SurgLAT: Learning Latent Surgical Attention for Depth-Aware Robotic Laparoscope Control

Rulin Zhou, Qiujie Song, Yujie Ma, An Wang, Guoheng Ma, Guankun Wang, Xingrong Diao, Jiankun Wang, *Senior Member, IEEE* Weizheng Li, Liyong Zhu, and Hongliang Ren, *Senior Member, IEEE*

***Abstract*— Autonomous laparoscopic camera control requires continuous understanding of the surgeon's operative intent in dynamic surgical scenes, where the target operative region is not a stable physical object but a latent and temporally evolving attention state. In this work, we present SurgLAT, a causal online framework for latent surgical attention modeling and autonomous laparoscopic view control. SurgLAT leverages a frozen DINOv3 encoder and a state-conditioned spatial token mixer to extract operative evidence under a memory-guided spatial prior, while a selective causal latent memory module jointly models short-term motion continuity and long-horizon surgical intent evolution through dynamic retrieval of current, recent, and historical latent states. The learned latent surgical attention state is decoded into a probabilistic attention heatmap and operative region for downstream endoscope guidance. Beyond perception, we further introduce a robotic deployment framework with explicit laparoscopic Remote Center of Motion (RCM) constrained control based on virtual-axis formulation, together with redundancy-aware null-space initialization for stable and smooth manipulator motion. We validate the full system on real laparoscopic surgical videos and a physical robotic laparoscope platform. Experimental results demonstrate robust online operative-region tracking and stable autonomous endoscopy adjustment under occlusion, rapid motion, and target transitions, highlighting the effectiveness of latent surgical intent modeling for surgical autonomy.**

***Index Terms*— Surgical Attention Tracking, Autonomous Laparoscope Control, Vision-based Surgical Robotics**

I. Introduction

MINIMALLY invasive surgery (MIS) has become a standard paradigm in modern surgical practice due to its reduced trauma, faster recovery, and improved patient outcomes compared with open surgery [1], [2]. In laparoscopic surgery, the endoscopic video stream is the surgeon's primary perceptual interface to the operative field, making a stable and clinically meaningful field of view (FoV) essential for safe manipulation and efficient surgical workflow. However, conventional laparoscope positioning is commonly performed by an assistant following verbal instructions, which is susceptible to communication latency, fatigue, hand tremor, and inconsistent viewpoint adjustment [3], [4]. These limitations may cause image instability, loss of critical anatomy, and interruptions to the surgical flow, motivating increasing interest in autonomous robotic laparoscope control.

Existing FoV control methods mainly rely on explicit human commands or automatic image-guided strategies. Voice-, gaze-, head-motion-, or foot-pedal-based interfaces allow direct camera regulation, but still impose additional cognitive and operational burden on the surgeon [5]. Automatic approaches typically define the desired camera target using hand-crafted visual proxies, such as tool centroids, instrument distributions, or visual-servoing objectives [6]–[8]. Although effective for basic tool-following scenarios, these geometric objectives do not directly capture the task-dependent operative focus, which may correspond to a tool–tissue interaction, anatomical boundary, suturing region, or inspection area. Recent learning-based studies have attempted to infer intent-aware viewpoints, and SurgAtt-Tracker [9] further formulates online surgical attention tracking with expert-annotated operative attention regions. Nevertheless, existing attention-guided methods are still largely limited to 2D localization or image-plane recentering, without explicitly modeling surgical attention as a causal latent state or converting it into physically feasible robotic laparoscope motion.

Learning surgical attention for autonomous laparoscope control remains challenging. Surgical attention is not an explicit object category, but a latent operative intent shaped by surgical phase, tool–tissue interaction, and anatomical context. Meanwhile, laparoscopic videos are often affected by smoke, specular highlights, bleeding, tissue deformation, instrument occlusion, and rapid camera motion, which makes frame-wise prediction unstable. The operative focus also evolves with both temporal continuity and abrupt transitions: it should remain stable during continuous manipulation, yet adapt quickly when the surgeon changes instruments or shifts the target region.

This work was supported by the Ministry of Science and Technology (MOST) of China Key Project 2025YFE0122500, the Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant SGCX20250526153900001, and the Hong Kong Research Grants Council Collaborative Research Fund (CRF C4026-21G). (R. Zhou, Q. Song, and Y. Ma are co-first authors. Corresponding authors: *L. Zhu and H. Ren*.)

R. Zhou, Q. Song, A. Wang, G. Wang, and H. Ren are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR 999077, China (email: zhourulin@connect.hku.hk; 22060067@zju.edu.cn; wa09@link.cuhk.edu.hk; gkwang@link.cuhk.edu.hk; hlren@ee.cuhk.edu.hk).

Y. Ma is with the College of Mechatronics and Engineering, Shenzhen University, Shenzhen 518060, China (email:2024110162@mails.szu.edu.cn).

G. Ma, X. Diao, and J. Wang are with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China (email: 12211611@mail.sustech.edu.cn; 12332163@mail.sustech.edu.cn; wangjk@sustech.edu.cn).

W. Li and L. Zhu are with the Third Xiangya Hospital, Central South University, Changsha 410013, China. (email: 602594@csu.edu.cn; zly8128@csu.edu.cn).

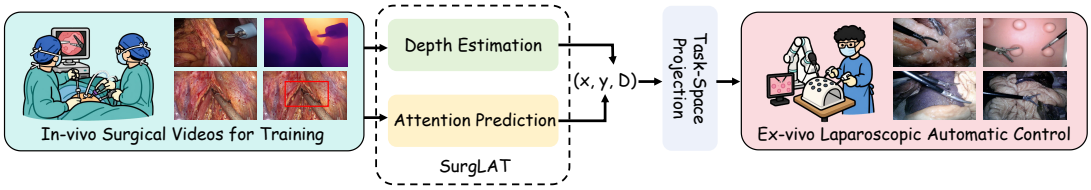

Fig. 1. Overview of the proposed perception-to-control pipeline: in-vivo surgical videos are used to learn surgical attention and depth-aware scale estimation, which are converted into task-space targets $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h})$ for ex-vivo robotic laparoscopic automatic control.

In addition, autonomous laparoscope control requires accurate and temporally stable visual targets that can be converted into physically feasible robot motion under the trocar-centered remote-center-of-motion (RCM) constraint.

To address these challenges, we introduce **SurgLAT**, a perception-to-control framework for autonomous laparoscopic FoV control. Given streaming endoscopic video, SurgLAT models surgical attention as a causal latent operative state rather than an instrument-following point or a generic object box. This latent state is decoded into a probabilistic attention heatmap, from which an attention center and an attention-centered operative region are derived for image-plane FoV guidance. In parallel, a depth-aware branch aggregates relative depth responses within the predicted attention region to estimate the operative scale, enabling adaptive zoom-in/out regulation along the viewing direction.

For robotic deployment, the predicted attention target and depth-aware scale cue are converted into task-space commands (x, y, D), where (x, y) guides lateral recentering and D regulates depth-wise camera motion. These commands are executed by a virtual-axis-based RCM-constrained controller, while a redundancy-aware null-space initialization strategy improves the feasible workspace and smoothness of the 7-DoF manipulator. In this way, SurgLAT integrates latent attention prediction, depth-aware FoV regulation, and physically constrained robotic execution into a unified closed-loop laparoscope control framework.

The main contributions are summarized as follows: (i) We design a **causal latent surgical attention model** that estimates operative focus from streaming laparoscopic video through memory-conditioned spatial reasoning and selective temporal updating. (ii) We introduce a **depth-aware operative scale estimation branch** that converts monocular relative depth responses into an adaptive cue for zoom-in/out laparoscope regulation. (iii) We develop a **virtual-axis-based RCM-constrained robotic control strategy** with redundancy-aware null-space initialization for real-time FoV adjustment on a physical 7-DoF laparoscope platform. (iv) Extensive experiments on **SurgAtt-1.16M [9] and real-world robotic validation** demonstrate accurate surgical attention prediction and stable closed-loop FoV control under occlusion, multi-instrument interference, and dynamic tissue interaction.

