Bridging Vision and Language for Robust Context-Aware Surgical Point Tracking: The VL-SurgPT Dataset and Benchmark

Rulin Zhou[1,3,5*], **Wenlong He**[4*], **An Wang**[1*], **Jianhang Zhang**[4], **Xuanhui Zeng**[4], **Xi Zhang**[4], **Chaowei Zhu**,[2], **Haijun Hu**,[2†], **Hongliang Ren**[1,5†]

[1]The Chinese University of Hong Kong
[2]Division of Gastrointestinal Surgery, Shenzhen People's Hospital
[3]The University of Hong Kong
[4]Shenzhen University
[5]The Chinese University of Hong Kong, Shenzhen Research Institute
zhourulin@connect.hku.hk, wa09@link.cuhk.edu.hk, hu.haijun@szhospital.com, hlren@ee.cuhk.edu.hk

Abstract

Accurate point tracking in surgical environments remains challenging due to complex visual conditions, including smoke occlusion, specular reflections, and tissue deformation. While existing surgical tracking datasets provide coordinate information, they lack the semantic context necessary to understand tracking failure mechanisms. We introduce **VL-SurgPT**, the first large-scale multimodal dataset that bridges visual tracking with textual descriptions of point status in surgical scenes. The dataset comprises 908 *in vivo* video clips, including 754 for tissue tracking (17,171 annotated points across five challenging scenarios) and 154 for instrument tracking (covering seven instrument types with detailed keypoint annotations). We establish comprehensive benchmarks using eight state-of-the-art tracking methods and propose **TG-SurgPT**, a text-guided tracking approach that leverages semantic descriptions to improve robustness in visually challenging conditions. Experimental results demonstrate that incorporating point status information significantly improves tracking accuracy and reliability, particularly in adverse visual scenarios where conventional vision-only methods struggle. By bridging visual and linguistic modalities, VL-SurgPT enables the development of context-aware tracking systems crucial for advancing computer-assisted surgery applications that can maintain performance even under challenging intraoperative conditions.

Project — https://szupc.github.io/VL-SurgPT/

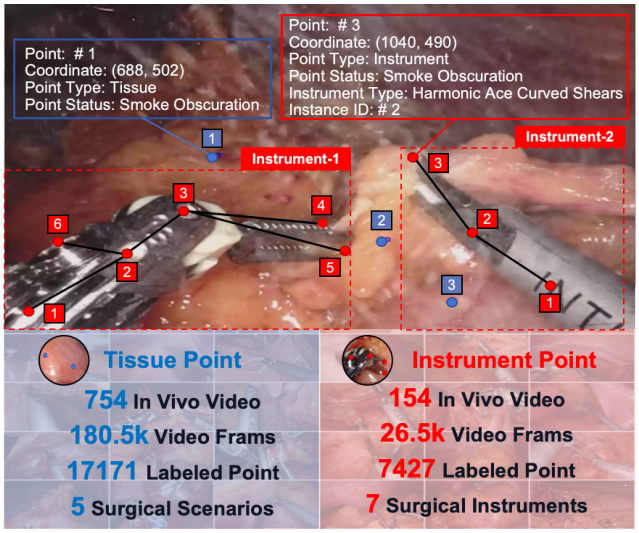

Figure 1: Overview of our Vision-Language Surgical Point Tracking (VL-SurgPT) dataset, a large-scale multimodal dataset containing visual and textual annotations for tissue and instrument points across diverse challenging scenarios.

Introduction

Accurate point tracking in surgical environments is critical for advancing computer-assisted interventions, enabling applications like motion understanding and scene perception (Schmidt et al. 2024a,b; Wu et al. 2025). Unlike general computer vision scenarios where point tracking has achieved remarkable success (Zheng et al. 2023; Doersch et al. 2022), surgical environments introduce extraordinary visual challenges, including dense electrocautery smoke, specular reflections from wet anatomical surfaces, dynamic instrument occlusions, and substantial tissue deformation. These challenges severely compromise the tracking performance of general tracking approaches (Doersch et al. 2023, 2024; Karaev et al. 2024b).

Existing tracking datasets treat point tracking as purely geometric problems with visual annotations alone. General benchmarks like TAP-Vid (Doersch et al. 2022) and PointOdyssey (Zheng et al. 2023), along with surgical-specific datasets like SurgT (Cartucho et al. 2024), STIR (Schmidt et al. 2024b), and SurgPose (Wu et al. 2025), provide visual data but lack semantic context explaining why tracking fails under specific conditions.

Furthermore, these datasets typically focus on either tis-

*These authors contributed equally.
†Corresponding authors.

sue or instrument tracking in isolation, failing to capture the holistic nature of surgical procedures where both elements interact dynamically. Recent efforts like SurgMotion (Zhan et al. 2024) began addressing this gap but remained unimodal, providing only coordinate annotations without semantic descriptions of visual conditions. This semantic gap represents a critical barrier to developing robust surgical tracking systems. Existing datasets overlook the point status where rich contextual information describing both the point's condition and the surrounding environmental state. We hypothesize that incorporating semantic descriptions of visual conditions can substantially enhance tracking performance by providing contextual guidance that pure visual methods cannot capture.

To address this, we introduce **VL-SurgPT**, the first comprehensive multimodal dataset that bridges visual point tracking with semantic descriptions of environmental conditions. As shown in Fig. 1, our dataset combines coordinate annotations with synchronized textual descriptions characterizing each point's visual status, enabling fine-grained analysis of tracking behavior and developing context-aware methods. Our contributions include: **(1) VL-SurgPT**, a large-scale multimodal dataset containing visual and textual annotations for surgical point tracking across diverse challenging scenarios; **(2)** comprehensive benchmarks evaluating eight tracking methods with condition-specific performance analysis; and **(3) TG-SurgPT**, a text-guided tracking approach that leverages semantic descriptions to achieve superior performance, particularly under adverse visual conditions where conventional methods struggle.

