PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

A Survey of Artificial Intelligence in Endoscopic Surgery Workflow: From Perception to Surgical Support

Juyan Ba[1], Hao Chen[2], Xiaohan Xing[3*], Yi Wang[1*]
[1]Shenzhen University
[2]Cambridge University
[3]Stanford University
2410245002@mails.szu.edu.cn, hc666@cam.ac.uk, xhxing@stanford.edu, onewang@szu.edu.cn

Abstract

Endoscopic surgery demands continuous real-time visual decision-making under severe constraints, including a limited field of view, motion blur, and dynamically deforming anatomy. These factors impose substantial cognitive load on surgeons and motivate the integration of artificial intelligence (AI) throughout the endoscopic surgical workflow. This survey reviews recent progress in AI for endoscopic surgery and organizes the literature into four stages that span perception to action: (1) image enhancement and analysis methods that improve visual perception; (2) multimodal video understanding approaches that model and reason about surgical instruments and anatomical structures over space and time; (3) 3D reconstruction techniques that enable robust tracking and interpretation of deformable anatomy; and (4) emerging paradigms of embodied surgical intelligence, where action-conditioned world models link perception to intraoperative assistance. Across these stages, we summarize current capabilities and limitations and identify key open challenges for clinical deployment. In addition, we provide an overview of 18 publicly available datasets, highlighting their scope and annotations. We hope this survey will stimulate further research toward reliable and clinically deployable AI systems for endoscopic surgery.

1 Introduction

Endoscopic surgery reduces patient trauma and accelerates recovery compared to open procedures, but this benefit comes at a perceptual cost. Unlike open surgery, where clinicians have a global anatomical view and direct tactile feedback, endoscopic procedures require surgeons to navigate complex anatomy through a narrow monocular (or near-monocular) video stream that is frequently degraded by motion blur, tissue occlusion, and organ deformation. The requirement to understand a dynamic surgical environment using this 2D video on screen imposes cognitive load, requiring surgeons to mentally reconstruct spatial relationships and infer tissue-instrument interactions from fragmented visual cues. These challenges motivate the development of AI-based tools capable of improving visual quality and extracting geometric information from endoscopic video to assist in instrument localization and manipulation.

Recent advances in generative models, multimodal large language models (MLLMs), and embodied intelligence have inspired broader research interest in supporting endoscopic surgery. Moving beyond treating surgical videos as passive inputs for offline analysis, these advances point toward a new paradigm of active, spatially grounded surgical assistance.

In this paradigm, learning systems are expected to reconstruct 3D anatomy, localize surgical instruments and lesions in the physical workspace, and ultimately support or automate robotic actions. These capabilities align more closely with actual clinical workflows, offering direct support that is genuinely in demand, especially given concerns about patient safety and procedural variability.

The path toward realizing this vision, however, remains unclear, despite prominent progress across different areas. Advances have been made in perception, understanding, and robotic control, but these developments are often pursued in isolation. The resulting fragmentation obscures key system-level bottlenecks, the dependencies between components, and the research directions most likely to yield a clinically meaningful impact. To address this gap, we present a systematic review that synthesizes progress across these interconnected areas and frames them within a unified embodied intelligence perspective. Accordingly, we ask: *what are the key steps toward embodied intelligence in endoscopic surgery?*

We organize this pathway into four progressively higher-level stages as shown in Figure 1. (1) Visual Perception and Processing focuses on learning robust visual representations through image quality enhancement, domain and style normalization, depth estimation, and segmentation. (2) Surgical Scene Understanding builds upon these representations to reason about surgical entities and context, including tissues, instruments, and procedural stages. (3) 3D Scene Reconstruction lifts perception into a spatially grounded representation by recovering 3D structure and aligning reconstructed models with the physical surgical workspace. (4) Embodied Surgical Intelligence closes the perception–action loop by integrating visually-conditioned control, decision-making, and

*Corresponding author.