II. Related Work

A. Surgical Attention Modeling

1) Attention-Aware Field-of-View Guidance: Autonomous laparoscope control requires a reliable visual target that reflects the surgeon's operative intent. Early image-guided methods commonly define this target using hand-crafted geometric cues, such as a single-tool centroid, a multi-tool center, or an enclosing region around visible instruments [6], [7]. These methods are simple and effective for basic tool-following scenarios, but they implicitly assume that the desired field of view (FoV) is determined by instrument geometry. In real procedures, however, the clinically informative region may correspond to a tissue–tool interaction, an exposed anatomical boundary, a suturing site, or an inspection area rather than the geometric center of visible tools. To move beyond purely geometric objectives, recent studies have explored intent-aware FoV guidance using gaze, voice commands, surgeon preference modeling, imitation learning, and action-aware prediction [5], [10], [11]. These methods provide more semantically meaningful camera targets, but many of them still rely on explicit user input, platform-specific action prediction, or predefined control policies. SurgAtt-Tracker [9] formulates surgical FoV guidance as online surgical attention tracking and introduces expert-annotated operative attention regions.

2) Temporal Robustness in Surgical Video Prediction: Surgical attention estimation is inherently temporal, since the operative focus should remain stable during continuous manipulation while adapting to instrument changes, target shifts, and unexpected events. Prior studies have improved the stability of image-guided laparoscope control through temporal smoothing, probabilistic modeling, and predictive tracking. Visual-servoing-based systems regularize target trajectories or camera motion to suppress high-frequency jitter [12], probabilistic frameworks encode surgeon preferences or instrument-state distributions from historical observations [13], [14], and hybrid tracking modules combine online tracking with future-position prediction to compensate for latency [15]. However, these methods mainly stabilize explicit visual proxies, such as tools, centroids, or selected geometric points. In contrast, surgical attention is a task-dependent latent state that may shift between instruments, tissues, anatomical structures, and interaction regions, motivating a causal temporal model that preserves attention continuity while allowing rapid state updates [16].

B. RCM-Constrained Laparoscope Control

Autonomous laparoscope control must satisfy the remote-center-of-motion (RCM) constraint imposed by the trocar port, which restricts the laparoscope shaft to rotate around the incision point and is essential for safe minimally invasive manipulation. Prior studies on image-based visual servoing and model-based robotic camera control typically define image-plane errors using tool tips, instrument centroids, or selected visual features, and then compute camera motion to minimize

the error [6], [17], [18]. Other methods incorporate camera-quality assessment, heuristic policies, constrained optimization, RCM-aware planning, or redundant manipulator control to improve view stability and deployment feasibility [12], [14], [19]–[23]. However, these approaches are often limited to 2D recentering and remain strongly coupled with instrument geometry. Consequently, the camera may satisfy the geometric tracking objective while failing to maintain the most informative operative view.

Comprehensive integration of learned surgical attention prediction, depth-aware FoV regulation, RCM-constrained execution, and real-world robotic validation remains insufficiently explored. Different from prior work that treats perception and control as loosely connected components, our framework converts the predicted attention center and depth-aware operative scale cue into task-space commands for lateral recentering and zoom-in/out regulation. These commands are executed through a virtual-axis-based RCM-constrained controller with redundancy-aware null-space initialization, enabling physically feasible and intention-aware laparoscope motion.

III. Methodology

A. Problem Formulation and Overview

Given a streaming laparoscopic video $\{I_1, \ldots, I_t\}$, our goal is to estimate an operative attention target that indicates where the laparoscope should look and convert it into physically feasible camera motion. Instead of directly mapping visual observations to camera velocities, SurgLAT decouples perception from control. The perception model predicts an image-plane operative attention region and an axial scale cue, while the downstream RCM-constrained controller maps them into robot-executable FoV adjustment commands. This formulation is motivated by the fact that surgical attention is not a persistent object with stable appearance or boundaries, but a latent and temporally evolving operative focus. SurgLAT therefore maintains a causal latent attention state from streaming video, integrating current visual evidence, short-term motion continuity, and long-term surgical-intent memory. The resulting attention representation is decoded into an operative region for image-plane guidance and combined with depth-aware scale regulation for robotic laparoscope control.

B. Latent Surgical Attention Tracking Model

1) Visual Token Encoding: For each incoming laparoscopic frame I_t, we first extract dense visual tokens using a frozen DINOv3 [24] visual foundation encoder. The encoder maps the current frame into a patch-level feature grid F_t, where each token encodes local semantic and appearance evidence from the endoscopic scene. Since the model is designed for causal online inference, this encoding is performed independently for each frame and does not use future observations or bidirectional temporal context. We then project the DINOv3 features into a unified hidden dimension and add two-dimensional positional embeddings, producing spatially aware visual tokens for subsequent attention modeling.

2) Memory-Conditioned Spatial Token Mixer: In laparoscopic surgery, the operative attention region is usually a spatially concentrated focus rather than a uniformly distributed scene cue. Although DINOv3 provides dense and semantically rich visual tokens over the entire laparoscopic scene, these tokens are not explicitly biased toward the current operative focus. Meanwhile, the attention region in the current frame is often correlated with its previous location, as tool–tissue interaction and camera motion evolve continuously over time. Therefore, instead of treating each frame as an independent global search problem, SurgLAT uses the previous latent attention state Z_{t-1} to generate a memory-conditioned spatial prior for the current frame. This prior provides a coarse reference of where the operative focus is likely to appear, enabling more focused feature refinement while preserving global contextual reasoning.

As shown by the **Memory-conditioned Prior Center** and **Spatial Token Mixer** in Fig. 2, the previous latent state tokens Z_{t-1} are first used to predict a normalized prior center p_t, which is then converted into a Gaussian spatial prior A_t over the DINO token grid. Instead of cropping or masking the image, the prior softly modulates the current visual tokens:

$$F_t^{\text{prior}} = F_t \odot (1 + \alpha A_t), \tag{1}$$

where F_t denotes the projected DINOv3 token grid with 2D positional encoding, and α controls the prior strength. In this way, tokens around the expected operative focus are enhanced, while tokens outside the prior region remain available for full-scene reasoning.

The modulated token grid is then processed by a spatial transformer to obtain refined dense visual tokens.

$$F'_t = \text{SpatialTransformer}(F_t^{\text{prior}}). \tag{2}$$

The spatial transformer maintains two token streams: dense image tokens and a small set of learnable spatial evidence tokens. The dense tokens perform self-attention over the full spatial grid to exchange global context while preserving spatial layout. The learnable spatial tokens then serve as task-specific evidence queries and cross-attend to the refined dense tokens:

$$S_t = \text{CrossAttn}(Q_s, F'_t, F'_t), \tag{3}$$

where Q_s denotes the learnable spatial queries. These spatial evidence tokens are further refined through self-attention and feed-forward layers, allowing interactions among multiple evidence slots. Finally, the refined dense grid F'_t is passed to the state-conditioned heatmap decoder, while the compact spatial evidence tokens S_t are sent to the selective latent memory module for causal state updating.

3) Selective Causal Latent Memory: Surgical attention evolves with both local motion continuity and long-horizon operative intent. During continuous manipulation, the attention state should change smoothly with instrument motion and tissue deformation; when the surgical intent shifts, it should also adapt to a new operative region. To capture these two properties, SurgLAT updates its latent attention state by selectively combining current visual evidence, short-term memory, and long-term memory in a causal manner.