Related Works

Surgical Point Tracking Datasets

Existing surgical tracking datasets have primarily focused on either tissue points or instrument keypoints. Early **tissue tracking** datasets, such as SuPer (Li et al. 2020) and Semantic SuPer (Lin et al. 2023), were largely confined to *ex vivo* settings and limited annotation quantities. SurgT (Cartucho et al. 2024) curated *in vivo* surgical tissue tracking dataset but only focused on bounding box tracking rather than precise, continuous point trajectories. STIR (Schmidt et al. 2024b) advanced the field by utilizing infrared fluorescent markers for more accurate point localization in both *in vivo* and *ex vivo* scenarios; however, its annotations were typically sparse and endpoint-focused. Regarding **instrument keypoint tracking** datasets, RMIT (Sznitman et al. 2012) and the EndoVis15 dataset (Bodenstedt et al. 2018) provided instrument keypoint labels, predominantly from *ex vivo* setups with limited semantic depth regarding the visual status of the keypoints. SurgPose (Wu et al. 2025) improved annotation efficiency for *ex vivo* data using UV markers, primarily targeting 6DoF pose estimation rather than direct, continuous keypoint tracking. Recently, SurgMotion (Zhan et al. 2024) offered comprehensive annotations for tissue and instrument point tracking, yet it adhered to a unimodal, vision-only annotation paradigm.

A critical limitation across these datasets is their unimodal nature: while they provide geometric coordinates, they lack synchronized semantic descriptions of visual conditions (e.g., smoke, reflection, occlusion) that directly impact tracking performance in complex *in vivo* surgical environments. This absence of multimodal information hinders the development and evaluation of tracking algorithms that can understand and adapt to these specific visual challenges. VL-SurgPT addresses this gap by offering rich textual annotations of point status in conjunction with precise spatial coordinates.

Surgical Point Tracking Methods

Recent advances in point tracking like TAPIR (Doersch et al. 2023), BootsTAP (Doersch et al. 2024), CoTracker (Karaev et al. 2024b,a), SEARAFT (Wang, Lipson, and Deng 2024), MFT (Neoral, Šerỳch, and Matas 2024), MFTIQ (Serych, Neoral, and Matas 2025), and Track-On (Aydemir et al. 2025) demonstrate strong performance on general tracking benchmarks like TAP-Vid (Doersch et al. 2022). Surgical-specific tracking methods have been developed to address domain challenges. Semantic SuPer (Lin et al. 2023) incorporated geometric and semantic cues for tissue tracking. SENDD (Schmidt et al. 2023) used graph neural networks for sparse keypoint matching. More recently, AdaTracker (Guo et al. 2024) and Endo-TTAP (Zhou et al. 2025) showed promising results on surgical datasets through adaptive matching and attention mechanisms, respectively. Besides, SurgMotion (Zhan et al. 2024) proposes to track surgical points of tissue and instruments simultaneously with mask and As Rigid As Possible (ARAP) constraints, and demonstrated strong performance on tissue tracking tasks. However, these methods still rely solely on visual cues and do not leverage semantic information to enhance robustness under challenging conditions. This limitation motivates our approach of providing explicit visual status annotations to enable condition-specific analysis and method development.

Vision-Language in Surgical Applications

Vision-Language (VL) integration has emerged as a powerful paradigm in surgical computer vision (Min, Lai, and Ren 2025), advancing video pretraining (Yuan et al. 2024b,a), scene captioning (Xu et al. 2021; Xu, Islam, and Ren 2022), and visual question answering (Seenivasan et al. 2022, 2023; Bai et al. 2025). However, this integration remains largely unexplored for low-level tasks like point tracking. Current surgical VL (Zeng et al. 2025) approaches typically interpret visible elements rather than leveraging language to enhance context awareness under challenging visual conditions.

Our work, **VL-SurgPT**, addresses this gap by systematically integrating vision and language for surgical point tracking. By providing synchronized visual coordinates with textual status descriptions and developing text-guided tracking methodology, we establish a foundation for semantically-aware tracking systems that overcome purely visual limitations. This opens new avenues for developing more reliable and interpretable tools for computer-assisted interventions.