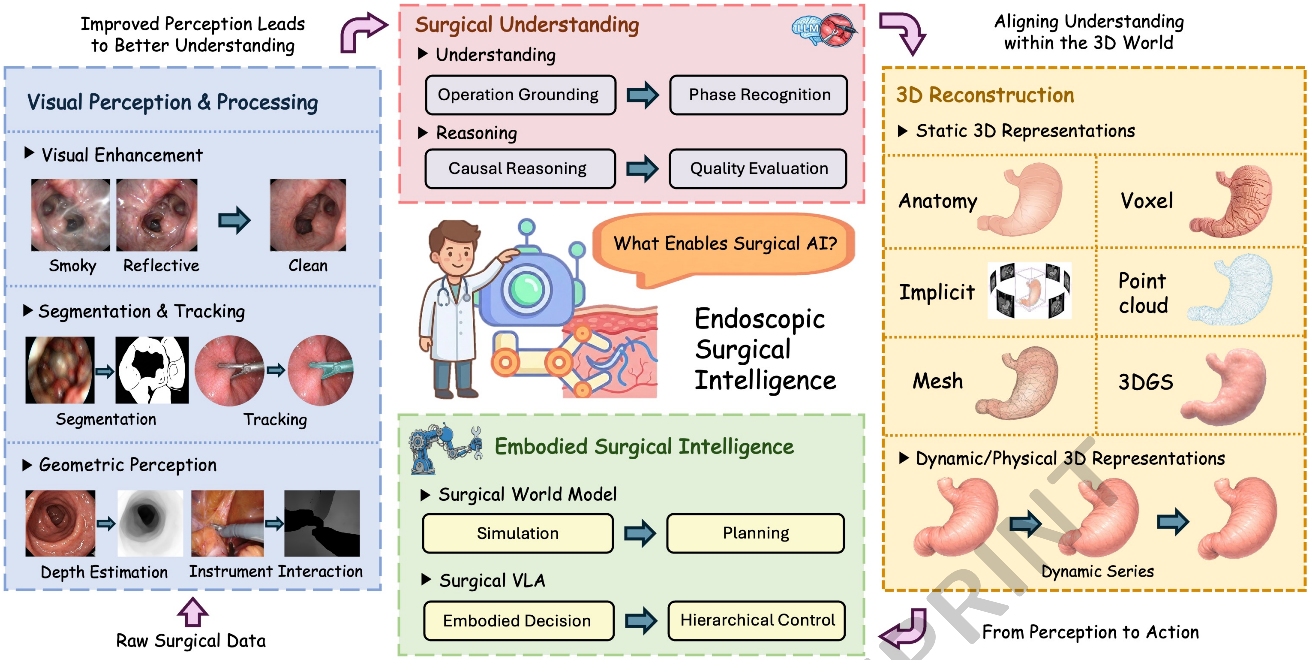

Figure 1: **Toward endoscopic surgical intelligence.** Starting from raw surgical data, endoscopic AI systems progressively transform perception into understanding, align high-level reasoning with the 3D surgical world, and ultimately enable embodied intelligence that links perception to planning and action in endoscopic surgery.

action-centric world models. For each stage, we analyze the current methods, identify fundamental bottlenecks and open challenges, and discuss how progress across stages can be jointly leveraged rather than pursued in isolation.

Finally, to advance embodied intelligence research, we first review existing datasets to highlight limitations in current data regimes (e.g., scale, diversity, and action-level supervision) that hinder progress in embodied intelligence. We then outline open research directions and articulate a long-term vision for embodied intelligence in endoscopic surgery and its integration within the operating room.

2 Visual Perception and Processing

Visual perception forms the foundation of surgical intelligence. However, raw surgical footage is often degraded by blood, motion blur, and poor lighting, which can severely compromise downstream analysis. Effective image processing is therefore required to reveal informative visual cues and stabilize subsequent interpretation. Addressing these challenges is essential for enabling higher-level understanding and decision-making. Accordingly, this section focuses on three core problems in visual perception: visual enhancement, segmentation, and geometric perception.

2.1 Visual Enhancement

Visual enhancement transforms degraded endoscopic footage into high-fidelity video suitable for downstream analysis.

Frame Processing. Early methods operate at the frame level. LighTDiff [Chen *et al.*, 2024] employs a diffusion architecture for low-light enhancement, processing at a lower resolution before recovering fine details. LVQIS [Zheng *et al.*, 2023] uses MPRNet to mitigate motion blur and restore sharpness in laparoscopic videos. EndoL2H [Almalioglu *et al.*, 2020] addresses resolution limitations in capsule endoscopy through conditional GANs with spatial attention, achieving up to 12× upscaling.

