2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SurgAM: Surgical Affordance Map Prediction with Multimodal Feature Fusion for Robot Autonomy

Lei Song, Yonghao Long, Mengya Xu, Jiayi Geng, Xiuyuan Chen†, Qi Dou†

Abstract— **Surgical automation is being increasingly studied, yet bridging visual scene understanding with autonomous action planning remains a fundamental challenge. While much research effort has been made on scene perception (e.g., tool recognition and scene segmentation), understanding and predicting actionable possibilities for surgical automation is still underexplored. In this paper, we introduce surgical affordance prediction, which identifies actionable regions for fundamental surgical actions from visual data. Specifically, a novel adaptive feature fusion framework is proposed that leverages the complementary strengths of a self-supervised vision transformer encoder for its superior semantic understanding and a large-scale generative model encoder for its spatially-aware capability. Furthermore, we introduce a hierarchical prompt learning mechanism to adapt to varying procedural contexts. Finally, a scene-guided attention decoder is proposed to focus on critical surgical areas while suppressing background distractions. To validate the effectiveness, we established a new dataset, derived from publicly available surgical datasets with affordance annotations for three basic surgical actions: aspiration, clipping, and retraction. Extensive experiments demonstrate that our approach achieves state-of-the-art performance. Moreover, we validate our framework's applicability for downstream automation on a realistic lung and prostate phantom, and results show that the predicted affordance maps successfully enable autonomous surgical actions.**

I. INTRODUCTION

Robotic surgery has been increasingly adopted in modern healthcare, demonstrating promising clinical benefits. Building on this success, there is a growing trend toward surgical automation [1], which promises to further reduce surgeon workload while improving procedural efficiency toward higher-level autonomy and human-robot collaboration [2]. To this end, autonomous robotic surgical systems must not only recognize what is in the surgical scene (e.g., surgical instruments or anatomical structure), but also interpret the scene in a way that directly supports manipulation, understanding what can be done where, and how to do it safely and effectively [3]. Typical perception tasks, such as key point extraction, instrument detection, and scene segmentation [4], have been widely studied, which provided valuable scene information. However, they just describe what is in the field of view, not where the robot should act [5]. The bridge between visual scene understanding and autonomous action planning remains a challenge for surgical automation [6].

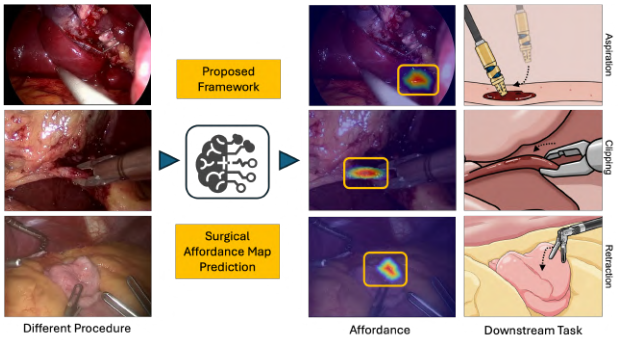

Fig. 1. Overall concept for surgical scene affordance map prediction: Model generates affordance map to identify optimal manipulation regions for specific surgical tasks, enabling downstream applications in robotic surgery and surgical planning.

To address this challenge, traditional methods often rely on simplistic visual cues or static pre-operative models, which are insufficient for the highly dynamic scenes encountered in surgery. Recent attempts have explored different learning-based paradigms for surgical automation. For example, end-to-end approaches directly map visual inputs to robotic actions through imitation learning [7], [8], demonstrating promise in controlled scenarios but suffering from overfitting to specific training demonstrations with poor generalization. Multistage learning methods decompose the problem into sequential modules, such as perception followed by planning and control [9], where perception is often used for providing semantic mask and depth information, and control to plan the action. Although promising, these methods do not provide direct, interpretable guidance linking visual understanding to safe and effective actions across variable surgical contexts.

Affordance prediction in industrial robotics directly addresses such a challenge by predicting actionable possibilities of manipulating objects or environments. Specifically, affordance maps identify regions that support specific interactions, forming an interpretable bridge from perception to action [10], [11]. For instance, grasping affordance maps highlight optimal contact points and orientations for a gripper [12], while pushing affordance maps indicate where applied forces will yield desired object motions [13]. These approaches are effective because they encode both geometric constraints and functional relationships between objects and actions, enabling generalization across objects

L. Song, Y. Long, M. Xu, and Q. Dou are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. J. Geng and X. Chen are with the Department of Thoracic Surgery, Peking University People's Hospital, Beijing, China. J. Geng and X. Chen are with the Thoracic Oncology Institute, Peking University People's Hospital, Beijing, China. J. Geng and X. Chen are with the Research Unit of Intelligence Diagnosis and Treatment in Early Non-small Cell Lung Cancer, Chinese Academy of Medical Sciences, 2021RU002, Peking University People's Hospital, Beijing, China. J. Geng and X. Chen are with the Institute of Advanced Clinical Medicine, Peking University, Beijing, China. J. Geng and X. Chen are with Beijing Key Laboratory of Innovative Application of Big Data in Lung Cancer, Peking University People's Hospital, Beijing, China. Corresponding authors: Qi Dou (qidou@cuhk.edu.hk), Xiuyuan Chen (dr_chenxy@pku.edu.cn).

and environments for robust robot manipulation.