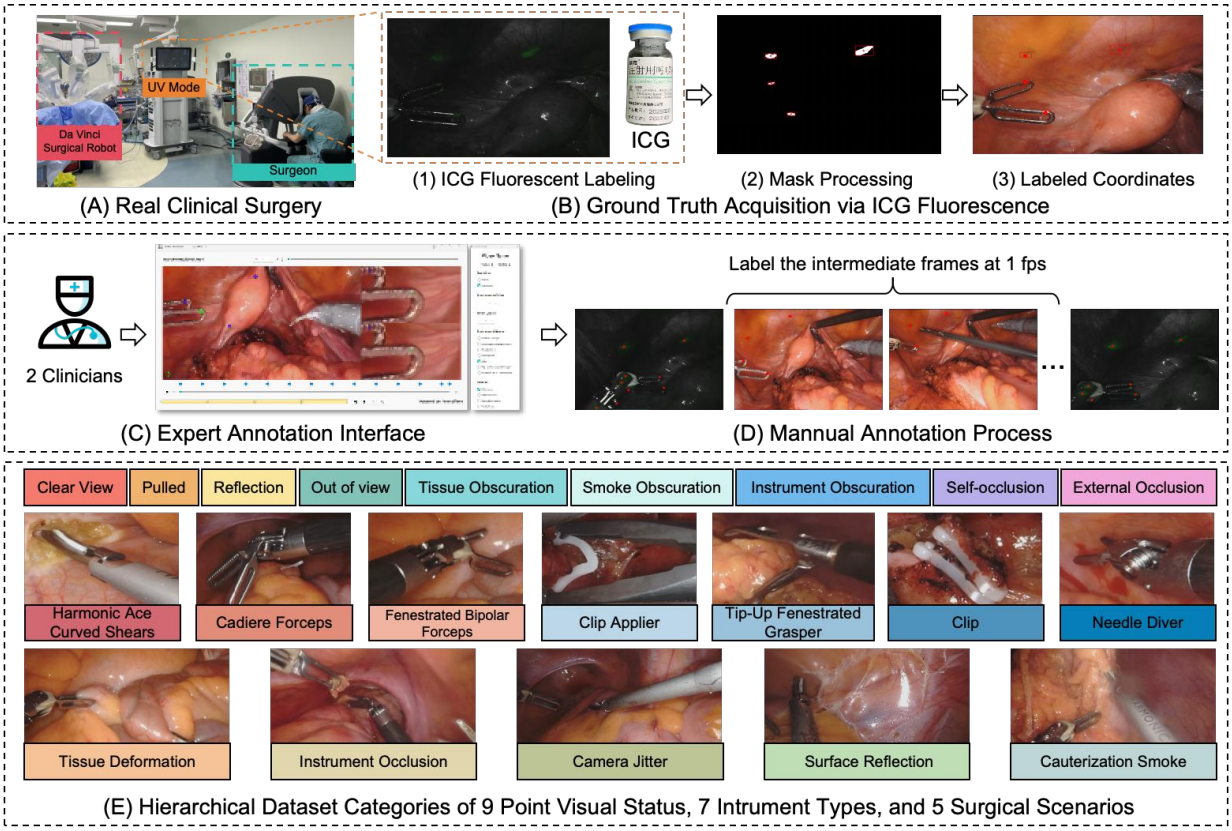

Figure 2: Data collection and annotation workflow for VL-SurgPT. (A) In vivo surgical setup using the da Vinci Xi system. (B) Ground truth acquisition using Indocyanine Green (ICG) fluorescent markers under UV illumination. (C-D) Annotation interface for point tracking and semantic labeling at 1 fps. (E) Coverage of 7 types of surgical instruments, 9 distinct visual status descriptions, and 5 representative challenging scenarios across our dataset.

VL-SurgPT Dataset

The VL-SurgPT dataset was meticulously constructed from real-world surgical procedures, involving a multi-stage process of data acquisition, processing, and comprehensive multimodal annotation, as illustrated in Fig. 2.

Data Collection

The VL-SurgPT dataset originates from *in vivo* robotic surgical procedures conducted at the Department of Gastrointestinal Surgery of the Shenzhen People's Hospital[1]. All procedures, illustrated in Fig. 2 (A), were performed using the da Vinci Xi Surgical System. The source footage encompasses a range of complex operations, including radical gastrectomy, radical resection of gastrointestinal tumors, radical resection of rectal cancer, and radical colectomy. Initially, data from 20 distinct surgeries, totaling approximately 33 hours of video, were collected. This raw footage underwent a rigorous quality screening process, prioritizing video clarity, minimal visual artifacts, and procedural consistency relevant to point tracking challenges. Following this curation, the final dataset comprises 115 minutes of high-quality surgical video selected for detailed annotation.

Data Processing and Annotation

The creation of VL-SurgPT involved a detailed two-stage data processing and annotation pipeline, combining intraoperative Indocyanine Green (ICG) marking with postoperative expert manual labeling, as depicted in Fig. 2 (B)-(D).

Intraoperative ICG Marking and Ground Truth Acquisition. To establish reliable ground truth for point locations, ICG dye, visible under near-infrared (NIR) or UV fluorescence imaging, was utilized. *For tissue points:* During surgery, a clinician applied ICG dye to 2-4 discrete locations on the tissue surface using an ICG-tipped needle holder. *For instrument keypoints:* Before surgical use, ICG dye was applied to predefined keypoints (e.g., tip, joints, shaft) on seven types of surgical instruments, typically marking 2–7 keypoints per instrument.

Recognizing the dynamic nature of *in vivo* surgery, which precludes the controlled repetitions possible in *ex vivo* setups (e.g., SurgPose (Wu et al. 2025)), we adopted a strategy to capture ground truth at the beginning and end of short video clips. The protocol involved (see Fig. 2 (B)): (1) Acti-

[1]All data collection protocols were approved by No. LL-KY-2023121-01 with appropriate patient consent and privacy protection.

vating the da Vinci system's UV/fluorescence mode with instruments and tissues held static, allowing for clear recording of ICG-marked keypoint coordinates. (2) Switching to standard white light and performing normal surgical maneuvers for 5–10 seconds. (3) Re-activating UV/fluorescence mode with static positioning to record the endpoint ICG coordinates. These ICG-derived coordinates from the first and last frames of each short clip, i.e., the **query points** and the **end points**, serve as high-fidelity ground truth anchors. Fluorescent markers were extracted by isolating activated ICG regions via binary thresholding under UV illumination.

Postoperative Manual Annotation and Semantic Labeling. Following *in vivo* data acquisition, the collected video clips were manually annotated by two experienced clinical doctors using a custom-designed labeling tool (Fig. 2 (C)-(D)). For each clip: 1. The ICG-marked points in the first frame served as initial query points for tracking. 2. Clinicians manually tracked these points frame-by-frame through the white-light portion of the clip, annotating their 2D coordinates at 1 frame per second (fps). The ICG-marked points in the final frame provided a reference to validate the trajectory endpoint. 3. Crucially, alongside coordinate data, each annotated point was assigned multiple semantic labels:

- A **point type** label (tissue or instrument).
- A **point status** textual description (e.g., "Clear View", "Smoke Obscuration", "Pulled", "Reflection", etc., from a predefined vocabulary shown in Fig. 2 (E)).
- For instrument points, additional labels included an **instance identifier** and the **specific instrument type**.

Dataset Details

VL-SurgPT provides a substantial collection of *in vivo* surgical tracking data spanning diverse clinical scenarios, as summarized in Fig. 1. Illustrative examples of these scenarios and instruments are visualized in Fig. 2 (E).