Video Processing. While efficient, frame-based methods cannot enforce temporal consistency, resulting in visual jitter and flickering across frames. This limitation motivates recent work that shifts toward video-based modeling. DGGAN [Xu *et al.*, 2025a] propagates degradation representations across neighboring frames to address multiple degradations including uneven illumination and motion blur. RVDNet [Zhu *et al.*, 2025] utilizes smoke-free reference frames to guide video desmoking in surgery, maintaining temporal coherence.

Discussion. Despite these advances, several key challenges remain. First, unified models for compound degradations (e.g., simultaneous smoke, blur, and low light) common in surgical settings are still lacking. Second, the unresolved cost-consistency trade-off limits real-time deployment. Third, current methods are largely confined to synthetic environments, and generalization to diverse real-world scenarios remains insufficiently explored.

2.2 Segmentation and Tracking

Surgical segmentation and tracking aim to parse anatomical structures and instruments from endoscopic videos, under challenges that differ fundamentally from natural scenes.

Anatomical Segmentation. Unlike natural images, surgical anatomy exhibits weak texture cues and ambiguous boundaries between adjacent tissues. Existing methods primarily

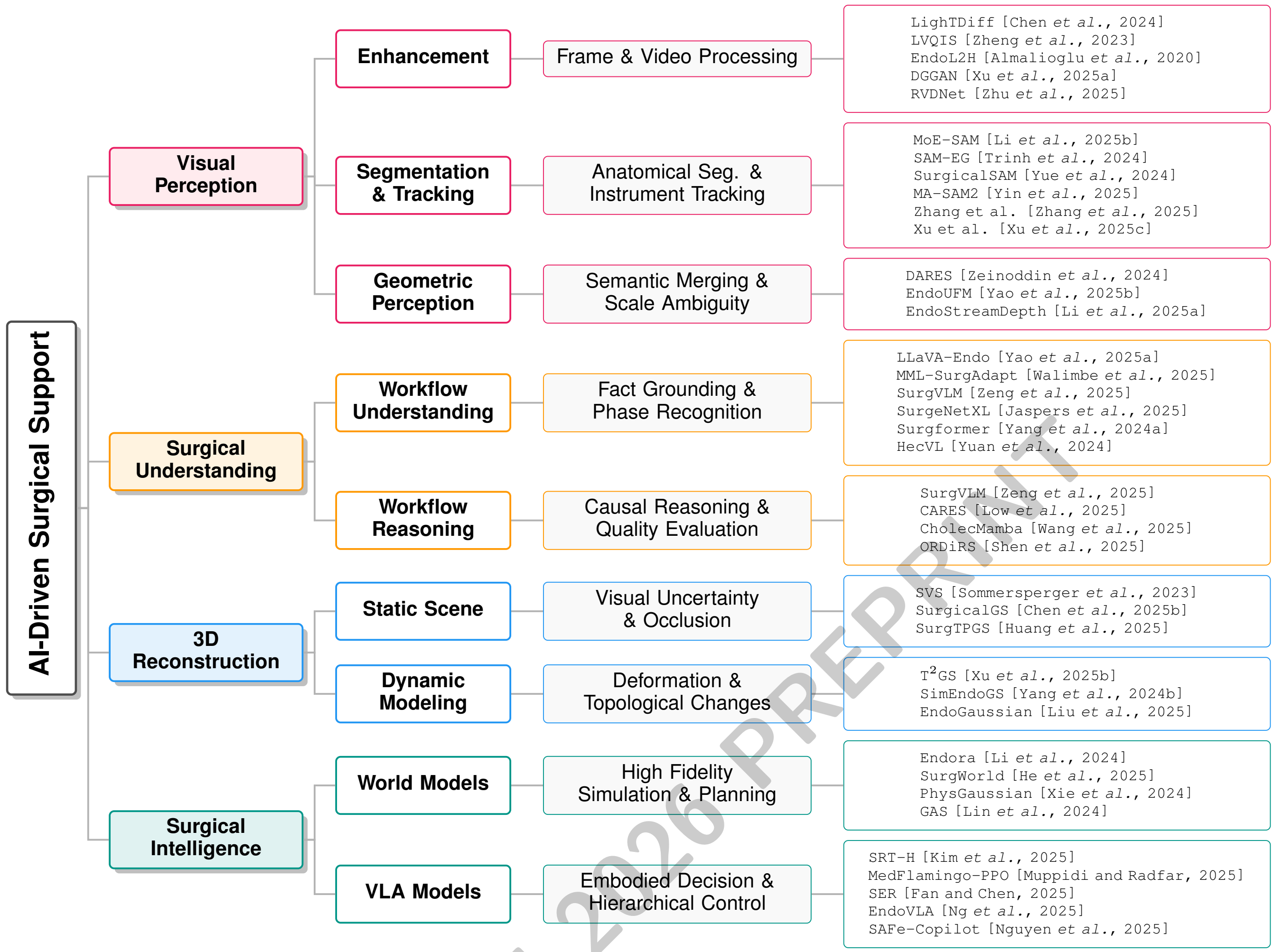

Figure 2: Taxonomy of AI-Driven Surgical Support: Perception, Understanding, Reconstruction, and Surgical Intelligence.

improve *what can be segmented* by injecting task-specific inductive biases. MoE-SAM [Li *et al.*, 2025b] routes features through multiple organ-specialized experts, enabling finer discrimination when texture statistics vary across organs. While it is effective for known anatomical categories, it assumes stable organ-specific appearance, limiting robustness to unseen anatomy. SAM-EG [Trinh *et al.*, 2024] instead focuses on *why boundaries fail* by adding explicit edge supervision, improving localization in low-contrast regions. While effective for boundary refinement, it cannot resolve semantic ambiguity when tissues share similar edge patterns, as it does not introduce higher-level anatomical reasoning.

Instrument Tracking. Instrument tracking faces a different set of constraints: identity preservation under occlusion and fast motion. Memory-based approaches such as MA-SAM2 [Yin *et al.*, 2025] explicitly address temporal discontinuity by re-identifying instruments after occlusion. However, such mechanisms rely on appearance consistency and degrade under severe motion blur. Zhang et al. [Zhang *et al.*, 2025] improve temporal stability by decoupling mask generation from category prediction, reducing jitter during rapid motion, but do not resolve inter-class ambiguity. SurgicalSAM [Yue *et al.