Despite the success of affordance prediction in industrial robotics, its application for surgical robotics still remains unexplored. In this paper, we propose to predict the surgical affordance map: estimating spatial target regions suitable for fundamental surgical actions to provide direct, actionable guidance for surgical robotics. To the best of our knowledge, this is the first comprehensive study of surgical affordance map prediction. An illustration in Fig. 1 shows affordance map prediction for fundamental surgical actions including aspiration, clipping, and retraction, where each map identifies optimal manipulation regions to guide robotic surgical interventions. Specifically, we propose a novel feature fusion framework that generates powerful visual representation by fusing dense semantic features with spatial features. Our key insight is that leveraging pre-trained vision models can provide complementary strengths for surgical affordance prediction. Specifically, transformers trained with self-supervised objectives demonstrate strong feature representation capabilities for fine-grained visual understanding, while generative models trained on diverse visual data excel at capturing spatial layout and providing smooth, coherent representations of anatomical structures. By combining these with an adaptive fusion strategy, hierarchical prompt learning, and a cross-modal alignment decoder, our framework generates robust and accurate surgical affordance maps.

Our main contributions are summarized as follows:

- We present the first study on affordance map prediction for surgical applications, bridging visual scene understanding and actionable robotic guidance.
- We propose a new feature fusion framework leveraging semantic and spatial features with adaptive fusion, hierarchical prompt learning, and cross-modal alignment.
- We establish a surgical affordance dataset with three fundamental surgical actions (aspiration, clipping, retraction) and demonstrate significant improvements over state-of-the-art baselines.
- We validate our method's practical applicability through integration with an existing surgical automation framework and realistic phantom experiments on da Vinci Research Kit (dVRK).

II. RELATED WORK

A. Surgical Scene Perception

Traditional surgical scene perception has focused on component-level understanding through specific subtasks such as instrument segmentation [14], anatomical structure detection [15], and surgical workflow recognition [16]. While these methods have achieved notable success in controlled scenarios, they provide isolated component identification that lacks holistic spatial-semantic reasoning necessary for autonomous surgical manipulation [3]. Segmentation-based approaches can identify anatomical structures but fail to indicate where and how a robot should interact with these structures. Similarly, detection methods locate surgical instruments but provide limited guidance on feasible manipulation spaces within dynamic surgical environments [17]. This limitation creates a critical gap between scene understanding and actionable intelligence, motivating affordance-based approaches for surgical manipulation.

B. Autonomy in Robotic Surgery

The pursuit of autonomy in robotic surgery aims to enhance precision, reduce surgeon fatigue, and standardize procedural outcomes [18]. Early approaches often relied on extensive, task-specific programming for subtasks like suturing, which limited their adaptability [19]. More recently, learning-based methods have demonstrated impressive capabilities. A landmark achievement by Kim et al. showed an autonomous robotic system successfully performing laparoscopic small bowel anastomosis in porcine models, outperforming human surgeons in consistency and accuracy [20]. However, even in state-of-the-art systems, the perception-to-action pipeline often relies on simplified representations such as tracking fiducial markers or registering anatomy to a pre-operative plan [21]. These methods typically lack a deep semantic understanding of the tissue itself, making them vulnerable to unexpected deformations or changes in the surgical scene. Explicitly identifying actionable regions as affordance maps remains largely unexplored, representing a critical missing link for robust surgical autonomy.

C. Affordance Map Prediction

Affordance prediction, rooted in ecological psychology [10], provides a natural bridge between perception and manipulation by identifying regions in an environment that support specific actions. In computer vision and robotics, this concept has evolved from early CNN-based methods [22] to recent approaches that leverage large-scale foundation models. For instance, AffordanceLLM [23] utilizes the world knowledge of Vision Language Models for affordance grounding, and other works adapt models like CLIP for nuanced visual understanding [24]. Concurrent research has also begun to develop foundation models specifically tailored for the surgical domain, demonstrating the growing interest in affordance prediction for robotic surgery. However, translating these general-domain methods to surgery remains challenging. Surgical scenes are characterized by deformable anatomy, occlusions, and dynamic lighting, which demand feature representations with both high semantic precision and spatial coherence [25]. Recent work [26] has highlighted the complementary nature of features from models like Stable Diffusion [27], which excel in spatial layout, and DINOv2 [28], which provides robust semantic understanding. This suggests that a strategic fusion of such features holds significant potential for tackling the unique challenges of surgical affordance prediction.

III. METHODS

In this section, we present our novel framework for surgical scene affordance map prediction that leverages complementary foundation models and hierarchical prompt learning. Our approach addresses key challenges in surgical affordances understanding by combining the semantic precision