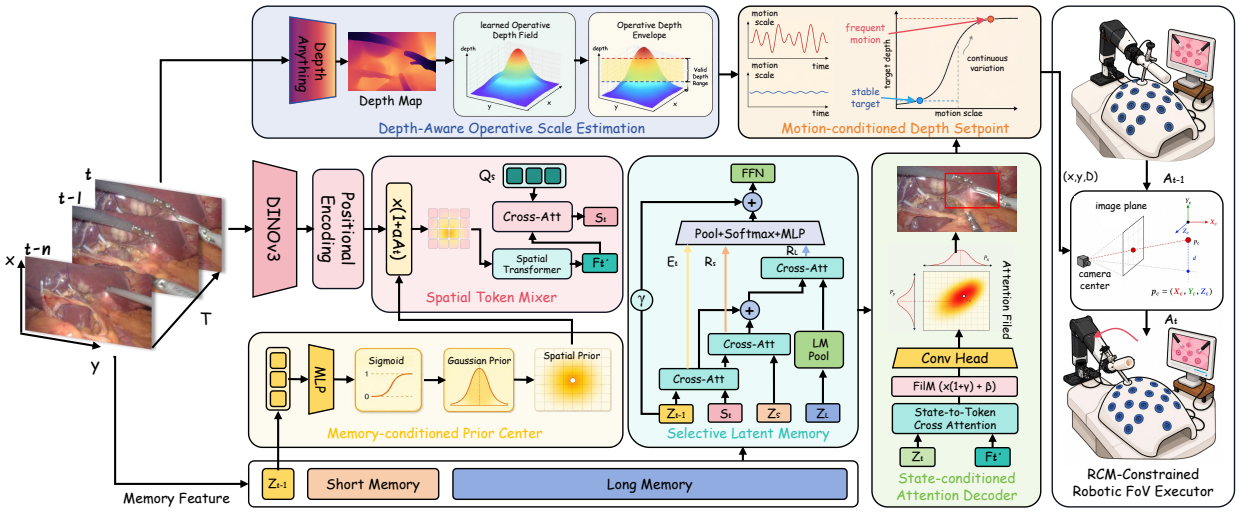

Fig. 2. SurgLAT predicts the image-plane operative target $(\boldsymbol{x}, \boldsymbol{y})$ from streaming laparoscopic video through memory-conditioned spatial reasoning and selective latent memory. In parallel, a depth-aware branch estimates the operative depth envelope, from which a motion-conditioned depth-domain regulation policy derives the viewing-direction depth target $\boldsymbol{D}$ for adaptive zoom-in/out control. The combined task-space target $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{D})$ is executed by an RCM-constrained robotic FoV executor for closed-loop laparoscopic viewpoint adjustment.

As shown in Fig. 2, the selective latent memory takes the previous state Z_{t-1}, the current spatial evidence tokens S_t, a short-term state cache $\mathcal{Z}_s$, and a long-term state cache $\mathcal{Z}_l$ as inputs. We first use learnable latent queries to cross-attend to S_t, aligning current spatial evidence with the latent state space and producing the current-frame evidence residual E_t. In parallel, the recent state cache $\mathcal{Z}_s$ is attended to obtain a short-term residual R_s, which captures local temporal continuity such as smooth instrument motion and small frame-to-frame shifts of the operative focus.

For long-term context, older latent states in $\mathcal{Z}_l$ are compressed by a long-memory pooling module into a compact set of memory prototypes. The current evidence, enhanced by the short-term residual, then queries these prototypes through cross-attention to obtain a long-term residual R_l. This branch provides transition-aware context and helps the model relax local continuity when the operative focus changes.

The three residual sources are fused adaptively rather than with fixed weights. A source router predicts the contribution of current evidence, short-term continuity, and long-term context:

$$[w_c, w_s, w_l] = \mathrm{Softmax}\left(\mathrm{MLP}([\bar{E}_t, \bar{R}_s, \bar{R}_l])\right), \tag{4}$$

where $\bar{E}_t$, $\bar{R}_s$, and $\bar{R}_l$ denote pooled representations of the three residual sources. The fused residual is computed as:

$$\Delta Z_t = w_c E_t + w_s R_s + w_l R_l. \tag{5}$$

The latent state is then updated with a residual formulation:

$$\hat{Z}_t = Z_{t-1} + \gamma \Delta Z_t, \qquad Z_t = \hat{Z}_t + \mathrm{FFN}(\hat{Z}_t), \tag{6}$$

where γ is a learnable residual scale initialized to 0.1 for stable early training.

This update is causal because it only uses the current spatial evidence and past latent states. The short-term memory supports stable attention tracking during continuous manipulation, while the long-term memory provides transition-aware context when the operative intent changes. The updated latent state Z_t is then used by the state-conditioned decoder to predict the attention heatmap and ROI.

4) State-Conditioned Heatmap and ROI Decoding: After the selective memory update, SurgLAT decodes the latent attention state into an attention heatmap and an operative ROI. Since surgical attention corresponds to a functional operative focus rather than an object with explicit boundaries, directly regressing all box coordinates (x, y, w, h) from a global feature can be unstable. We therefore decouple center localization from scale estimation: the attention center is inferred from a spatial heatmap, while a lightweight scale head predicts the ROI extent. As shown in Fig. 2, given the refined dense tokens F'_t and the updated latent state tokens Z_t, the decoder first conditions the visual grid on the current latent state using state-to-token cross-attention and FiLM-style channel modulation. The conditioned dense grid is then passed to a convolutional head to predict the attention heatmap H_t. This heatmap formulation preserves spatial layout and provides a smoother alternative to unconstrained coordinate regression.

To obtain a differentiable attention center, we marginalize the heatmap into horizontal and vertical distributions:

$$P_x = \mathrm{Softmax}(\mathrm{LogSumExp}_y(H_t)), \tag{7}$$

$$P_y = \mathrm{Softmax}(\mathrm{LogSumExp}_x(H_t)). \tag{8}$$

The center is then computed as the expectation of the two axis-wise distributions:

$$\hat{c}_t = (\mathbb{E}_{P_x}[x], \mathbb{E}_{P_y}[y]), \tag{9}$$

which avoids the discontinuity of hard argmax and yields continuous center coordinates from the spatial response.

For ROI scale estimation, a lightweight scale head predicts the width and height $\hat{s}_t = (\hat{w}_t, \hat{h}_t)$ conditioned on the latent attention state. The final attention box is constructed by combining the heatmap-derived center and the predicted scale:

$$\hat{B}_t = [\hat{x}_t - \hat{w}_t/2, \hat{y}_t - \hat{h}_t/2, \hat{x}_t + \hat{w}_t/2, \hat{y}_t + \hat{h}_t/2]. \tag{10}$$

This design allows the heatmap decoder to focus on stable operative-center localization, while the scale head adaptively estimates the spatial extent of the attention region.

5) Loss Function and Training Protocol: We train SurgLAT with heatmap-assisted box supervision. Since the attention center is the primary variable for laparoscope recentering, while the ROI extent is adaptively estimated by the scale head, the objective jointly supervises the attention heatmap, decoded center, predicted scale, and temporal center displacement. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{hm}} + \mathcal{L}_{\text{center}} + \mathcal{L}_{\text{scale}} + \mathcal{L}_{\text{temp}}. \tag{11}$$

The heatmap loss supervises the predicted attention field using a Gaussian target centered at the ground-truth center:

$$\mathcal{L}_{\text{hm}} = \text{BCEWithLogits}(H_t, G(c_t^*)), \tag{12}$$

where H_t denotes the predicted heatmap logits, c_t^* is the ground-truth center, and $G(c_t^*)$ is the corresponding Gaussian heatmap. The center loss further constrains both the axis-wise center distributions and the continuous expected center:

$$\mathcal{L}_{\text{center}} = \mathcal{L}_{\text{axis}} + \text{SmoothL1}(\hat{c}_t, c_t^*), \tag{13}$$

where $\mathcal{L}_{\text{axis}}$ is the cross-entropy loss over horizontal and vertical center bins, and $\hat{c}_t$ is the expected center decoded from the axis distributions.

