produces a $(B, 1152)$ embedding, which an MLP projects into a shared space. Concurrently, the Video Swin Transformer processes input frames (B, C, T, H, W) to extract spatiotemporal features (B, C, H, W). The AdaLN decoder then fuses these representations to output a probabilistic affordance heatmap.

3 Experiment

(i) Dataset: We curate a custom dataset of 15638 video clips from 103 cases: Youtube (21 cases), Cholec-80 (34 cases) [26], HeiChole (11 cases) [28], Comprehensive Robotic Cholecystectomy Dataset (8 cases) [20], and SurgVU (29 cases) [32]. Table 1 details the distribution of video clips across datasets and train/validation/test splits. Here, a *case* denotes a full recording from a unique

Table 1. Distribution of video clips per tool–action pair across datasets. Each entry describes the total number of clips, with (train/val/test) split shown in parentheses.

Tool-action pair	Dataset type				
	Youtube (21 cases)	**Cholec80** (34 cases)	**HeiChole** (11 cases)	**CHEC** (8 cases)	**SurgVU** (29 cases)
Dissect - Hook	1280 (1066/41/173)	4726 (3454/741/531)	1106 (875/97/134)	628 (476/0/152)	2685 (2120/270/295)
Dissect - Grasper	493 (462/0/31)	43 (36/0/7)	101 (92/5/4)	0	295 (234/61/1)
Dissect - Scissors	209 (209/0/0)	12 (12/0/0)	100 (97/3/0)	0	350 (350/0/0)
Grasp - Grasper	584 (538/3/43)	931 (677/189/65)	321 (286/9/26)	136 (97/0/39)	799 (689/66/44)
Clip - Clipper	125 (111/4/10)	185 (143/26/16)	87 (73/6/8)	0	99 (76/14/9)
Cut - Scissors	69 (63/0/6)	145 (106/20/19)	79 (70/4/5)	0	49 (46/3/0)

operation, while a *clip* represents a short temporal window capturing a single tool-action pair within a case. To prevent data leakage, dataset splitting is enforced strictly at the case level. Because each case comes from an independent procedure with unique conditions (e.g., tissue features, light conditions), the resulting dataset diversity ensures a reliable evaluation of model generalization. Each video clip is annotated for tool-action pairs (language) and tissue affordance. To define target regions, each case was manually annotated with 4 action-specific keypoints outlining the safe tool–tissue interaction zone. We used a two-stage pipeline where initial annotations by medical and engineering students were audited by the core research team and clinicians. To minimize inter-annotator variance, semi-automatic point-tracking tools anchored annotations to anatomical landmarks, filtering out instrument-induced drift. Although labeled as uniform polygons, affordance regions are modeled as Gaussian distributions for dense prediction, where boundary distance represents a transition toward non-affordance. This gradient is a modeling choice rather than a reflection of varying safety levels; thus, the task is satisfied if the model recovers the total area, regardless of intensity deviations from the Gaussian prior. Crucially, to predict affordance prior to instrument interaction, only frames preceding the onset of the surgical action are utilized.

(ii) Training and inference: The language and vision encoders are frozen during training, with optimization restricted to the decoder parameters. The model is trained on a single NVIDIA A100 GPU for 100 epochs, using a combination of binary cross-entropy loss and soft intersection-over-union (IoU) loss. Optimization is performed using AdamW [15] with an initial learning rate of 1×10^{-4} and a cosine learning rate scheduler. During training, target frames are randomly sampled within the pre-action window. The model inputs the target frame alongside 256 past frames (stride of 8), providing approximately 10.6 seconds of temporal context to capture tool-tissue dynamics.

(iii) Evaluation Metrics: Evaluation is split into logits- and boundary- conditioned metrics. *Logits-conditioned* metrics assess probabilistic overlap of target and predicted heatmaps using a "soft" Dice score [18] computed directly on continuous logits. *Boundary-conditioned* metrics evaluate spatial alignment, and include Precision at K (PCK) [3], Hausdorff Distance (HD) in pixels, and Average Symmetric Surface Distance (ASSD) in pixels [8]. Performance is evaluated across eight pre-action frames, with frames sampled well before the action occurs.

Table 2. Comparison of our model against baselines: U-Net [22], SAM3 [1], Molmo-VLM [7], and Qwen-VLM [30].