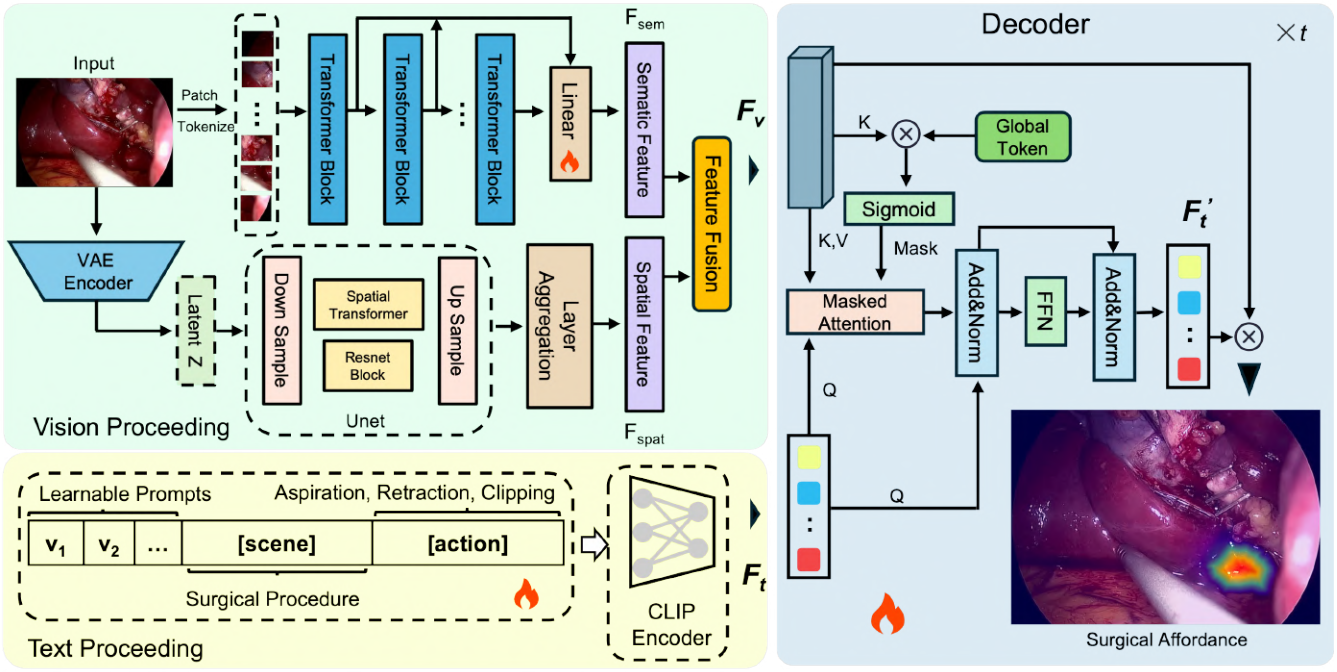

Fig. 2. Framework of surgical scene affordance map prediction. The architecture features two parallel streams on the left for multimodal feature extraction: a dual-vision encoder that fuses complementary features from a self-supervised vision transformer and a diffusion-based generative model, and a text encoder that leverages hierarchical prompt learning. On the right, a surgical scene guided cross-modal decoder integrates these features to generate the final affordance map.

of self-supervised vision transformers [28] with the spatial awareness of diffusion-based generative models [27], while incorporating surgical context through structured prompt learning to enhance domain-specific understanding.

A. Overall Framework

The prediction of surgical affordances in dynamic, deformable environments requires a visual representation that simultaneously captures two critical properties: fine-grained semantic precision and holistic spatial coherence. Semantic precision is vital for distinguishing between different anatomical structures, while spatial coherence is essential for defining smooth, physically actionable regions on tissue. The proposed learning framework, presented in Fig. 2, is designed to meet this dual requirement through four key components: dual vision encoders, a complementary feature fusion module, a hierarchical text prompt encoder, and a surgical scene guided cross-modal decoder.

Central to our approach, two complementary vision encoders are used: one from a self-supervised vision transformer for dense semantic features, and the other from a diffusion-based generative model for spatially-aware representations. The fusion module then strategically combines these representations into a unified feature map. For text processing, we use a hierarchical prompt learning strategy to incorporate surgical context, enabling the model to distinguish between affordances in varying scenarios. Finally, a lightweight surgical scene guided cross-modal decoder generates affordance predictions by using a global context-guided attention mechanism to fuse the visual and text features, focusing on relevant surgical areas.

The following subsections detail each component's design and implementation, demonstrating how they collectively address the challenges of surgical affordance prediction.

B. Complementary Visual Feature Extraction and Fusion

A single type of pre-trained feature is often insufficient for the nuanced requirements of surgical affordance prediction [29]. Recent studies have highlighted a fundamental trade-off in mainstream foundation models: features from generative models (diffusion models) provide spatially smooth but semantically imprecise representations, whereas features from self-supervised vision transformers are semantically accurate but often sparse and noisy [26].

Therefore, our core technical contribution is a fusion strategy that synergizes these complementary feature types to form a more robust representation. We extract features celebrated for their semantic correspondence from a vision transformer [28] and features known for their spatial layout understanding from a diffusion-based generative model [27].

Multi-Layer Semantic Feature Aggregation. To capture fine-grained semantic details at multiple granularity levels, we aggregate features from the last j layers of the vision transformer:

$$\mathbf{F}_{sem} = \sum_{i=1}^{j} \alpha_i \mathbf{F}_i, \tag{1}$$

where $\mathbf{F}_{sem}$ represents the semantic features, $\mathbf{F}_i$ denotes the i-th layer features, and the learnable weights with $\sum_i \alpha_i = 1$.

Multi-Scale Spatial Feature Extraction. To capture the holistic spatial layout and context, we extract and concatenate features from multiple decoder layers of the generative

model's U-Net:

$$\mathbf{F}_{spat} = \text{Concat}[\mathbf{F}^{(2)}_{spat}, \mathbf{F}^{(5)}_{spat}, \mathbf{F}^{(8)}_{spat}], \tag{2}$$

where $\mathbf{F}_{spat}$ represents the spatial features from layers 2, 5, and 8, which are processed through appropriate dimensionality reduction.