To supervise the adaptive ROI extent, we apply a scale regression loss to the width and height predicted:

$$\mathcal{L}_{\text{scale}} = \text{SmoothL1}\left(\log(\hat{s}_t + \epsilon), \log(s_t^* + \epsilon)\right), \tag{14}$$

where $\hat{s}_t = (\hat{w}_t, \hat{h}_t)$ and $s_t^* = (w_t^*, h_t^*)$ denote the predicted and ground-truth ROI scales, respectively. The logarithmic form reduces sensitivity to absolute box size and stabilizes scale learning.

To encourage temporally coherent predictions, we regularize the predicted inter-frame center displacement:

$$\mathcal{L}_{\text{temp}} = \frac{1}{T-1} \sum_{t=2}^{T} \text{SmoothL1}\left(\hat{c}_t - \hat{c}_{t-1}, c_t^* - c_{t-1}^*\right). \tag{15}$$

This term penalizes inconsistent motion while allowing the prediction to follow the ground-truth attention trajectory. It improves temporal stability without altering the causal inference setting, since the model still predicts each frame using only the current observation and past latent states.

We adopt a two-stage training protocol to balance optimization stability and online consistency. In the first stage, SurgLAT is trained on short video clips to stabilize the spatial reader, latent memory update, state-conditioned heatmap decoder, and scale head on top of frozen DINO features. In the second stage, we switch to an online streaming protocol that matches deployment. Consecutive segments from the same video are processed in temporal order, and the latent memory is carried across segment boundaries instead of being reset at each clip. This streaming protocol aligns training with causal inference and enables stable memory propagation during continuous video prediction.

C. Depth-Aware Operative Scale Regulation

1) Attention-Weighted Operative Depth Estimation: The predicted attention center $(\hat{x}_t, \hat{y}_t)$ specifies the lateral recentering target, but does not determine the appropriate axial viewing scale. To provide a scale cue, we estimate the relative operative depth within the predicted attention region using a pretrained monocular depth model. Since monocular depth is scale-ambiguous, the predicted depth map D_t is normalized within each frame by robust percentile normalization and used only as a per-frame geometric cue rather than a metric camera-to-tissue distance. Given the predicted heatmap H_t, we compute an attention-weighted operative depth score:

$$d_t^{\text{op}} = \frac{\sum_{i,j} \alpha_t(i,j) \tilde{D}_t(i,j)}{\sum_{i,j} \alpha_t(i,j) + \epsilon}, \qquad \alpha_t(i,j) = \sigma(H_t(i,j)), \tag{16}$$

where $\tilde{D}_t$ denotes the normalized relative depth map. This score focuses the depth measurement on the operative region instead of the full image. From training videos, we estimate a valid operative depth envelope $\mathcal{E} = [d^{\text{low}}, d^{\text{high}}]$, which represents the typical range of normalized depth responses associated with stable laparoscopic viewing.

2) Motion-Conditioned Depth Setpoint: The desired axial scale depends on the recent motion of the attention target. A stable target allows a closer view for fine manipulation, while frequent target motion benefits from a wider FoV. We therefore compute a normalized motion scale from the recent attention trajectory:

$$m_t = \frac{1}{K d_{\text{img}}} \sum_{k=1}^{K} \|\hat{c}_{t-k+1} - \hat{c}_{t-k}\|_2, \tag{17}$$

where d_{img} is the image diagonal. The motion scale is mapped to a depth interpolation factor:

$$\gamma(m_t) = \text{clip}\left(\frac{m_t - m_{\min}}{m_{\max} - m_{\min} + \epsilon}, 0, 1\right). \tag{18}$$

The depth setpoint is then selected within the valid envelope:

$$d_t^{\text{set}} = d^{\text{low}} + \gamma(m_t)(d^{\text{high}} - d^{\text{low}}). \tag{19}$$

Thus, low target motion shifts the setpoint toward a closer operative view, whereas high target motion shifts it toward a wider view for robust tracking and exploration.

IV. Robotic Laparoscope Control

SurgLAT predicts an image-plane target and an axial depth setpoint for autonomous laparoscopic viewpoint adjustment. To execute these commands on a physical robot, the laparoscope must satisfy the Remote Center of Motion (RCM) constraint imposed by the trocar point p_{trocar}. In addition, the initial manipulator configuration strongly affects the feasible rotational workspace during subsequent tracking. We therefore combine RCM-constrained target tracking with redundancy-aware initial configuration optimization.

A. RCM-Constrained Target Tracking

1) RCM Modeling: To model insertion motion while maintaining the trocar constraint, we augment the 7-DoF manipulator configuration q with a virtual insertion coordinate λ. Let $p_{\mathrm{ee}}(q)$ and $p_{\mathrm{cam}}(q)$ denote two points on the laparoscope axis. The RCM point is defined as:

$$p_{\mathrm{rcm}}(q,\lambda) = p_{\mathrm{ee}}(q) + \lambda\big(p_{\mathrm{cam}}(q) - p_{\mathrm{ee}}(q)\big). \tag{20}$$

Since the laparoscope is rigidly attached to the manipulator, differentiating p_{rcm} gives $\dot{p}_{\mathrm{rcm}} = J_{\mathrm{rcm}}\begin{bmatrix}\dot{q}^T & \dot{\lambda}\end{bmatrix}^T$, where $J_{\mathrm{rcm}} = \begin{bmatrix}(1-\lambda)J_{p_{\mathrm{ee}}} + \lambda J_{p_{\mathrm{cam}}} & p_{\mathrm{cam}} - p_{\mathrm{ee}}\end{bmatrix}$. The RCM constraint is enforced by driving p_{rcm} to the fixed p_{trocar}.

2) Target-Oriented Laparoscope Rotation: The tracking objective is to keep the SurgLAT-predicted target region near the image center. Given the predicted image target r_{pix}, its estimated depth, camera intrinsic matrix K, and camera pose $T_{\mathrm{cam}}(q)$, we back-project the target to a world-frame point r_w. Let p_w be the RCM point, O_w the camera optical center, and a_w the current optical-axis direction. During target tracking, λ is kept fixed, so the task reduces to finding an incremental laparoscope rotation $R_\Delta = \exp([\omega]_\times)$. The rotated optical axis is $a'_w = R_\Delta a_w$, and the target direction from the rotated optical center is $b'_w = \frac{r_w - p_w - R_\Delta(O_w - p_w)}{\|r_w - p_w - R_\Delta(O_w - p_w)\|}$. We solve the following alignment objective:

$$\min_{\omega} \|a'_w \times b'_w\|^2 + \eta \max(0, -{a'_w}^T b'_w), \tag{21}$$

where the first term minimizes angular misalignment and the second term avoids the degenerate case where the optical axis points away from the target. The optimized rotation defines the desired laparoscope orientation as $R_{\mathrm{des}} = R^*_\Delta R_{\mathrm{cur}}$.

B. Initial Configuration Optimization

Because the laparoscope rotates during tracking, a favorable initial joint configuration can enlarge the rotational workspace. For a fixed initial laparoscope pose, the 7-DoF manipulator has one redundant DoF. Since axial rotation around the laparoscope does not change the viewing direction or violate the RCM constraint, we relax this axial constraint and search in a two-dimensional redundant space.

Let $p_{\mathrm{ee},0}$ and u_0 denote the initial end-effector position and laparoscope-axis direction. The relaxed 5-DoF initial pose constraint is $e_{\mathrm{ini}}(q) = \begin{bmatrix}p_{\mathrm{ee}}(q) - p_{\mathrm{ee},0} & u(q) \times u_0\end{bmatrix}^T$, $J_{\mathrm{ini}}(q) = \frac{\partial e_{\mathrm{ini}}}{\partial q}$. To evaluate the local rotational capability under the RCM constraint, we project the axis-direction Jacobian into the RCM null space: $J_{\mathrm{rot}}(q) = J_u(q) N^q_{\mathrm{rcm}}(q)$, where $J_u(q)$ is the Jacobian of the laparoscope-axis direction and N^q_{rcm} denotes the null-space of the RCM Jacobian.