Metrics	DICE↑	PCK@0.05↑	PCK@0.1↑	HD(px)↓	ASSD(px)↓
Ours	0.124	**0.517**	**0.667**	**79.763**	**20.557**
U-Net [22]	0.106	0.444	0.619	98.243	44.390
SAM3 [1]	**0.18**	0.128	0.221	150.320	81.138
Molmo-VLM [7]	0.026	0.095	0.320	129.494	60.184
Qwen-VLM (8B) [30]	0.014	0.031	0.022	203.214	111.271

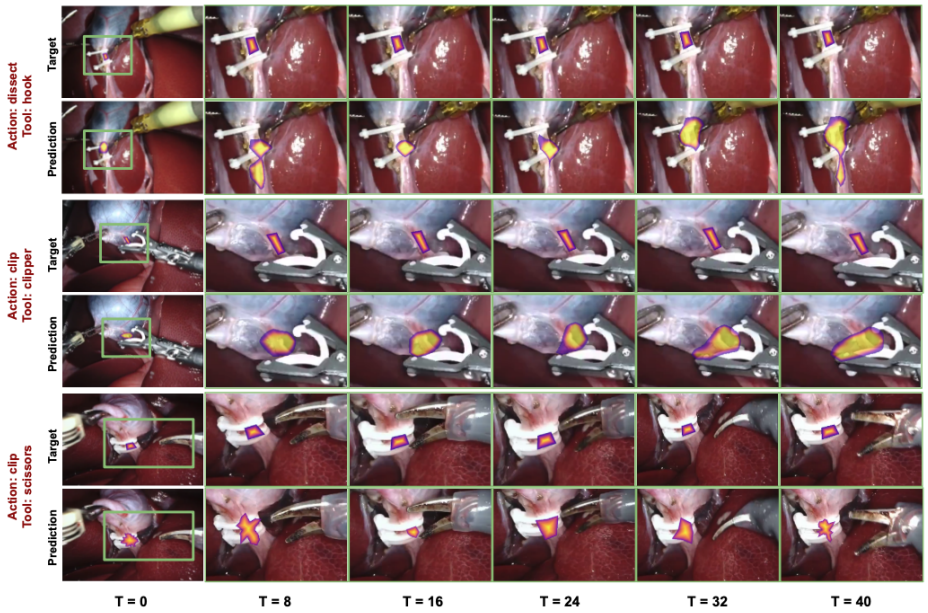

Fig. 2. Qualitative comparison between ground truth and predicted heatmaps for three tool–action pairs: (i) hook - dissection, (ii) clipper - clipping, and (iii) scissors -clipping, across six representative timestamps.

This protocol ensures the model relies on tissue context rather than grounding its predictions on the instrument's position as it approaches the affordance region.

4 Results and Ablation Study

(i) Baseline Comparison: We quantitatively evaluate our pipeline's performance in detecting safe tool-tissue interaction against the following baseline models: U-Net [22], SAM3 [1], Molmo-VLM [7] and QWEN-8B VLM [30]. From Table 2 we observe that the model achieves an **ASSD of 20.557 px** and **PCK@0.05 and PCK@0.1 of 0.517 and 0.667**, respectively. This strong spatial alignment between predicted and ground-truth heatmaps observed quantitatively is supported by a qualitative analysis in Fig. 2. In contrast, the Haus-

forff Distance (HD) performance is relatively high. Qualitative analysis showed that it is primarily caused by occasional secondary, low-confidence heatmap activations that appear on the tool surface. We suggest that in future studies we add a carefully designed maximum-distance penalty to solve this problem. The DICE score is comparatively low across cases. As illustrated in Fig. 2, this is largely due to differences in intensity distributions rather than boundary misalignment. While baseline methods like SAM3 produce binary outputs that are post-hoc fitted to a Gaussian to artificially match the target distribution, our model predicts continuous probability values directly. Consequently, although our model accurately localizes the affordance areas – as evidenced by low ASSD and HD metrics – it receives a lower Dice score because it does not perfectly replicate the idealized Gaussian prior. We argue that this intensity mismatch is a mathematical artifact of our design choice rather than a performance failure; precise spatial localization is far more critical for surgical safety than exact replication of a probabilistic distribution. The standard IoU metric for localization is not added to our evaluation experiments, as it is included in our training loss. Our model substantially outperforms all evaluated baselines. Relative to our approach, the two strongest competitors – a fully-supervised U-Net and a fine-tuned Molmo-VLM – exhibit 1.16× and 1.93× higher ASSD, alongside 23.16% and 62.34% increases in HD, respectively. These results indicate that neither standard medical segmentation networks nor large-scale foundation models do not match the performance of our task-specific architecture for dense heatmap prediction in laparoscopic data.