Adaptive Feature Fusion. To intelligently combine the strengths of both feature types, we implement an adaptive fusion strategy:

$$\mathbf{F}_v = \text{LayerNorm}(\beta \mathbf{F}_{sem} + (1-\beta)\mathbf{F}_{spat}), \tag{3}$$

where β is a learnable balancing parameter that is optimized during training to balance semantic precision and spatial coherence, and features are normalized before fusion to ensure training stability.

C. Hierarchical Prompt Learning

Manually designing prompts for surgical affordances presents significant challenges, as traditional templates [30] like "somewhere to [affordance]" fail to capture the complex contextual information inherent in surgical procedures. For instance, "clipping" may refer to hemostatic control of active bleeding vessels, prophylactic sealing of at-risk vasculature, or tissue division for exposure—each requiring different spatial considerations and anatomical landmarks. Moreover, CLIP exhibits limited understanding of fine-grained affordances, particularly in the specialized domain of surgical robotics, where the same action may have different meanings across various procedural contexts.

To address these limitations, we extend the Context Optimization (CoOp) [31] method with a hierarchical prompt learning strategy:

$$\mathbf{P} = [\mathbf{v}_1, \ldots, \mathbf{v}_p, \mathbf{s}_{context}, \mathbf{s}_{action}], \tag{4}$$

where $\mathbf{s}_{context}$ represents surgical context descriptors, $\{\mathbf{v}_i\}$ are learnable vectors, and $\mathbf{s}_{action}$ denotes the target affordance class.

The hierarchical prompts are processed through the CLIP text encoder to generate text embeddings:

$$\mathbf{F}_t = \text{CLIP}_{\text{text}}(\mathbf{P}), \tag{5}$$

where $\mathbf{F}_t \in \mathbb{R}^{N \times D}$ represents the final text embeddings for N affordance classes.

D. Surgical Scene Guided Cross-Modal Attention

The decoder takes as input the fused visual features $\mathbf{F}_v$ and text embeddings $\mathbf{F}_t$ to generate precise affordance predictions. Drawing inspiration from recent advances in vision-language alignment [32], we design a scene guided cross-attention mechanism that leverages global surgical context for precise affordance localization.

The attention mask is computed by projecting the global scene representation onto local visual features:

$$\mathbf{M} = \sigma\left(\frac{\mathbf{S}_{global}\mathbf{K}^T}{\sqrt{d}}\right), \tag{6}$$

where $\mathbf{S}_{global}$ represents the global surgical scene token (CLS token) extracted from $\mathbf{F}_v$, and $\mathbf{K} = \text{Linear}(\mathbf{F}_v)$ denotes the key projections from visual features.

The masked cross-attention mechanism integrates text queries with visual representations:

$$\mathbf{F}_{out} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} \odot \mathbf{M}\right)\mathbf{V} + \mathbf{F}_t, \tag{7}$$

where $\mathbf{Q} = \text{Linear}(\mathbf{F}_t)$, $\mathbf{V} = \text{Linear}(\mathbf{F}_v)$, and the residual connection preserves the original text information.

The final surgical affordance prediction combines the refined embeddings:

$$\mathbf{P}_{aff} = \text{softmax}\left(\mathbf{F}_{out}\mathbf{F}_v^T\right), \tag{8}$$

where $\mathbf{P}_{aff} \in \mathbb{R}^{N \times L}$ represents the spatial probability distribution over N surgical actions across L image locations.

IV. EXPERIMENTS

In this section, we conduct a series of experiments to systematically evaluate our proposed framework. We aim to answer three key questions: (1) How does our full model perform against state-of-the-art methods on our challenging surgical affordance dataset? (2) What is the individual contribution of each core component of our design, particularly the dual-feature fusion strategy? (3) Can the predicted affordance maps be effectively translated into actionable guidance for a physical robot in surgical manipulation tasks?

A. Dataset

To address the scarcity of specialized surgical data, we constructed a new affordance dataset from a collection of diverse, publicly available surgical video corpora. Our annotation methodology is uniquely designed to capture surgical *intent*. For each interaction, we identify the "pre-contact" frame just before the tool touches the tissue and annotate a single point indicating the intended interaction location. This point then serves as a visual prompt for the Segment Anything Model (SAM) [33] to generate a high-fidelity segmentation mask of the actionable and safe tissue region, which is subsequently converted into a heatmap to serve as the ground truth.

Our final dataset consists of annotations from 1,915 surgical video sequences from 7 public datasets: AutoLaparo [34], CholecT50 [35], HeiChole [36], MultiBypass140 [37], SurgicalActions160 [38], Endovis18 [39], and MESAD-Real [40], categorized into three fundamental affordance types: **Retraction**, **Clipping**, and **Aspiration**. The distribution of these clips is as follows: 904 for Retraction, 231 for Clipping, and 780 for Aspiration. This imbalanced distribution is representative of the natural frequency of these actions during actual surgical procedures, where tissue retraction and aspiration are more common maneuvers than clipping. This dataset provides a robust foundation for training and evaluating affordance prediction models in a variety of surgical contexts.