The optimal initial configuration is obtained by maximizing the rotation score:

$$\begin{aligned} q^* = \arg\max_q \big(\sigma^+_{\min}(J_{\mathrm{rot}}) - w_\kappa \kappa(J_{\mathrm{rot}}) + w_m m(q)\big), \\ \text{s.t.} \quad e_{\mathrm{ini}}(q) = 0, \qquad q_{\min} \le q \le q_{\max}. \end{aligned} \tag{22}$$

Here, $\sigma^+_{\min}$ is the smallest non-zero singular value of J_{rot}, encouraging larger local rotational capability; $\kappa(\cdot)$ is the condition number, encouraging isotropic motion; and $m(q)$ is a joint-limit margin that avoids joint limits.

To reduce sensitivity to local optima, we first sample candidate seeds in the null space of $J_{\mathrm{ini}}(q_0)$. Using two null-space basis vectors n_1 and n_2, we generate $q_{\mathrm{seed}} = q_0 + \alpha n_1 + \beta n_2$, where $\alpha, \beta \in [-r, r]$. Each seed is projected back to the initial pose constraint by iterative correction:

$$q^{k+1} = \mathrm{clip}\left(q^k - J_{\mathrm{ini}}(q^k)^{\#} e_{\mathrm{ini}}(q^k), q_{\min}, q_{\max}\right). \tag{23}$$

The best projected candidate is then used as the initialization for solving Eq. (22), yielding the final configuration q^*.

C. Hierarchical Task-Space Control

Both initialization and online target tracking are executed using a hierarchical task-space controller. The primary task is solved first; the secondary task is projected into its null space:

$$\dot{\xi}^* = J^{\#}_{\mathrm{pr}} v_{\mathrm{pr}} + N_{\mathrm{pr}} \left(J_{\mathrm{s}} N_{\mathrm{pr}}\right)^{\#} \left(v_{\mathrm{s}} - J_{\mathrm{s}} J^{\#}_{\mathrm{pr}} v_{\mathrm{pr}}\right), \tag{24}$$

where $\xi = [q^T, \lambda]^T$, $N_{\mathrm{pr}} = I - J^{\#}_{\mathrm{pr}} J_{\mathrm{pr}}$, $v_{\mathrm{pr}} = K_{\mathrm{pr}} e_{\mathrm{pr}}$, and $v_{\mathrm{s}} = K_{\mathrm{s}} e_{\mathrm{s}}$. During initialization, the primary task enforces the RCM constraint, while the secondary task drives the manipulator toward q^*. During online tracking, the primary task jointly enforces the RCM constraint and the desired laparoscope orientation R_{des}, while the secondary task regulates insertion depth using $e_{\mathrm{s}} = \lambda_{\mathrm{target}} - \lambda$. This hierarchy ensures trocar safety is maintained while the robot follows the SurgLAT-predicted target and adaptive depth command.

V. Experiments

A. Experimental Setup and Details

1) Dataset and Baselines: We evaluate SurgLAT on the public SurgAtt-1.16M benchmark [9], which contains three subsets: SurgAtt-SZPH, SurgAtt-AutoLaparo, and SurgAtt-Hamlyn. SurgAtt-SZPH includes approximately 1M frames from gastrointestinal surgery and is used as the primary in-domain benchmark. We follow the original annotation protocol, where each frame is annotated with a bounding box indicating the operative attention region. We compare SurgLAT with representative tracking-based and detection-based baselines. Tracking-based methods include AQATrack [25], ODTrack [26], SPMTrack-B [27], LoRAT-B [28], LoRATv2-B [29], and MCITrack [30], which are adapted by treating the operative attention region as a temporally tracked target. Detection-based methods include YOLOv11-S/M [31], YOLOv12-S/M [32], YOLOv26-S/M [33], RT-DETR [34], RT-DETRv2 [35], and SurgAtt-Tracker [9], which directly predict the operative attention region from visual observations.

2) Evaluation Metrics: Following SurgAtt [9], we use Intersection over Union (IoU) to measure ROI overlap and Mean Center Error (MCE) to measure center localization error in pixels. On SurgAtt-SZPH, we additionally report $\mathrm{mAP}_{0.5}$, $\mathrm{mAP}_{0.75}$, and $\mathrm{mAP}_{0.5:0.95}$ to assess localization quality under different IoU thresholds. Runtime is measured in frames per second (FPS).

TABLE I

COMPARISON WITH DETECTION-BASED AND TRACKING-BASED BASELINES ON THREE SURGICAL ATTENTION TRACKING BENCHMARKS.

	SurgAtt-SZPH					SurgAtt-AutoLaparo		SurgAtt-Hamlyn		Runtime
Model	IoU↑	MCE↓	mAP@0.5 ↑	mAP@0.75 ↑	mAP@0.5 : 0.95 ↑	IoU↑	MCE↓	IoU↑	MCE↓	FPS↑
Detection-based										
YOLOv11-S [31]	0.470	79.45	0.372	0.040	0.111	0.283	193.93	0.311	69.24	41.5
YOLOv11-M [31]	0.471	82.00	0.370	0.040	0.112	0.285	221.66	0.239	80.71	40.0
YOLOv12-S [32]	0.471	80.79	0.347	0.035	0.099	0.292	203.64	0.322	67.90	39.9
YOLOv12-M [32]	0.480	79.12	0.396	0.043	0.119	0.304	207.36	0.328	68.20	42.0
YOLOv26-S [33]	0.471	84.45	0.315	0.043	0.102	0.316	211.64	0.235	81.79	43.2
YOLOv26-M [33]	0.466	87.89	0.328	0.042	0.106	0.313	215.88	0.323	65.57	40.3
RT-DETR [34]	0.499	77.80	0.452	0.058	0.148	0.349	162.84	0.239	52.54	**76.7**
RT-DETRv2 [35]	0.493	78.32	0.437	0.046	0.139	0.330	164.64	0.259	52.07	71.3
Object Tracking–based										
ODTrack [26]	0.413	77.20	0.304	0.026	0.084	0.320	186.30	0.328	51.27	25.0
SPMTrack-B [27]	0.486	79.51	0.524	0.079	0.178	0.403	161.02	0.406	46.27	26.7
LoRAT-B [28]	0.483	78.66	0.529	0.070	0.174	0.358	178.69	0.372	50.77	49.4
LoRATv2-B [29]	0.501	78.15	0.571	0.102	0.206	0.424	167.52	0.308	55.55	41.3
AQATrack [25]	0.528	76.56	0.602	0.162	0.246	0.417	167.69	0.418	47.91	25.7
MCITrack [30]	0.539	74.07	0.612	0.183	0.258	0.459	154.44	0.420	46.19	17.8
SurgAtt-Tracker [9]	0.566	49.92	0.656	0.220	0.280	0.462	121.12	0.443	42.48	12.4
SurgLAT (Ours)	**0.604**	**41.24**	**0.669**	**0.268**	**0.322**	**0.527**	**113.75**	**0.479**	**37.97**	34.5

3) Implementation Details: For fair comparison, all models are trained for 10 epochs on an NVIDIA H100 GPU. For baseline methods, we follow their official implementations. For detector-based baselines, we apply the same post-processing protocol to all models, with a confidence threshold of $\tau_{\text{conf}} = 0.001$ and an NMS IoU threshold of $\tau_{\text{nms}} = 0.25$. For IoU and MCE evaluation, each method is required to output one attention box per annotated frame. Tracking-based baselines are initialized with the ground-truth attention box in the first frame of each sequence and then evaluated on subsequent frames under the same per-frame averaging protocol. SurgLAT uses a frozen DINOv3 ViT-B/16 encoder [24]. Input frames are resized to 512×512, yielding a 32×32 token grid, and the extracted features are projected to a hidden dimension of $C = 256$. The spatial mixer uses 16 learnable evidence tokens, while the selective latent memory maintains 4 latent state tokens with short- and long-term caches of 16 and 64 previous states, respectively. Training follows a two-stage schedule: the first stage trains on short clips of length $T = 32$ for 5 epochs with AdamW, a learning rate of 2×10^{-5}, and weight decay of 10^{-4}; the second stage adopts online streaming training with segment length $T = 64$ for another 5 epochs and a learning rate of 1×10^{-5}. Consecutive segments from the same video are processed chronologically, and the latent memory is carried across segment boundaries to match causal inference.