Tissue Tracking Subset: This subset consists of 754 *in vivo* video clips, accumulating to 180.5k frames. From these, 7,117 frames were manually annotated at 1 fps, yielding 1,862 distinct point trajectories of the query points and a total of 17,171 visible tissue points across all the frames of this subset. For tissue tracking, VL-SurgPT encompasses five challenging *in vivo* surgical scenarios: *Tissue Deformation*, *Instrument Occlusion*, *Camera Jitter*, *Surface Reflection*, and *Cauterization Smoke*. Each scenario is well-represented, typically with over 120 video clips, more than 1,200 annotated frames, and between 300 to 450 tracked point trajectories. This balanced distribution across varied intraoperative conditions is designed to facilitate robust evaluation of point tracking models.

Instrument Tracking Subset: This subset includes 154 *in vivo* video clips, totaling 26,490 frames. Manual annotations were performed on 1,108 of these frames. The dataset features 7 distinct types of surgical instruments, with up to 7 predefined keypoints annotated per visible instrument instance in a labeled frame. For instrument tracking, seven commonly used surgical instruments are annotated: Harmonic Ace Curved Shears, Cadiere Forceps, Fenestrated Bipolar Forceps, Clip Applier, Clip, Tip-Up Fenestrated Grasper, and Needle Driver. Instruments like the Harmonic Ace Curved Shears and Fenestrated Bipolar Forceps have extensive coverage (hundreds of labeled frames and thousands of keypoints each), while less frequently appearing but procedurally critical tools such as the Clip Applier and Needle Driver are also included to ensure comprehensive representation of tool diversity.

Context-aware Textual Point Label Unlike traditional datasets that focus solely on geometric coordinates, VL-SurgPT associates each tracked point with descriptive textual labels, providing crucial semantic context, as demonstrated in Fig. 1. Specifically, each annotated point includes:

- **2D Coordinates**: Precise pixel location (x, y) in the frame or "null" in case the point is invisible.
- **Point Type**: Categorical label ("Tissue" or "Instrument").
- **Point Status**: A textual description from a predefined vocabulary indicating the point's current visual condition ("Clear View", "Pulled", "Reflection", "Smoke Obscuration", "Instrument Obscuration", "Tissue Obscuration", "Out of View", "External Occlusion", "Self-occlusion").
- **Instrument-Specific Labels**: **Instrument Type** (e.g., "Clip Applier", "Cadiere Forceps", etc.) and **Instance ID** used to distinguish instruments within the same frame.

The distribution of semantic labels across different visual conditions reveals important patterns in the challenges present during surgical procedures. For tissue points, while "Clear View" and "Pulled" statuses predominate, a substantial portion experiences obscuration from instruments and surrounding tissue. Specular reflections from moist tissue surfaces also appear frequently, representing a significant tracking challenge unique to surgical environments. For instrument points, the majority appear in clear view—a natural consequence of surgical protocols that prioritize instrument visibility for safe operation. Nevertheless, approximately one-third of annotated instrument points exhibit challenging visual conditions, including external occlusion by tissue or smoke, self-occlusion between instrument parts, and instances where instruments move partially or completely out of the field of view. This significant proportion of visually compromised points underscores the complexity of tool point tracking in dynamic *in vivo* procedures.

Multimodal Tracking Baseline: TG-SurgPT

To address the lack of text-guided reasoning and underutilization of sparse semantic labels in existing point tracking models, we propose Text-Guided Surgical Point Tracking (TG-SurgPT). Our approach builds upon the Track-On (Aydemir et al. 2025) framework, chosen for its balance of efficiency and performance, and extends it to enable text-guided tracking of arbitrary points.

As illustrated in Fig. 3, TG-SurgPT systematically integrates visual and textual modalities through a dual-branch architecture with cross-modal attention mechanisms. **Textual Branch:** For each tracked point, we encode two types of semantic attributes: **Point Type** (tissue/instrument) and **Point Status** (e.g., "Clear View", "Smoke Obscuration",

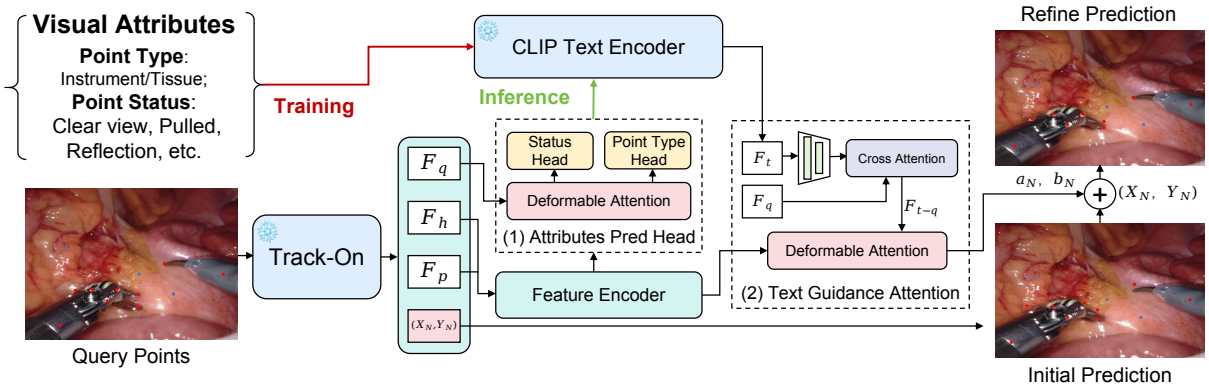

Figure 3: Overview of our Text-guided Surgical Point Tracking (TG-SurgPT). The method builds upon Track-On (Aydemir et al. 2025) by integrating visual features with semantic text descriptions through cross-modal attention.

"Pulled"). These descriptions are processed through a frozen CLIP Text Encoder (Radford et al. 2021) to produce textual feature embeddings $F_t \in \mathbb{R}^{2\times512}$. **Visual Branch:** The query points and video frames are processed by the frozen Track-On model, yielding: (1) feature representations of the query points (F_q), (2) dense visual features of the current frame (F_h), (3) coarse patch-level matching positions for each query point (F_p), and (4) initial predicted coordinates (X_N, Y_N), where N is the number of query points.