*, 2024] tackles this by embedding category priors, enabling discrimination between visually similar tools. Still, these methods remain fundamentally 2D and thus fail to reason about physical interaction. Xu et al. [Xu *et al.*, 2025c] extend tracking to 3D, enabling estimation of instrument–tissue contact and reasoning beyond appearance; however, the approach requires predefined geometry, which restricts scalability and generalization across instruments.

Discussion. Existing methods address isolated failure modes, such as weak texture, boundary ambiguity, or occlusion, through task-specific priors, but often sacrifice generalizability. Moreover, heavy reliance on two-dimensional appearance constrains the ability to reason about interactions between organs and instruments. Overcoming these limitations requires more realistic modeling in real surgical settings.

2.3 Geometric Perception

Geometric perception in surgical video aims to recover spatial structure from monocular endoscopy. Deformable tissues, specular highlights, and the lack of fixed landmarks make depth inference ill-posed, forcing surgeons to rely on subjective estimation rather than reliable geometric cues.

Monocular Depth Estimation. Most approaches focus on improving *relative depth consistency* under adverse visual conditions. DARES [Zeinoddin *et al.*, 2024] enhances texture cues to sharpen depth discontinuities, improving local boundary perception under complex lighting. However, operating purely on appearance, it cannot distinguish instruments from tissue, leading to spatially plausible but semantically inconsistent depth predictions. EndoUFM [Yao *et al.*, 2025b] addresses this limitation by coupling depth with semantic labels, allowing geometry to be interpreted in a context-aware manner. While this improves spatial coherence within a frame, it does not address instability across time.

Temporal Stability. EndoStreamDepth [Li *et al.*, 2025a] explicitly targets instability in depth estimation by enforcing temporal smoothness, producing more stable depth streams. Nevertheless, like other monocular methods, it remains limited to relative depth and cannot recover metric scale.

Discussion. Current approaches largely treat depth as a visual reconstruction problem rather than a physical one. Relative depth may suffice for perception enhancement, but it fails to capture the true geometry required for safe interaction, such as estimating clearance, force, or tool reach. In surgical settings, where millimeter-level errors can be critical, the lack of metric depth undermines reliability and limits automation. This gap highlights the core geometric challenge: recovering physically grounded 3D structure under deformable, reflective, and dynamic conditions, which remains unresolved by monocular depth estimation alone.