B. Quantitative Evaluation of SurgAtt-1.16M Dataset

1) Overall Benchmark Comparison: Table I compares SurgLAT with representative detection-based and tracking-based baselines on three surgical attention tracking benchmarks. On the in-domain SurgAtt-SZPH benchmark, SurgLAT achieves the best performance, improving IoU from 0.566 to 0.604 and reducing MCE from 49.92 to 41.24 pixels compared with the strongest baseline, SurgAtt-Tracker. The improvement is also consistent under stricter localization criteria, where mAP@0.75 increases from 0.220 to 0.268 and mAP@0.5:0.95 from 0.280 to 0.322. Under cross-domain evaluation, SurgLAT improves IoU/MCE from 0.462/121.12 to 0.527/113.75 on SurgAtt-AutoLaparo and from 0.443/42.48 to 0.479/37.97 on SurgAtt-Hamlyn. These results demonstrate more accurate operative-focus localization and better generalization across surgical domains. Fig. 3 further shows that SurgLAT produces attention regions more consistently aligned with the ground-truth operative focus, especially under multi-instrument interference, cluttered tissue backgrounds, and target transitions.

The performance gap reflects the limitation of treating surgical attention as either a frame-wise detection target or a conventional tracked object. Detection-based methods are efficient but lack explicit temporal state modeling, while tracking-based methods exploit temporal continuity but assume a persistent target. In laparoscopic surgery, however, the operative focus may shift among instruments, tissues, anatomical boundaries, and tool–tissue interaction regions. By modeling attention as a causal latent state, SurgLAT preserves temporal continuity during stable manipulation while adapting to attention transitions when the operative intent changes.

2) Temporal Accuracy and Stability: The temporal evaluation further supports the stability of SurgLAT. As shown in Fig. 4, SurgLAT maintains the lowest normalized center error over the full normalized video progress. This indicates that the proposed model provides stable online attention localization throughout the procedure, rather than achieving improvement only on isolated frames. Such temporal consistency is important for autonomous laparoscope control, where transient prediction errors may be directly converted into unnecessary camera motion. Fig. 5 compares the motion-aware temporal instability across validation videos. SurgLAT achieves the lowest and most compact instability distribution among the compared methods, demonstrating smoother and more ground-truth-consistent attention prediction. This result is particularly relevant to robotic field-of-view control, because unstable attention predictions can lead to jittery image-plane commands and oscillatory camera adjustment.

C. Real-World Robotic Validation

1) Closed-Loop Laparoscope Control on Ex Vivo Tasks: We further validate SurgLAT on a physical laparoscopic robotic platform under four ex vivo manipulation settings:

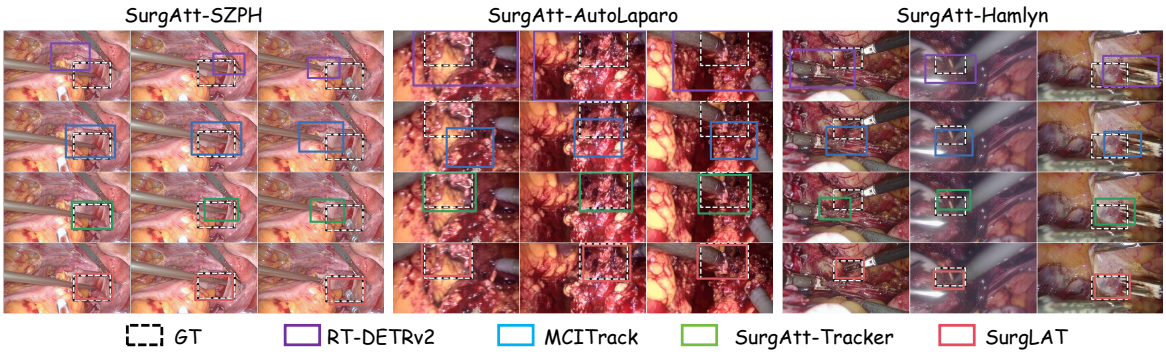

Fig. 3. Qualitative comparison of surgical attention localization on SurgAtt-1.16M. SurgLAT shows more consistent alignment with the operative focus across in-domain and cross-domain surgical scenes.

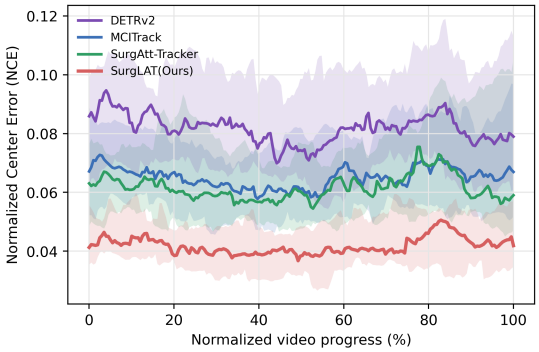

Fig. 4. Temporal comparison of NCE across validation videos. Lower values indicate more accurate attention localization over time.

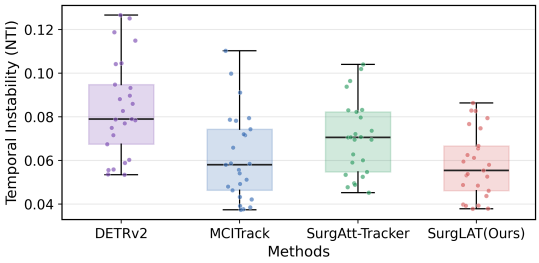

Fig. 5. Comparison of motion-aware temporal instability across 25 validation videos. Lower values indicate smoother and more GT-consistent attention prediction.

trajectory following, target reaching, tissue manipulation, and tissue manipulation under simulated bleeding. As shown in Fig. 7, the predicted attention targets are converted into closed-loop camera control commands and evaluated by center-reference IoU and trajectory smoothness. SurgLAT consistently achieves the best performance across all settings. Compared with SurgAtt-Tracker, SurgLAT improves IoU/smoothness from 0.490/0.186 to 0.520/0.238 in trajectory following, from 0.423/0.174 to 0.503/0.204 in target reaching, from 0.383/0.146 to 0.498/0.154 in tissue manipulation, and from 0.361/0.106 to 0.484/0.127 under simulated bleeding. These results indicate that the proposed latent attention model provides more reliable visual targets for stable closed-loop FoV adjustment across both phantom and ex vivo tissue manipulation scenarios.