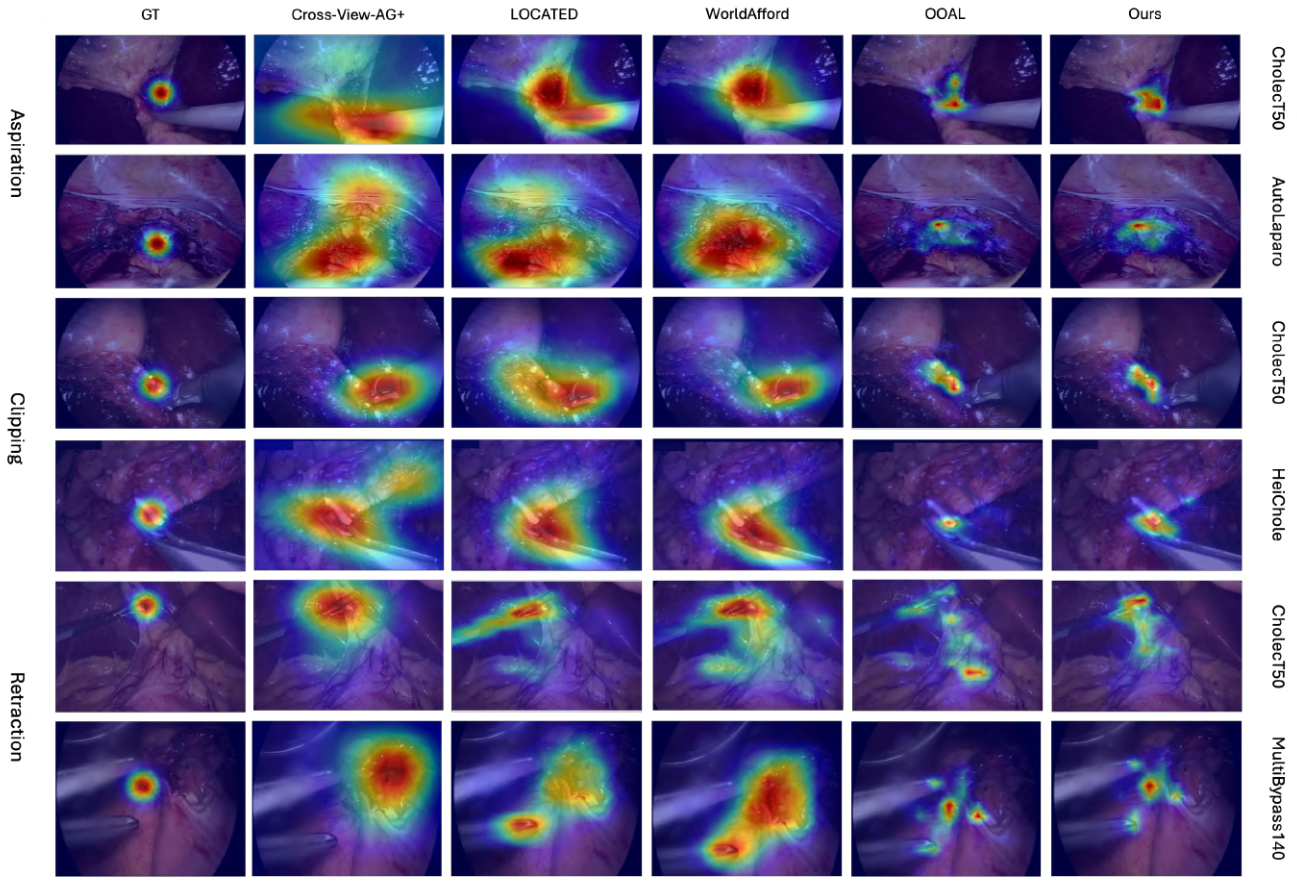

Fig. 3. Qualitative comparison of affordance prediction methods across different surgical tasks (aspiration, clipping, retraction) and datasets showing baseline methods vs. our approach

B. Experimental Setup

Metrics. We evaluate our model on our affordance dataset (details in Sec. IV-A) using a comprehensive set of established affordance prediction metrics [41], [42]. To assess the similarity between the predicted and ground-truth probability distributions, we use the Kullback-Leibler Divergence (KLD), where a lower value indicates less divergence and a better match. To measure spatial overlap, we use the Similarity (SIM) metric, for which higher is better. We also employ Normalized Scanpath Saliency (NSS), which evaluates the model's ability to assign high values to the ground truth affordance regions; a higher NSS score signifies a more accurate prediction. Furthermore, we propose Centroid Localization Accuracy (CLA) to quantify spatial precision in a normalized manner. Unlike raw Average Centroid Distance (ACD), which depends on image resolution, CLA normalizes by the image diagonal D: $CLA = 1 - ACD/D$. This produces an interpretable score between 0 and 1, where higher values indicate better spatial accuracy.

Implementation Details. We implement our framework using DINOv2 [28] as the self-supervised vision transformer and Stable Diffusion [27] as the diffusion-based generative model. Features are extracted using pre-trained checkpoints without fine-tuning backbone networks. For DINOv2, we aggregate features from the last 4 layers (j=4) with learnable weights. For Stable Diffusion, we extract features from U-Net decoder layers 2, 5, and 8, applying linear projections to align dimensions before fusion. The adaptive fusion parameter β is initialized to 0.5 and learned during training.

TABLE I

QUANTITATIVE RESULTS ON THE WHOLE SURGICAL DATASET.

Methods	KLD ↓	SIM ↑	NSS ↑	CLA ↑
Cross-View-AG [43]	1.795	0.260	1.227	0.765
Cross-View-AG+ [44]	1.793	0.253	1.265	0.782
LOCATE [41]	1.696	0.295	1.276	0.881
WorldAfford [45]	1.559	0.307	1.446	0.883
OOAL [32]	1.665	0.274	1.425	0.884
SurgAM (Ours)	**1.362**	**0.367**	**1.642**	**0.895**

C. Qualitative results and Analysis

Fig. 3 shows qualitative results on our test datasets. We compare our approach SurgAM, with several state-of-the-art methods that learn affordances from 2D images. All baseline models are trained and tested on the same datasets to ensure a fair comparison.