(ii) Ablation study: We perform an ablation study to validate each component of our workflow. Table 3 demonstrates the impact of the SigLIP language encoder, the Swin vision encoder, and the AdaLN decoder. The decoder and temporal vision encoder choices have the most crucial effect on model's performance. Replacing AdaLN decoder with cross-attention decoder [27] reduces memory efficiency and increases HD by 35.45% and ASSD by 39.78%. This highlights the importance of adaptive feature modulation in our setting, suggesting that global conditional normalization is better suited for structured heatmap prediction than token-level cross-attention fusion. Similarly, replacing the Swin Transformer with a 3D ResNet-18 [10] leads to a significant performance degradation - (HD +42.43%, ASSD +18.34%), justifying our choice of vision backbone. Table 4 evaluates the language encoder and dataset augmentations. The importance of language encoding is evident as the model's performance drops drastically when it is removed: **ASSD increases by 109.8% and HD by 213.7%**.

Table 3. Ablation study of our model with a different (i) language encoder, (ii) vision encoder, and (iii) decoder.

Metrics	DICE↑	PCK@0.05↑	PCK@0.1↑	HD(px)↓	ASSD(px)↓
Ours	**0.124**	**0.517**	**0.667**	**79.763**	**20.557**
Ours with Bert-based language encoder	0.106	0.502	0.651	84.403	23.891
Ours with vision encoder	0.086	0.422	0.609	113.613	24.329
Ours with cross-attention decoder	0.085	0.443	0.571	108.039	28.736

Table 4. Ablation study of our model without (i) A (image augmentations) and (ii) L (SigLip language encoder).

Metrics	DICE↑	PCK@0.05↑	PCK@0.1↑	HD(px)↓	ASSD(px)↓
Ours	**0.124**	**0.517**	**0.667**	**79.763**	**20.557**
Ablation (w/o A)	0.094	0.491	0.625	127.101	29.181
Ablation (w/o L)	0.068	0.205	0.348	170.482	43.135

Table 5. Ablation study of our model without (i) action, (ii) previous frames, and (iii) tool specification in input.

Metrics	DICE↑	PCK@0.05↑	PCK@0.1↑	HD(px)↓	ASSD(px)↓
Ours	**0.124**	**0.517**	**0.667**	**79.763**	**20.557**
Ablation (w/o action)	0.112	0.350	0.542	104.733	22.087
Ablation (w/o previous frames)	0.103	0.490	0.635	85.932	24.973
Ablation (w/o tool)	0.101	0.320	0.495	93.702	27.302

Table 6. Ablation study of our model on early pre-action frames without (i) previous frames in input.

Metrics	DICE↑	PCK@0.05↑	PCK@0.1↑	HD(px)↓	ASSD(px)↓
Ours on early pre-action frames	**0.114**	**0.493**	**0.641**	**85.710**	**22.930**
Ablation (w/o previous frames)	0.092	0.317	0.609	102.70	31.95

Because the dataset contains various tool–action pairs, the model must be given a mechanism to differentiate between them, making the language encoder a crucial part of the pipeline. In contrast, image augmentations provide a modest but consistent improvement, showing that while they contribute to robustness, they are secondary to model performance.

Table 5 provides an ablation study for different input structures that we use in our pipeline: temporal vision context and action and tool specification in the prompt. Removing the tool specification leads to a substantial performance drop, with ASSD increasing by 32.81%. This result is expected, since many samples in our dataset include frames with multiple instruments, making it confusing for the model to understand for which tool the affordance region should be predicted. In comparison, removing temporal context across all pre-action frames yields a more moderate performance decline (+21.48% ASSD, +7.73% HD). To isolate the impact of temporal data, we evaluate our model solely on early pre-action frames—a phase where historical context is critical for establishing reliable guidance. As shown in Table 6, excluding temporal data in such samples leads to a much more noticeable degradation (**ASSD +39.34%, HD +19.82%**). These results demonstrate that tracking multi-viewpoint tool-tissue dynamics allows the model to capture tissue deformation and operator intent. Ultimately, this provides robustness against occlusions and motion blur while significantly refining the spatial consistency of predictions at the start of the maneuver.