Attributes Pred Head To enable autonomous semantic understanding during inference, we introduce an Attributes Prediction Head that predicts point status from visual features alone. As illustrated in Fig. 3 (1), we first fuse F_p and F_h using a multi-scale correlation module. The resulting fused representation $F_{p\text{-}h}$ serves as both Key and Value in a Multi-Scale Deformable Attention module, with query features F_q as the Query. This attention mechanism allows each query to attend to a small set of spatially offset sampling points across multiple feature levels. We further introduce two parallel classification heads: a *Point Type Head*, which predicts a $2 \times N$ vector indicating the spatial location of each point; and a *Point Status Head*, which outputs a $7 \times N$ or $4 \times N$ matrix depending on the point type (tissue or instrument) for point-wise status classification.

Text-guided Attention We design a guided attention module for text-vision feature fusion, as illustrated in Fig. 3 (2). We first project the textual features F_t to match the dimensionality of visual query features F_q using learned linear transformations. We then perform standard cross-attention with F_q as the Query and F_t as the Key and Value, producing a fused representation $F_{t\text{-}q}$. This text-enhanced features $F_{t\text{-}q}$ are combined with the visual-spatial features $F_{p\text{-}h}$ through another deformable attention layer to produce refinement offsets (a_N, b_N). The final predicted coordinates are computed as:

$$(\hat{X}_N, \hat{Y}_N) = (X_N + a_N, Y_N + b_N). \quad (1)$$

It should be noted that during training, we use ground-truth textual descriptions of the dataset to guide visual understanding. At inference, these descriptions are replaced by the model's own predicted status labels, enabling text-guided reasoning without relying on manually provided annotations. This design allows the model to improve tracking accuracy in a fully automatic manner during deployment.

Loss Function Transformer-based models such as Track-On (Aydemir et al. 2025) normally require training on densely annotated synthetic datasets like MOVi-F (Greff et al. 2022), which uses a large number of query points (e.g., 2048 points sampled over 25-frame sequences). Our method is capable of training with sparsely labeled real-world surgical data without relying on dense point annotations, benefiting from textual annotations. During training, our loss function $\mathcal{L}$ consists of three key components: the point distance loss $\mathcal{L}_p$, the trajectory smoothness loss $\mathcal{L}_s$, and the textual classification loss $\mathcal{L}_{\text{text}}$. The total training loss $\mathcal{L}$ is accumulated across all frames with GT as:

$$\mathcal{L} = \sum_{t\in\mathcal{T}} \Big(\underbrace{\mathcal{H}_\delta(\hat{\mathbf{p}}_t - \mathbf{p}_t)}_{\mathcal{L}_p} + \underbrace{\|\Delta^2 \hat{\mathbf{p}}_t\|_1}_{\mathcal{L}_s} + \underbrace{\mathcal{L}_{\text{CE}}(\hat{\mathbf{s}}_t, \mathbf{s}_t)}_{\mathcal{L}_{\text{text}}} \Big), \quad (2)$$

where $\mathcal{T}$ denotes the set of selected frames for supervision, $\hat{\mathbf{p}}_t$ and $\mathbf{p}_t$ are the predicted and ground-truth point coordinates, and $\hat{\mathbf{s}}_t$, $\mathbf{s}_t$ denote the predicted and target status labels. Here, $\mathcal{L}_p$ measures the point-wise distance using the robust Huber loss (Meyer 2021) $\mathcal{H}_\delta$. $\mathcal{L}_s$ promotes temporal smoothness by minimizing the second-order difference $\Delta^2 \hat{\mathbf{p}}_t$ of predicted trajectories. Finally, $\mathcal{L}_{\text{text}}$ applies cross-entropy loss $\mathcal{L}_{\text{CE}}$ between $\hat{\mathbf{s}}_t$ and $\mathbf{s}_t$ to supervise semantic status classification, such as "Occluded" or "Off-camera".

Experiments

Experimental Setup

We conduct comprehensive experiments to establish baseline performance on VL-SurgPT and demonstrate the effectiveness of text-guided tracking. Our evaluation encompasses eight state-of-the-art vision-based tracking methods (RAFT (Teed and Deng 2020), SEARAFT (Wang, Lipson, and Deng 2024), MFT (Neoral, Šerỳch, and Matas 2024), MFTIQ (Serych, Neoral, and Matas 2025), TAPIR (Doersch et al. 2023), BootsTAP (Doersch et al. 2024), CoTrackerV3 (Karaev et al. 2024a), and Track-On (Aydemir

Method	Tissue Subset				Instrument Subset				Mean fps
	AJ ↑	$< \delta^x_{avg}$ ↑	OA ↑	EPE ↓	AJ ↑	$< \delta^x_{avg}$ ↑	OA ↑	EPE ↓	
RAFT (Teed and Deng 2020)	27.81	30.37	85.81	99.73	23.67	33.20	76.49	138.37	12.47
SEA-RAFT (Wang, Lipson, and Deng 2024)	20.49	24.08	83.14	89.52	18.11	27.07	78.02	103.18	13.25
TAPIR (Doersch et al. 2023)	40.01	46.63	70.63	56.99	41.81	49.49	76.15	75.78	13.41
BootsTAP (Doersch et al. 2024)	56.93	62.77	87.87	23.52	46.26	53.71	84.76	44.20	**14.25**
CotrackerV3 (Karaev et al. 2024a)	43.27	38.11	76.52	56.13	42.03	38.31	79.23	52.12	9.87
MFT (Neoral, Šerỳch, and Matas 2024)	57.61	67.12	90.46	18.07	45.63	55.18	86.13	40.97	1.42
MFTIQ (Serych, Neoral, and Matas 2025)	61.52	63.44	87.80	19.81	46.56	56.47	74.21	39.42	0.71
Track-On (Aydemir et al. 2025)	58.55	66.27	88.81	13.79	46.97	59.18	85.07	41.67	10.85
TG-SurgPT (Ours)	**62.88**	**67.77**	**91.04**	**11.02**	**49.52**	**62.94**	**89.79**	**39.14**	9.72

Table 1: Benchmark of tracking performance on tissue and instrument points. Our multimodal (Vision-Language) TG-SurgPT shows superiority over unimodal (vision-only) baselines, with a decent inference speed.