2) RCM-Constrained Initial Configuration Optimization: As shown in Fig. 6, compared with the initial configuration q_0, the optimized configuration q^* increases the rotation score

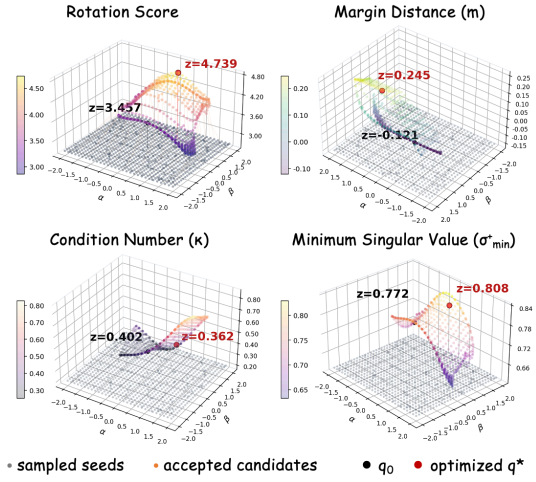

Fig. 6. Visualization of the null-space search for initial configuration optimization. $\boldsymbol{q}^*$ achieves improved rotational capability and a larger joint-limit margin while preserving comparable local conditioning properties.

TABLE II
ABLATION STUDY OF THE MEMORY-CONDITIONED SPATIAL PRIOR AND SELECTIVE LATENT MEMORY COMPONENTS.

Module			Metrics		
Prior Center	Short Memory	Long Memory	IoU↑	MCE↓	mAP@0.5↑
×	×	×	0.526	54.87	0.604
✓	×	×	0.571	48.03	0.638
✓	✓	×	0.594	43.72	0.661
✓	×	✓	0.586	45.26	0.653
✓	✓	✓	0.604	41.24	0.669

from 3.457 to 4.739 and improves the joint-limit margin from -0.121 to 0.245. This indicates a larger usable rotational workspace and increased clearance from joint limits. Meanwhile, $\sigma^+_{\min}$ increases from 0.772 to 0.808, and κ decreases from 0.402 to 0.362, suggesting that the optimization improves rotational feasibility with increased local kinematic conditioning. These results show that the proposed null-space search effectively exploits redundancy to obtain a more favorable initial configuration for RCM-constrained laparoscope tracking.

D. Ablation Study

1) Effect of Latent Memory and Spatial Prior: We first analyze the contribution of the memory-conditioned spatial prior and latent memory components on SurgAtt-SZPH. As shown in Table II, removing all prior and memory modules leads to a clear performance drop, with an IoU of 0.526 and an MCE

Models	Phantom Task 1	Phantom Task 2	Tissue Task 1	Tissue Task 2
DETRv2	0.286 / 0.095	0.252 / 0.101	0.212 / 0.090	0.170 / 0.078
MCITrack	0.389 / 0.153	0.363 / 0.126	0.304 / 0.113	0.281 / 0.091
SurgAtt-Tracker	0.490 / 0.186	0.423 / 0.174	0.383 / 0.146	0.361 / 0.106
SurgVLT	0.520 / 0.238	0.503 / 0.204	0.498 / 0.154	0.484 / 0.127

Fig. 7. Ex vivo validation of algorithm-driven laparoscopic control across phantom and tissue manipulation tasks. Blue values denote center-reference IoU, and red values denote trajectory smoothness.

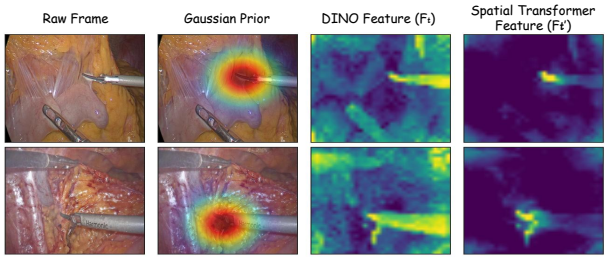

Fig. 8. Visualization of memory-conditioned spatial reasoning. The Gaussian prior provides a coarse focus, and the spatial transformer refines broad DINO responses into a more concentrated feature activation around the operative interaction region.

TABLE III
ABLATION STUDY OF DIFFERENT HEATMAP SIZES AND CENTER DECODING STRATEGIES.

Module		Metrics		
Heatmap Size	Center Decoding	IoU↑	MCE↓	FPS↑
–	Direct center regression	0.548	53.12	**48.6**
16×16	Axis soft-argmax	0.581	46.87	41.2
64×64	Axis soft-argmax	**0.608**	**40.92**	22.8
32×32	Full soft-argmax	0.592	44.71	31.4
32×32	Soft-argmax	0.598	43.18	33.1
32×32	Axis soft-argmax	0.604	41.24	34.5

of 54.87 pixels. Introducing the prior center improves IoU to 0.571 and reduces MCE to 48.03, suggesting that a coarse memory-conditioned spatial cue helps guide the model toward the operative region. Adding short-term memory further improves IoU to 0.594 and MCE to 43.72 pixels, showing its importance for local motion continuity. Long-term memory also brings consistent gains, reaching 0.586 IoU and 45.26 MCE when combined with the prior center. The full model achieves the best performance, with 0.604 IoU, 41.24 MCE, and 0.669 mAP@0.5, indicating that spatial prior, short-term continuity, and long-term intent memory are complementary. Fig. 8 provides a qualitative explanation of this improvement. The Gaussian prior offers a coarse focus around the expected operative region, while the DINO feature map captures broad semantic responses from instruments and tissues. After the spatial transformer, the refined feature response becomes more concentrated around the operative interaction region, supporting more accurate attention localization.

2) Heatmap Resolution and Center Decoding: We further evaluate the heatmap resolution and center decoding strategy in Table III. Direct center regression is computationally efficient but yields substantially lower localization accuracy, confirming that surgical attention is better represented as a spatial probability field rather than an unconstrained coordinate vector. Increasing the heatmap size from 16×16 to 64×64 improves accuracy but significantly reduces runtime. The 32×32 heatmap with axis soft-argmax provides the best accuracy–efficiency trade-off, achieving 0.604 IoU and 34.5 FPS. Compared with full soft-argmax, axis soft-argmax produces lower MCE while preserving real-time performance, suggesting that factorized center decoding provides a stable and efficient approximation for online attention localization.

TABLE IV
ABLATION STUDY OF HEATMAP, CENTER, AND TEMPORAL SUPERVISION IN SURGICAL ATTENTION LOCALIZATION.

Loss			Metrics		
$\mathcal{L}_{\text{hm}}$	$\mathcal{L}_{\text{center}}$	$\mathcal{L}_{\text{temp}}$	IoU↑	MCE↓	mAP@0.5↑
✓	×	×	0.575	48.25	0.634
✓	×	✓	0.583	45.72	0.642
✓	✓	×	0.597	43.90	0.663
✓	✓	✓	**0.604**	**41.24**	**0.669**

3) Effect of Loss Components: Table IV evaluates the contribution of loss components. Using only heatmap supervision achieves 0.575 IoU and 48.25 MCE, indicating that heatmap-level supervision alone is insufficient for precise center localization. Adding the center loss improves the performance to 0.597 IoU and 43.90 MCE, showing that directly constraining the decoded attention center is important for reducing localization error. When the temporal displacement loss is further introduced, the full objective achieves the best results, with 0.604 IoU, 41.24 MCE, and 0.669 mAP@0.5. These results confirm that heatmap supervision, center-level constraint, and temporal regularization provide complementary benefits for improved surgical attention prediction.

VI. CONCLUSION

In this work, we presented SurgLAT, a perception-to-control framework for autonomous robotic laparoscope control. By modeling surgical attention as a causal latent state, SurgLAT integrates memory-conditioned spatial reasoning, selective short- and long-term latent memory, and heatmap-assisted ROI decoding to produce accurate and temporally stable operative

targets from streaming laparoscopic video. The predicted attention target is further combined with depth-aware operative scale estimation and executed through a virtual-axis-based RCM-constrained controller with redundancy-aware null-space initialization. Experiments on SurgAtt-1.16M demonstrate improved attention localization, temporal stability, and cross-domain generalization over detection- and tracking-based baselines, while ex vivo robotic validation shows stable closed-loop FoV adjustment across phantom and tissue manipulation tasks. Although the current robotic experiments are conducted mainly in phantom and ex vivo settings, future work will extend SurgLAT toward larger-scale in vivo validation, more complex surgical workflows, and stronger safety-aware control constraints for clinical deployment.