5 Discussion and Conclusion

We present AffordTissue, a multimodal framework combining a temporal vision encoder, language conditioning, and a DiT-style decoder for predicting tool-action specific tissue affordance regions as dense heatmaps for six unique tool-action pairs critical to the cholecystectomy procedure. By explicitly conditioning on tool-action specifications and temporal context, our approach achieves more precise affordance localization than vision/vision-language baseline models such as U-Net and Molmo-VLM. The predicted affordance maps potentially unlock two complementary modes for safe surgical automation: a conditioning signal guiding learned policies toward appropriate tissue regions, and a safety layer enabling early automated stopping when instruments deviate outside predicted affordance regions. Future directions include extending affordance prediction to ground on the surgical phase for stage-specific reasoning, modeling diverse tool-tissue actions, and integrating with VLAs for closed-loop surgical automation.

Disclosure of Interests. The authors have no competing interests.

Acknowledgments. This work was in part supported by the Johns Hopkins University SURPASS Grant, Malone Center Seed Grant, and Johns Hopkins University internal funds.

References

1. AI, M., et al.: Sam 3: Segment anything with concepts. In: International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=r35clVtGzw
2. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
4. Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., Valdastri, P.: Autonomy in surgical robotics. Annual Review of Control, Robotics, and Autonomous Systems **4**(1), 651–679 (2021)
5. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: pi_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
6. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025)
7. Deitke, M., Schwenk, D., Salvador, J., VanderBilt, L., Aniol, K., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. arXiv preprint arXiv:2409.17146 (2024)
8. Gerig, G., Jomier, J., Chakos, M.: Valmet: A new software tool for assessing and visualizing segmentation accuracy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2001)

9. Haidegger, T.: Autonomy for surgical robots: Concepts and paradigms. IEEE Transactions on Medical Robotics and Bionics **1**(2), 65–76 (2019)
10. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6546–6555 (2018)
11. Kim, J.W., Chen, J.T., Hansen, P., Shi, L.X., Goldenberg, A., Schmidgall, S., Scheikl, P.M., Deguet, A., White, B.M., Tsai, D.R., et al.: Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. Science robotics **10**(104), eadt5254 (2025)
12. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
14. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
16. Lu, D.V., Hershberger, D., Smart, W.D.: Layered costmaps for context-sensitive navigation. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 709–715. IEEE (2014)
17. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al.: Surgical data science–from concepts toward clinical translation. Medical image analysis **76**, 102306 (2022)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV) (2016)
19. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8688–8697 (2019)
20. Oh, K.H., Borgioli, L., Mangano, A., Valle, V., Di Pangrazio, M., Toti, F., Pozza, G., Ambrosini, L., Ducas, A., Žefran, M., et al.: Expanded comprehensive robotic cholecystectomy dataset (crcd). arXiv preprint arXiv:2412.12238 (2024)
21. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
23. Scheikl, P.M., Tagliabue, E., Gyenes, B., Wagner, M., Dall'Alba, D., Fiorini, P., Mathis-Ullrich, F.: Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery. IEEE Robotics and Automation Letters **8**(2), 560–567 (2022)
24. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
25. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2:

Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
26. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems (2017)
28. Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical image analysis **86**, 102770 (2023)
29. Xu, R., Zhang, J., Guo, M., Wen, Y., Yang, H., Lin, M., Huang, J., Li, Z., Zhang, K., Wang, L., et al.: A0: An affordance-aware hierarchical model for general robotic manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13491–13501 (2025)
30. Yang, A., Yang, B., Zhang, B., Hui, B., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
31. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11975–11986 (2023)
32. Zia, A., Berniker, M., Nespolo, R., Perreault, C., Wang, Z., Mueller, B., Schmidt, R., Bhattacharyya, K., Liu, X., Jarc, A.: Surgical visual understanding (surgvu) dataset. arXiv preprint arXiv:2501.09209 (2025)
33. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)