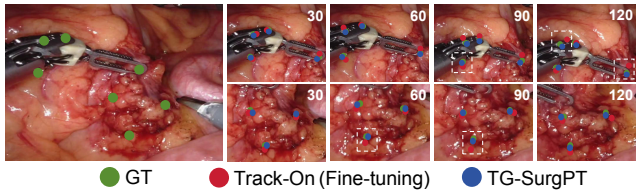

Figure 4: Qualitative comparison of tracking performance of the finetuned Track-On (Aydemir et al. 2025) and our TG-SurgPT across sequential frames (30, 60, 90, 120) for both instrument (top row) and tissue (bottom row) point tracking.

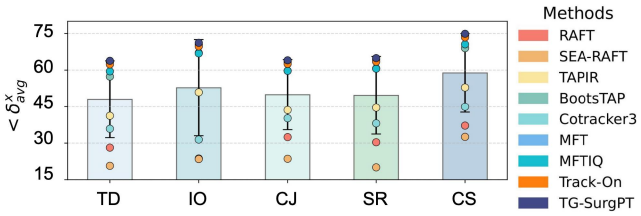

Figure 5: Scenario-specific comparison ($< \delta^x_{avg}$) across Tissue Deformation (TD), Instrument Occlusion (IO), Camera Jitter (CJ), Surface Reflection (SR), and Cauterization Smoke (CS). Box height indicates the mean performance.

et al. 2025)) and our proposed **TG-SurgPT**, a novel text-guided approach that leverages semantic descriptions. We follow standard metrics from point tracking literature (TAP-Vid (Doersch et al. 2022) and STIR (Schmidt et al. 2024b)) for reliable evaluation, including *Average Position Accuracy* ($< \delta^x_{\text{avg}}$), *Average Jaccard* (AJ), *Occlusion Accuracy*, *End Point Error* (EPE), *Inference Speed* (fps).

The dataset is split 4:1 for training and testing, stratified by surgical scenarios (tissue subset) and instrument types (instrument subset) to promote balanced distribution. All input frames are resized to 720×480 resolution for consistent evaluation. We use default parameters recommended by the original authors for all baseline methods to ensure fair comparison, with consistent input and evaluation protocols.

Results and Analysis

Benchmark Comparison Table 1 presents a detailed performance comparison across eight state-of-the-art tracking methods and our proposed TG-SurgPT approach. The results reveal several important patterns in surgical point tracking performance. Traditional flow-based methods (RAFT, SEA-RAFT) demonstrate substantial limitations in surgical environments, with Average Jaccard scores consistently below 30%, highlighting their inability to handle complex tissue deformations and visual artifacts inherent to surgical procedures. In contrast, transformer-based approaches show markedly superior performance, underscoring the effectiveness of attention mechanisms in capturing the complex spatial-temporal relationships in surgical scenes. For tissue tracking, MFTIQ achieves the highest baseline AJ (61.52%), while MFT excels in position accuracy (67.12%) and occlusion detection (90.46%). Track-On demonstrates optimal trajectory completion with the lowest EPE (13.79 px). Similar patterns appear in instrument tracking, where Track-On leads in spatial accuracy metrics while MFT maintains superior occlusion handling capabilities. However, a critical efficiency-accuracy trade-off emerges: while MFT and MFTIQ achieve strong accuracy, their impractical inference speeds (1.42-0.71 fps) preclude real-time clinical deployment. Track-On provides the optimal balance at 10.85 fps, making it suitable for surgical integration.

Our multimodal TG-SurgPT consistently outperforms all baselines across every metric while maintaining practical processing speed (9.72 fps). The improvements are particularly pronounced for instrument tracking (5.4% AJ gain, 6.4% position accuracy improvement), demonstrating that semantic text guidance provides contextual information unavailable to purely visual methods. These results validate our core hypothesis that incorporating point status descriptions significantly enhances tracking robustness under challenging surgical conditions where conventional vision-only approaches fail. Figure 4 compares TG-SurgPT with finetuned Track-On across sequential frames, demonstrating TG-SurgPT's enhanced trajectory coherence and accuracy in tracking instruments and tissue points.

Scenario-specific Results Analysis Figure 5 reveals the relative difficulty hierarchy across five challenging surgi-

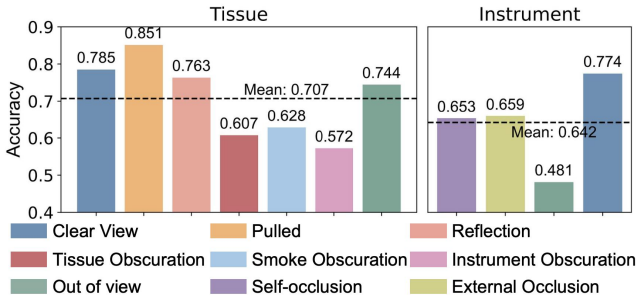

Figure 6: Classification accuracy of different visual status categories for both tissue and instrument subsets.

cal conditions. Tissue Deformation emerges as the most difficult scenario, where complex non-rigid transformations fundamentally challenge current visual tracking paradigms. Surprisingly, Cauterization Smoke (CS) yields the highest performance (58.62% mean) despite intuitive expectations of visual impairment, likely because smoke creates distinctive temporal patterns and edge features that attention mechanisms can effectively leverage. The consistent performance gaps between methods across scenarios highlight fundamental algorithmic differences. Flow-based methods show severe degradation in all conditions, while transformer-based approaches exhibit scenario-dependent variations. TG-SurgPT demonstrates consistent improvements across all scenarios, particularly in dynamic conditions like IO and CJ. This suggests that semantic descriptions provide crucial disambiguation when visual features become unreliable. The results strongly support our hypothesis that text guidance is most beneficial in situations where visual-only methods face the greatest challenges.