On the surgical test datasets, we observe distinct failure patterns across different baseline methods, highlighting the architectural limitations in surgical affordance prediction. **Cross-View-AG** [43] and **Cross-View-AG+** [44] produce overly expansive affordance maps due to their ACP strategy, despite achieving relatively high CLA scores (0.765-0.782). **LOCATE** [41] shows improved localization through its

PartSelect mechanism, yet its k-means clustering approach creates instability in complex scenes, leading to extensive anatomical coverage as a conservative recall strategy (CLA: 0.881). **WorldAfford** [45] leverages LLM reasoning through ARCoT but suffers from attention diffusion in its WCB module, producing over-activation patterns despite strong metrics (KLD: 1.559, CLA: 0.883). **OOAL** [32] presents distinctly different behavior with more concentrated predictions due to its DINOv2-based architecture and few-shot learning strategy, avoiding overfitting but struggling with complex multi-object scenarios due to single-modality limitations.

Our approach **SurgAM** consistently produces superior affordance maps that achieve an optimal balance between spatial precision and coverage completeness. Unlike baseline methods that achieve high confidence through over-activation strategies, our framework demonstrates genuine spatial accuracy with concentrated predictions that correspond to actual affordance regions. The synergistic fusion of Stable Diffusion's spatial understanding with DINOv2's semantic precision, enhanced by our surgical-aware hierarchical prompting strategy, enables accurate localization while maintaining spatial coherence. The substantial improvements across all evaluation metrics (KLD: 1.362, SIM: 0.367, NSS: 1.642, CLA: 0.895) validate the effectiveness of our dual-foundation model fusion approach. Notably, our method achieves the best KLD and NSS scores while maintaining competitive CLA performance, demonstrating that superior affordance prediction stems from architectural design rather than conservative prediction strategies employed by competing methods.

TABLE II

ABLATION RESULTS OF DIFFERENT VISUAL FOUNDATION MODELS.

Model	KLD ↓	SIM ↑	NSS ↑	CLA ↑
CLIP	2.660	0.156	0.276	0.819
DeiT III	2.231	0.165	0.292	0.531
SD	2.242	0.187	0.542	0.834
DINOv2	1.808	0.237	1.313	0.837

TABLE III

ABLATION RESULTS OF THE PROPOSED MODULES. HPL: HIERARCHICAL PROMPT LEARNING. SD-DINO: STABLE DIFFUSION AND DINOV2 FEATURE FUSION. TD: TRANSFORMER DECODER. SGM: SURGICAL SCENE GUIDED MASK

Components				Metrics			
HPL	SD-DINO	TD	SGM	KLD ↓	SIM ↑	NSS ↑	CLA ↑
				1.808	0.237	1.313	0.837
✓				1.742	0.268	1.240	0.847
✓	✓			1.473	0.334	1.528	0.885
✓	✓	✓		1.432	0.348	1.552	0.892
✓	✓	✓	✓	1.362	0.367	1.642	0.895

D. Ablation study

We conduct comprehensive ablation studies to validate the effectiveness of each component in our framework. The experiments are performed on our whole dataset, which encompasses diverse surgical scenarios and complex anatomical structures. **Visual Foundation Model Comparison.** Table II presents a systematic comparison of different visual foundation models for surgical affordance prediction. CLIP and DeiT III demonstrate limited effectiveness in surgical contexts, with poor KLD (2.660, 2.231) and NSS (0.276, 0.292) scores, indicating substantial distribution mismatch and weak affordance localization capabilities. Stable Diffusion features exhibit superior spatial understanding, achieving better NSS (0.542) and CLA (0.834) scores, validating that SD's spatial layout training provides valuable coherence for surgical scenes. However, SD underperforms in semantic precision (KLD: 2.242). DINOv2 demonstrates clear superiority across all metrics (KLD: 1.808, NSS: 1.313), with nearly 5× better NSS than CLIP. This superior performance stems from self-supervised training that learns part-aware representations, making it well-suited for fine-grained spatial understanding required in surgical affordance prediction.

Component-wise Analysis. Table III demonstrates the progressive improvement achieved by incorporating each proposed component. Starting from the DINOv2 baseline (KLD: 1.808), each module contributes meaningful performance gains, culminating in substantial improvements in the full model (KLD: 1.362, NSS: 1.642).

The most significant breakthrough comes from the SD-DINO fusion module, which represents the core innovation of our approach. This fusion strategy addresses the fundamental limitations observed in single-modality approaches by combining SD's spatial coherence with DINOv2's semantic precision. The substantial performance improvement validates our central hypothesis that surgical affordance prediction requires both spatial layout understanding and semantic correspondence capabilities, which qualities neither foundation model possesses independently.

The Hierarchical Prompt Learning (HPL) component provides essential domain adaptation, enabling the model to distinguish between affordances across different surgical contexts. The Transformer Decoder (TD) enhances contextual modeling by enabling sophisticated feature interactions, while the Surgical Scene Guided Mask (SGM) provides crucial attention focusing by leveraging global surgical context to suppress background distractions.

The final integrated system achieves optimal performance across all metrics, with particularly notable improvements in NSS (25.0% gain) and substantial KLD reduction (24.7% improvement), demonstrating that the synergistic combination of all components successfully addresses the multifaceted challenges of surgical affordance prediction.