3 Surgical Understanding

High-quality visual perception alone is insufficient for surgical intelligence; clinical decision-making requires structured semantic understanding and reasoning about procedural context. We distinguish surgical understanding into two complementary levels: *workflow understanding*, which captures objective surgical facts, and *workflow reasoning*, which evaluates procedural quality and safety.

3.1 Workflow Understanding

Workflow understanding aims to transform surgical video into objective procedural records and temporal logic, such as operative reports that accurately document what actions occurred, when they occurred, and which surgical phase is currently underway.

Fact Grounding. Recent work introduces domain-specific grounding to address this gap. LLaVA-Endo [Yao *et al.*, 2025a] uses professional terms to describe surgical steps. MML-SurgAdapt [Walimbe *et al.*, 2025] identifies operational intent by analyzing how instruments and tissues interact. SurgVLM [Zeng *et al.*, 2025] further demonstrates that a unified model can support multiple workflow tasks, improving record completeness. Nevertheless, these models remain sensitive to institutional variations in equipment and protocols, leading to unstable performance in unseen environments.

Phase Recognition. Phase recognition supports operating room scheduling and risk warnings by capturing the temporal logic and long-term dependencies of surgical workflows. SurgeNetXL [Jaspers *et al.*, 2025] and Surgformer [Yang *et al.*, 2024a] improve generalization via pretraining and modeling efficiency through hierarchical designs, respectively. Furthermore, by aligning visual observations with professional terminology (e.g., HecVL [Yuan *et al.*, 2024]) and incorporating instrument intent recognition, these methods convert complex visual evolution into structured clinical facts, enabling precise partitioning of complex surgical phases.

Discussion. Existing approaches effectively capture what happens during surgery but rely heavily on learned correlations tied to specific data distributions, making them sensitive to shifts in visual appearance or procedural conventions. Developing methods that are less dependent on institutional style and more deeply grounded in invariant surgical principles remains an important direction for future work.

3.2 Workflow Reasoning

Workflow reasoning shifts from factual description to evaluative judgment, addressing whether surgical actions are safe and clinically appropriate. Traditional approaches rely on low-level kinematic metrics, which capture motion patterns but often fail to reflect adherence to surgical standards.

Recent methods introduce explicit clinical reasoning. SurgVLM [Zeng *et al.*, 2025] and CARES [Low *et al.*, 2025] leverage language-based evaluation and causal modeling to assess procedural quality and identify hazards. Similarly, CholecMamba [Wang *et al.*, 2025] and ORDiRS [Shen *et al.*, 2025] ensure reasoning consistency and operational safety through temporal modeling and digital twin grounding, respectively.

Discussion. While language-based reasoning improves interpretability, current models struggle to maintain logical consistency over long procedures and to ground abstract judgments in actionable signals. The inability to translate reasoning outputs into executable safety constraints remains a key barrier, limiting the integration of workflow reasoning into autonomous surgical systems.

4 3D Reconstruction

Three-dimensional (3D) reconstruction provides spatial context that is absent in two-dimensional endoscopy, enabling intuitive perception of anatomical depth and structure. This geometric understanding serves as a foundation for robotic assistance and clinical decision-making. This section examines how different reconstruction paradigms balance fidelity, editability, and real-time performance in surgical settings.

4.1 3D Representations

The choice of 3D representation largely determines whether a reconstruction can support intraoperative interaction. Voxel grids explicitly encode volumetric information but scale poorly and incur high memory and computational costs, limiting real-time use. Mesh-based representations reduce storage by modeling surfaces and are well-suited for smooth anatomy; however, their fixed topology makes them fragile under large tissue deformations. Point clouds offer a lightweight alternative but are often too sparse to support precise boundary reasoning.

Implicit neural representations model (e.g., NeRF [Mildenhall *et al.*, 2020]) continuous geometry and appearance, achieving high visual fidelity and reduced artifacts. However, their reliance on iterative optimization and expensive rendering introduce latency that is incompatible with real-time surgical requirements. These limitations have motivated explicit alternatives such as 3D Gaussian Splatting [Kerbl *et al.*, 2023], which represents scenes using learnable Gaussian primitives.

Discussion. Selecting a 3D representation involves a fundamental trade-off between geometric expressiveness and computational efficiency. While continuous representations excel in visual quality, explicit geometric primitives are better aligned with real-time constraints. As a result, Gaussian-based representations have gained traction as a balanced solution that supports interactive performance while maintaining sufficient geometric realism for surgical applications.