REFERENCES

[1] T. Haidegger, S. Speidel, D. Stoyanov, and R. M. Satava, "Robot-assisted minimally invasive surgery—surgical robotics in the data age," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 835–846, 2022.

[2] P. E. Dupont, B. J. Nelson, M. Goldfarb, B. Hannaford, A. Menciassi, M. K. O'Malley, N. Simaan, P. Valdastri, and G.-Z. Yang, "A decade retrospective of medical robotics research from 2010 to 2020," *Science robotics*, vol. 6, no. 60, p. eabi8017, 2021.

[3] S. Merola, P. Weber, A. Wasielewski, and G. H. Ballantyne, "Comparison of laparoscopic colectomy with and without the aid of a robotic camera holder," *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, vol. 12, no. 1, pp. 46–51, 2002.

[4] M. Wagner, A. Bihlmaier, H. G. Kenngott, P. Mietkowski, P. M. Scheikl, S. Bodenstedt, A. Schiepe-Tiska, J. Vetter, F. Nickel, S. Speidel *et al.*, "A learning robot for cognitive camera control in minimally invasive surgery," *Surgical Endoscopy*, vol. 35, no. 9, pp. 5365–5374, 2021.

[5] K. Fujii, A. Salerno, K. Sriskandarajah, K.-W. Kwok, K. Shetty, and G.-Z. Yang, "Gaze contingent cartesian control of a robotic arm for laparoscopic surgery," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 3582–3589.

[6] T. Osa, C. Staub, and A. Knoll, "Framework of automatic robot surgery system using visual servoing," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 1837–1842.

[7] B. Yang, W. Chen, Z. Wang, Y. Lu, J. Mao, H. Wang, and Y.-H. Liu, "Adaptive fov control of laparoscopes with programmable composed constraints," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, no. 4, pp. 206–217, 2019.

[8] Y. Huang, J. Li, X. Zhang, K. Xie, J. Li, Y. Liu, C. S. H. Ng, P. W. Y. Chiu, and Z. Li, "A surgeon preference-guided autonomous instrument tracking method with a robotic flexible endoscope based on dvrk platform," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2250–2257, 2022.

[9] R. Zhou, G. Wang, A. Wang, Y. Ma, L. Ouyang, B. Cui, J. Li, C. Zhu, M. Li, M. Chen, X. Zhong, P. Lu, J. Wang, X. Liu, and H. Ren, "Surgatt-tracker: Online surgical attention tracking via temporal proposal reranking and motion-aware refinement," 2026. [Online]. Available: https://arxiv.org/abs/2602.20636

[10] J. Sandoval, M. A. Laribi, J.-P. Faure, C. Brèque, J.-P. Richer, and S. Zeghloul, "Towards an autonomous robot-assistant for laparoscopy using exteroceptive sensors: Feasibility study and implementation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6473–6480, 2021.

[11] B. Li, R. Wei, J. Xu, B. Lu, C. H. Yee, C. F. Ng, P.-A. Heng, Q. Dou, and Y.-H. Liu, "3d perception based imitation learning under limited demonstration for laparoscope control in robotic surgery," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7664–7670.

[12] C. Gruijthuijsen, L. C. Garcia-Peraza-Herrera, G. Borghesan, D. Reynaerts, J. Deprest, S. Ourselin, T. Vercauteren, and E. Vander Poorten, "Robotic endoscope control via autonomous instrument tracking," *Frontiers in Robotics and AI*, vol. 9, p. 832208, 2022.

[13] B. Li, B. Lu, Y. Lu, Q. Dou, and Y.-H. Liu, "Data-driven holistic framework for automated laparoscope optimal view control with learning-based depth perception," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 366–12 372.

[14] B. Li, Y. Lu, W. Chen, B. Lu, F. Zhong, Q. Dou, and Y.-H. Liu, "Gmm-based heuristic decision framework for safe automated laparoscope control," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1969–1976, 2024.

[15] E. Iovene, D. Cattaneo, J. Fu, G. Ferrigno, and E. De Momi, "Hybrid tracking module for real-time tool tracking for an autonomous exoscope," *IEEE Robotics and Automation Letters*, vol. 9, no. 7, pp. 6067–6074, 2024.

[16] J. Wang, Y. Wei, L. Jiang, X. Guo, A. Zheng, W. Zhao, and Z. Li, "Visionsafeenhanced vpc: Cautious predictive control with visibility constraints under uncertainty for autonomous robotic surgery," *IEEE Robotics and Automation Letters*, 2026.

[17] X. Ma, C. Song, P. W. Chiu, and Z. Li, "Visual servo of a 6-dof robotic stereo flexible endoscope based on da vinci research kit (dvrk) system," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 820–827, 2020.

[18] C. Zhang, W. Zhu, J. Peng, Y. Han, and W. Liu, "Visual servo control of endoscope-holding robot based on multi-objective optimization: System modeling and instrument tracking," *Measurement*, vol. 211, p. 112658, 2023.

[19] A. Bihlmaier and H. Woern, "Automated endoscopic camera guidance: A knowledge-based system towards robot assisted surgery," in *ISR/Robotik 2014; 41st International Symposium on Robotics*. VDE, 2014, pp. 1–6.

[20] Y. Sun, B. Pan, Y. Fu, and G. Niu, "Visual-based autonomous field of view control of laparoscope with safety-rcm constraints for semi-autonomous surgery," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 16, no. 2, p. e2079, 2020.

[21] K. Fozilov, J. Colan, A. Davila, K. Misawa, J. Qiu, Y. Hayashi, K. Mori, and Y. Hasegawa, "Endoscope automation framework with hierarchical control and interactive perception for multi-tool tracking in minimally invasive surgery," *Sensors*, vol. 23, no. 24, p. 9865, 2023.

[22] Z. Jiang, J. Li, and H. Zheng, "Robust predictive visual servoing of usvs with wave perturbations considering fov constraint," *IEEE Transactions on Industrial Electronics*, vol. 72, no. 8, pp. 8279–8289, 2025.

[23] N. Pasini, A. Mariani, A. Deguet, P. Kazanzides, and E. De Momi, "Grace: Online gesture recognition for autonomous camera-motion enhancement in robot-assisted surgery," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8263–8270, 2023.

[24] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[25] J. Xie, B. Zhong, Z. Mo, S. Zhang, L. Shi, S. Song, and R. Ji, "Autoregressive queries for adaptive tracking with spatio-temporal transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 300–19 309.

[26] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, and X. Li, "Odtrack: Online dense temporal token learning for visual tracking," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 7, 2024, pp. 7588–7596.

[27] W. Cai, Q. Liu, and Y. Wang, "Spmtrack: Spatio-temporal parameter-efficient fine-tuning with mixture of experts for scalable visual tracking," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16 871–16 881.

[28] L. Lin, H. Fan, Z. Zhang, Y. Wang, Y. Xu, and H. Ling, "Tracking meets lora: Faster training, larger model, stronger performance," in *ECCV*, 2024.

[29] L. Lin, H. Fan, Z. Zhang, Y. Huang, Y. Wang, Y. Xu, and H. Ling, "Loratv2: Enabling low-cost temporal modeling in one-stream trackers," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[30] B. Kang, X. Chen, S. Lai, Y. Liu, Y. Liu, and D. Wang, "Exploring enhanced contextual information for video-level object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 4194–4202.

[31] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[32] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[33] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "Yolo26: key architectural enhancements and performance benchmarking for real-time object detection," *arXiv preprint arXiv:2509.25164*, 2025.

[34] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," 2023.

[35] W. Lv, Y. Zhao, Q. Chang, K. Huang, G. Wang, and Y. Liu, "Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer," 2024. [Online]. Available: https://arxiv.org/abs/2407.17140