E. Deployment and Validation of Autonomy on Phantom

The ablation studies confirmed the effectiveness of our design choices in generating accurate surgical affordance maps. To further validate the practical utility, we integrated our affordance prediction module with a visual servoing controller based on our VPPV framework [46] and deployed the overall framework on a physical surgical robot platform to perform surgical tasks autonomously.

Experimental Setup. Our physical setup, shown in Fig. 4, consists of a da Vinci Research Kit (dVRK) and two high-

fidelity surgical phantoms: a torso phantom with a silicone lung model and a standalone prostate phantom with simulated vasculature. The dVRK is equipped with an Endoscope Camera Manipulator (ECM) for visual feedback and two Patient Side Manipulators (PSMs) to control surgical instruments, enabling us to replicate key visual and kinematic aspects of different surgical scenarios. For each phantom, we collected targeted datasets through teleoperated demonstrations and fine-tuned task-specific models while retaining generalized features. During autonomous execution, our framework processed real-time video to generate affordance maps that guided the visual servoing controller.

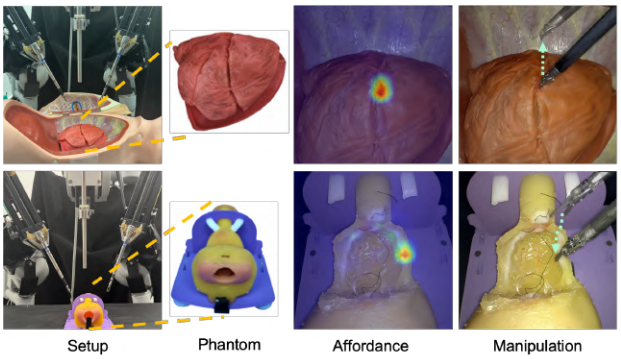

Fig. 4. Experimental setup for autonomous task validation using dVRK. Each row shows (from left to right): dVRK setup, phantom model, affordance prediction, and autonomous manipulation for lung retraction (top) and prostate clipping (bottom) tasks.

Autonomous Task Execution and Results. During the autonomous execution, our framework processed real-time video feeds from the ECM to generate affordance maps, which guide a visual servoing controller that commanded the PSM to perform the designated manipulation. For autonomous lung retraction with simulated circulation dynamics, our system consistently identified optimal retraction regions (as shown in the first row of Fig.4) while navigating tissue movement and deformation, achieving 100% success rate across 50 trials. For autonomous vessel clipping on the prostate phantom, we evaluated a clinically relevant workflow where expert surgeons first exposed target vessels through teleoperation, followed by autonomous clipping execution (prediction result shown in the bottom row of Fig.4), achieving 98% success rate (49/50 trials). The single failure occurred when camera repositioning temporarily occluded the target vessel, but affordance prediction immediately recovered once visibility was restored, demonstrating robust recovery capabilities. These results across two distinct surgical scenarios demonstrate that our framework provides reliable perception-to-action guidance for robotic surgical manipulation. However, complete surgical autonomy requires integration with comprehensive control systems and validation beyond controlled phantom environments.

V. CONCLUSION AND DISCUSSION

This paper presents a novel framework for surgical scene affordance map prediction that leverages the complementary strengths of diffusion-based generative model and vision transformer encoder. Our adaptive fusion strategy combines SD's spatial coherence with DINOv2's semantic precision, enhanced by hierarchical prompt learning and scene-guided attention mechanisms to address the unique challenges of surgical environments. Extensive experiments demonstrate the effectiveness of our approach across diverse surgical scenarios, with successful phantom implementations validating the practical applicability for autonomous surgical manipulation tasks. This work represents a significant step toward intelligent surgical robotic systems capable of adaptive decision-making in complex surgical environments. Future work will focus on incorporating temporal dynamics, expanding the annotated dataset scope, and conducting comprehensive validation studies on live surgical procedures.

ACKNOWLEDGMENT

This work was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 14208424), and in part by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

The work from J. Geng and X. Chen has been supported in part by the Peking University Medicine Plus X Pilot Program-Artificial Intelligence and Medical Development Initiative (BMU2025YXXLHAIYX002), Peking University People's Hospital Scientific Research Development Funds (RDX2024-07), and S&T Program of Xiongan New Area (XA202501102003K). The authors would like to thank Yiru Ye, Oranuch Ekkowit, Phutanate Pisutsin and Wong Pak Hin for their valuable contributions to the dataset annotation.

REFERENCES

[1] S. Schmidgall, J. D. Opfermann, J. W. Kim, and A. Krieger, "Will your next surgeon be a robot? autonomy and ai in robotic surgery," *Science Robotics*, vol. 10, no. 104, p. eadt0187, 2025.

[2] P. Fiorini, K. Y. Goldberg, Y. Liu, and R. H. Taylor, "Concepts and trends in autonomy for robot-assisted surgery," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 993–1011, 2022.

[3] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, 2021.

[4] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, E. De Momi, and N. Padoy, "EndoNet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[5] N. Ayobi, S. Rodríguez, A. Pérez, I. Hernández, N. Aparicio, E. Dessevres, S. Peña, J. Santander, J. I. Caicedo, N. Fernández, *et al.*, "Pixel-wise recognition for holistic surgical scene understanding," *Medical Image Analysis*, p. 103726, 2025.

[6] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," p. eaam8638, 2017.

[7] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger, "Surgical robot transformer (srt): Imitation learning for surgical tasks," *arXiv preprint arXiv:2407.12998*, 2024.