Visual Status Prediction Analysis Figure 6 reports the prediction accuracy for each visual status category across the tissue and instrument subsets with our TG-SurgPT. For tissue points (left), distinctive states like *Pulled* (85.1%) and *Clear View* (78.5%) achieve high accuracy, while ambiguous obscuration states show lower performance, suggesting difficulty in disambiguating between different types of visual obstruction when tissue points become partially or completely hidden. For instrument points (right), both *Self-occlusion* (65.3%) and *External Occlusion* (65.9%) achieve decent performance, showing the model's ability to distinguish between different types of occlusion scenarios. However, *Out of View* exhibits the poorest accuracy (48.1%), highlighting the challenge of correctly identifying when instrument points move outside the field of view, particularly during rapid instrument movements or camera transitions. Overall, while our TG-SurgPT model demonstrates promising semantic understanding capabilities across most visual status categories, the results reveal distinct challenges for each point type: tissue points struggle most with occlusion disambiguation, while instrument points face difficulties in out-of-view detection. These findings indicate specific areas for improvement in the classification of visual status.

Finetuning	Clip Size	Text	AJ ↑	$< \delta^x_{avg}$ ↑	OA ↑	EPE ↓
Tissue Subset						
×	×	×	58.55	66.27	88.81	13.79
✓	short	×	61.09	66.39	89.70	12.82
✓	long	×	59.91	66.73	89.41	12.94
✓	short	✓	**62.88**	**67.77**	**91.04**	**11.02**
✓	long	✓	62.09	67.46	90.88	11.77
Instrument Subset						
×	×	×	46.97	59.18	85.07	41.67
✓	short	×	47.48	60.75	86.43	40.41
✓	long	×	47.22	59.89	87.01	40.82
✓	short	✓	**49.52**	62.94	**89.79**	**39.14**
✓	long	✓	48.77	**63.08**	88.75	39.69

Table 2: Ablation study on fine-tuning, clip length, and text guidance for tissue and instrument subsets.

Ablation Study Table 2 summarizes ablation studies on the contributions of fine-tuning, training clip length, and text guidance in TG-SurgPT for tissue and instrument tracking. **Fine-tuning Impact:** Fine-tuning consistently improves performance across all metrics. For tissue tracking, short-clip fine-tuning increases AJ from 58.55 to 61.09 (+4.3%) and reduces EPE from 13.79 to 12.82 px. Similar gains are observed for instrument tracking (AJ: 46.97→47.48, +1.1%). **Clip Length Analysis:** Short clips (31 frames per clip) generally outperform long clips (181 frames per clip) in most metrics, particularly for AJ and EPE. This suggests that shorter temporal windows reduce noise and improve point correspondence stability, making them more suitable for sparse surgical annotations. **Text Guidance Effectiveness:** Text guidance provides substantial improvements across both subsets. For tissue tracking, adding text guidance to short-clip fine-tuning yields the best performance: AJ increases to 62.88 (+2.9% over visual-only), OA reaches 91.04 (+1.3%), and EPE drops to 11.02 px (-14.0%). For instrument tracking, text guidance achieves AJ of 49.52 (+4.3%) and OA of 89.79 (+3.9%), demonstrating that semantic context significantly enhances tracking robustness.

Discussion and Conclusion

We present **VL-SurgPT**, the first large-scale multimodal dataset for surgical point tracking that systematically integrates visual trajectories with semantic point status descriptions. Through comprehensive benchmarking and the proposed text-guided tracking method **TG-SurgPT**, we demonstrate the significant potential of incorporating semantic understanding into surgical scene analysis, paving the way for more robust and context-aware point tracking in computer-assisted surgery. Current limitations include the dataset's limited scope (primarily gastrointestinal procedures using the da Vinci system) and TG-SurgPT's inference speed of 9.72 fps, which approaches but does not meet real-time clinical requirements. Future work will expand the dataset to additional surgical domains, optimize inference speed for real-time clinical use, and develop more advanced cross-modal learning strategies to better leverage limited labeled data.

Acknowledgments

This work was supported in part by Hong Kong Research Grants Council (RGC) Collaborative Research Fund (CRF C4026-21GF), Research Impact Fund RIF R4020-22 and General Research Fund (GRF 14216022, 14204524, 14203323, 14206125), NSFC Young Scientists Fund - Category A T252500134, NSFC/RGC Joint Research Scheme N_CUHK420/22; Guangdong Basic and Applied Basic Research Foundation (GBABF) #2021B1515120035.

References

Aydemir, G.; Cai, X.; Xie, W.; and Güney, F. 2025. TrackOn: Transformer-based Online Point Tracking with Memory. In *The Thirteenth International Conference on Learning Representations*.

Bai, L.; Wang, G.; Islam, M.; Seenivasan, L.; Wang, A.; and Ren, H. 2025. Surgical-VQLA++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. *Information Fusion*, 113: 102602.

Bodenstedt, S.; Allan, M.; Agustinos, A.; Du, X.; Garcia-Peraza-Herrera, L.; Kenngott, H.; Kurmann, T.; Müller-Stich, B.; Ourselin, S.; Pakhomov, D.; et al. 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475*.

Cartucho, J.; Weld, A.; Tukra, S.; Xu, H.; Matsuzaki, H.; Ishikawa, T.; Kwon, M.; Jang, Y. E.; Kim, K.-J.; Lee, G.; et al. 2024. Surgt challenge: Benchmark of soft-tissue trackers for robotic surgery. *Medical image analysis*, 91: 102985.

Doersch, C.; Gupta, A.; Markeeva, L.; Recasens, A.; Smaira, L.; Aytar, Y.; Carreira, J.; Zisserman, A.; and Yang, Y. 2022. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35: 13610–13626.