[8] M. Moghani, N. Nelson, M. Ghanem, A. Diaz-Pinto, K. Hari, M. Azizian, K. Goldberg, S. Huver, and A. Garg, "Sufia-bc: Generating high quality demonstration data for visuomotor policy learning in surgical subtasks," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 4534–4541.

[9] J. E. Knudsen, U. Ghaffar, R. Ma, and A. J. Hung, "Clinical applications of artificial intelligence in robotic surgery," *Journal of robotic surgery*, vol. 18, no. 1, p. 102, 2024.

[10] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[11] P. Ardón, È. Pairet, K. S. Lohan, S. Ramamoorthy, and R. Petrick, "Affordances in robotic tasks–a survey," *arXiv preprint arXiv:2004.07400*, 2020.

[12] L. Chen, M. Niu, J. Yang, Y. Qian, Z. Li, K. Wang, T. Yan, and P. Huang, "Robotic grasp detection using structure prior attention and multiscale features," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[13] M. Zhao, G. Zuo, S. Yu, D. Gong, Z. Wang, and O. Sie, "Position-aware pushing and grasping synergy with deep reinforcement learning in clutter," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 3, pp. 738–755, 2024.

[14] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2018, pp. 624–628.

[15] D. Kitaguchi, N. Takeshita, H. Matsuzaki, T. Oda, M. Watanabe, K. Mori, E. Kobayashi, and M. Ito, "Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: experimental research," *International journal of surgery*, vol. 79, pp. 88–94, 2020.

[16] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim, and N. Navab, "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 343–352.

[17] S. Sen, A. Garg, D. V. Gealy, S. McKinley, Y. Jen, and K. Goldberg, "Automating multi-throw multilateral surgical suturing with a mechanical needle guide and sequential convex optimization," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 4178–4185.

[18] T. Haidegger, "Autonomy for surgical robots: Concepts and paradigms," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, no. 2, pp. 65–76, 2019.

[19] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Science translational medicine*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.

[20] J. W. Kim, J.-T. Chen, P. Hansen, L. X. Shi, A. Goldenberg, S. Schmidgall, P. M. Scheikl, A. Deguet, B. M. White, D. R. Tsai, *et al.*, "Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning," *Science robotics*, vol. 10, no. 104, p. eadt5254, 2025.

[21] B. Zhan, W. Zhao, Y. Fang, B. Du, F. Vasconcelos, D. Stoyanov, D. S. Elson, and B. Huang, "Tracking everything in robotic-assisted surgery," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 1–7.

[22] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 5882–5889.

[23] S. Qian, W. Chen, M. Bai, X. Zhou, Z. Tu, and L. E. Li, "Affordancellm: Grounding affordance from vision language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7587–7597.

[24] M. Wang, J. Xing, J. Mei, Y. Liu, and Y. Jiang, "Actionclip: Adapting language-image pretrained models for video action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[25] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[26] J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang, "A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence," *Advances in Neural Information Processing Systems*, vol. 36, pp. 45 533–45 547, 2023.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[28] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[29] Y. Li, M. E. H. Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec, "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Computers in Biology and Medicine*, vol. 177, p. 108635, 2024.

[30] 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 *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[31] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[32] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani, "One-shot open affordance learning with foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3086–3096.

[33] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[34] Z. Wang, B. Lu, Y. Long, F. Zhong, T.-H. Cheung, Q. Dou, and Y. Liu, "Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 486–496.

[35] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Medical Image Analysis*, vol. 78, p. 102433, 2022.

[36] L. Maier-Hein, M. Wagner, T. Ross, A. Reinke, S. Bodenstedt, P. M. Full, H. Hempe, D. Mindroc-Filimon, P. Scholz, T. N. Tran, *et al.*, "Heidelberg colorectal data set for surgical data science in the sensor operating room," *Scientific data*, vol. 8, no. 1, p. 101, 2021.

[37] J. L. Lavanchy, S. Ramesh, D. Dall'Alba, C. Gonzalez, P. Fiorini, B. P. Müller-Stich, P. C. Nett, J. Marescaux, D. Mutter, and N. Padoy, "Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery," *International journal of computer assisted radiology and surgery*, vol. 19, no. 11, pp. 2249–2257, 2024.

[38] K. Schoeffmann, H. Husslein, S. Kletz, S. Petscharnig, B. Muenzer, and C. Beecks, "Video retrieval in laparoscopic video recordings with dynamic content descriptors," *Multimedia Tools and Applications*, vol. 77, no. 13, pp. 16 813–16 832, 2018.

[39] 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.

[40] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo, *et al.*, "The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods," *arXiv preprint arXiv:2104.03178*, 2021.

[41] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 922–10 931.

[42] S. Qian and D. F. Fouhey, "Understanding 3d object interaction from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 753–21 763.

[43] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning affordance grounding from exocentric images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2252–2261.

[44] ——, "Grounded affordance from exocentric view," *International Journal of Computer Vision*, vol. 132, no. 6, pp. 1945–1969, 2024.

[45] C. Chen, Y. Cong, and Z. Kan, "Worldafford: Affordance grounding based on natural language instructions," in *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2024, pp. 822–828.

[46] Y. Long, A. Lin, D. H. C. Kwok, L. Zhang, Z. Yang, K. Shi, L. Song, J. Fu, H. Lin, W. Wei, *et al.*, "Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery," *Science Robotics*, vol. 10, no. 104, p. eadt3093, 2025.