Doersch, C.; Luc, P.; Yang, Y.; Gokay, D.; Koppula, S.; Gupta, A.; Heyward, J.; Rocco, I.; Goroshin, R.; Carreira, J.; et al. 2024. Bootstap: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision*, 3257–3274.

Doersch, C.; Yang, Y.; Vecerik, M.; Gokay, D.; Gupta, A.; Aytar, Y.; Carreira, J.; and Zisserman, A. 2023. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10061–10072.

Greff, K.; Belletti, F.; Beyer, L.; Doersch, C.; Du, Y.; Duckworth, D.; Fleet, D. J.; Gnanapragasam, D.; Golemo, F.; Herrmann, C.; et al. 2022. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3749–3761.

Guo, J.; Wang, J.; Li, Z.; Jia, T.; Dou, Q.; and Liu, Y.-H. 2024. Ada-Tracker: Soft Tissue Tracking via Inter-Frame and Adaptive-template Matching. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 15463–15470. IEEE.

Karaev, N.; Makarov, I.; Wang, J.; Neverova, N.; Vedaldi, A.; and Rupprecht, C. 2024a. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*.

Karaev, N.; Rocco, I.; Graham, B.; Neverova, N.; Vedaldi, A.; and Rupprecht, C. 2024b. Cotracker: It is better to track together. In *European Conference on Computer Vision*, 18–35. Springer.

Li, Y.; Richter, F.; Lu, J.; Funk, E. K.; Orosco, R. K.; Zhu, J.; and Yip, M. C. 2020. Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics. *IEEE Robotics and Automation Letters*, 5(2): 2294–2301.

Lin, S.; Miao, A. J.; Lu, J.; Yu, S.; Chiu, Z.-Y.; Richter, F.; and Yip, M. C. 2023. Semantic-super: a semantic-aware surgical perception framework for endoscopic tissue identification, reconstruction, and tracking. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 4739–4746. IEEE.

Meyer, G. P. 2021. An alternative probabilistic interpretation of the huber loss. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 5261–5269.

Min, Z.; Lai, J.; and Ren, H. 2025. Innovating robot-assisted surgery through large vision models. *Nature Reviews Electrical Engineering*, 1–14.

Neoral, M.; Šerỳch, J.; and Matas, J. 2024. Mft: Long-term tracking of every pixel. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6837–6847.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Schmidt, A.; Mohareri, O.; DiMaio, S.; and Salcudean, S. E. 2023. Sendd: Sparse efficient neural depth and deformation for tissue tracking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 238–248. Springer.

Schmidt, A.; Mohareri, O.; DiMaio, S.; Yip, M. C.; and Salcudean, S. E. 2024a. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 103131.

Schmidt, A.; Mohareri, O.; DiMaio, S. P.; and Salcudean, S. E. 2024b. Surgical tattoos in infrared: A dataset for quantifying tissue tracking and mapping. *IEEE Transactions on Medical Imaging*, 43(7): 2634–2645.

Seenivasan, L.; Islam, M.; Kannan, G.; and Ren, H. 2023. SurgicalGPT: end-to-end language-vision GPT for visual question answering in surgery. In *International conference on medical image computing and computer-assisted intervention*, 281–290. Springer.

Seenivasan, L.; Islam, M.; Krishna, A. K.; and Ren, H. 2022. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 33–43. Springer.

Serych, J.; Neoral, M.; and Matas, J. 2025. Mftiq: Multi-flow tracker with independent matching quality estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8079–8089. IEEE.

Sznitman, R.; Ali, K.; Richa, R.; Taylor, R. H.; Hager, G. D.; and Fua, P. 2012. Data-driven visual tracking in retinal microsurgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 568–575. Springer.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Wang, Y.; Lipson, L.; and Deng, J. 2024. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, 36–54. Springer.

Wu, Z.; Schmidt, A.; Moore, R.; Zhou, H.; Banks, A.; Kazanzides, P.; and Salcudean, S. E. 2025. Surgpose: a dataset for articulated robotic surgical tool pose estimation and tracking. *arXiv preprint arXiv:2502.11534*.

Xu, M.; Islam, M.; Lim, C. M.; and Ren, H. 2021. Learning domain adaptation with model calibration for surgical report generation in robotic surgery. In *2021 IEEE international conference on robotics and automation (ICRA)*, 12350–12356. IEEE.

Xu, M.; Islam, M.; and Ren, H. 2022. Rethinking surgical captioning: End-to-end window-based mlp transformer using patches. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 376–386. Springer.

Yuan, K.; Navab, N.; Padoy, N.; et al. 2024a. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37: 122952–122983.

Yuan, K.; Srivastav, V.; Navab, N.; and Padoy, N. 2024b. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 306–316. Springer.

Zeng, Z.; Zhuo, Z.; Jia, X.; Zhang, E.; Wu, J.; Zhang, J.; Wang, Y.; Low, C. H.; Jiang, J.; Zheng, Z.; et al. 2025. SurgVLM: A Large Vision-Language Model and Systematic Evaluation Benchmark for Surgical Intelligence. *arXiv preprint arXiv:2506.02555*.

Zhan, B.; Zhao, W.; Fang, Y.; Du, B.; Vasconcelos, F.; Stoyanov, D.; Elson, D. S.; and Huang, B. 2024. Tracking Everything in Robotic-Assisted Surgery. *arXiv preprint arXiv:2409.19821*.

Zheng, Y.; Harley, A. W.; Shen, B.; Wetzstein, G.; and Guibas, L. J. 2023. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19855–19865.

Zhou, R.; He, W.; Wang, A.; Yao, Q.; Hu, H.; Wang, J.; et al. 2025. Endo-TTAP: Robust Endoscopic Tissue Tracking via Multi-Facet Guided Attention and Hybrid Flow-point Supervision. *arXiv preprint arXiv:2503.22394*.