4.2 Static Scene Reconstruction

Static reconstruction aims to restore accurate geometry to support spatial reasoning in visually ambiguous surgical scenes, such as texture-poor endoluminal environments. A key challenge is depth ambiguity caused by single-source illumination, which weakens natural shading cues. SVS [Sommersperger *et al.*, 2023] addresses this issue by introducing artificial shadows, enhancing depth perception without altering underlying geometry.

Another major challenge arises from frequent occlusions caused by surgical instruments. SurgicalGS [Chen *et al.*, 2025b] addresses this issue by inferring occluded anatomy to preserve spatial continuity during tool interaction. More recent work moves beyond purely static reconstruction toward interactive representations. For example, SurgTPGS [Huang *et al.*, 2025] integrates geometry with language, allowing surgeons to query spatial structures directly.

Discussion. Static reconstruction assumes an unchanging anatomy, enabling spatial understanding but limiting its applicability to visualization. As a result, such models are limited in supporting tasks that require reasoning about procedural changes or surgical interventions.

4.3 Dynamic and Generative Modeling

Dynamic reconstruction [Chen *et al.*, 2025a] handles evolving surgical scenes involving tissue deformation, instrument interaction, and structural changes from cutting. T^2GS [Xu *et al.*, 2025b] enables localized updates by decoupling instrument motion from tissue deformation to reduce drift. SimEndoGS [Yang *et al.*, 2024b] integrates physical constraints to simulate geometric changes after tissue removal and improves scene alignment.

Recent studies increasingly incorporate generative models. EndoGaussian [Liu *et al.*, 2025] uses deformation layers to preserve geometric consistency and reduce flicker during rapid motion. However, drastic topological changes like excision or fragmentation remain difficult to model.

Discussion. Dynamic models move reconstruction from passive display toward interactive simulation, but at the cost of stability and efficiency. While deformation can be tracked, topological changes and long-term consistency remain unresolved. Bridging physical realism with real-time responsiveness is essential for deploying dynamic reconstruction in autonomous or semi-autonomous surgical systems.

5 Embodied Surgical Intelligence

Embodied surgical intelligence extends beyond perception to enable actionable decision support, with the goal of improving safety and efficiency through autonomous or semi-autonomous systems. A central challenge is the scarcity of paired data that jointly captures surgical video, language, and robot-executable action representations. To address this limitation, surgical world models simulate procedural dynamics to generate scalable multimodal supervision, while vision–language–action (VLA) models leverage this structure to translate clinical intent into executable behavior.

5.1 Surgical World Models

Surgical world models aim to predict how the surgical environment evolves in response to actions, enabling systems to evaluate outcomes before execution. This predictive capability is essential for risk-aware planning, such as anticipating tissue deformation during cutting or visibility changes caused by bleeding. Existing approaches largely diverge into two complementary routes, differing in how explicitly they model visual and physical detail.

High-Fidelity Simulation Route. High-fidelity approaches prioritize pixel-level realism, treating visual accuracy as the foundation for reliable prediction. Endora [Li *et al.*, 2024] achieves realistic endoscopic rendering, and SurgWorld [He *et al.*, 2025] advances this by learning action-controllable dynamics from unlabeled videos, while PhysGaussian [Xie *et al.*, 2024] further constrains models with physical laws to ensure plausible tissue deformation.

Efficient Planning Route. In contrast, efficient planning approaches operate in compact latent spaces, assuming that high-resolution visual detail is largely redundant for control. GAS [Lin *et al.*, 2024] exemplifies this strategy by optimizing robotic actions through forward prediction without full image synthesis, enabling high-frequency control suitable for delicate manipulation.

Discussion. High fidelity simulations model complex phenomena but cause latency unsuitable for real time control. These methods suit offline evaluation rather than direct surgical decision making. Conversely, planning routes prioritize speed for interactive assistance. However, compressed representations may discard clinical details and raise safety concerns. Balancing model compactness with essential medical information remains a major challenge.

Surgical world models have transitioned from theory into essential foundations for modern robotics. The current trajectory involves merging efficient planning with realistic physics. This allows robots to run parallel simulations to anticipate complications and optimize precise tissue handling before execution.

Dataset	Type	Organs	Supported Tasks			Scale / Resolution
			Rec.	Loc.	Seg.	
Synthetic Datasets (Computer-Generated)						
Endo-SLAM [Ozyoruk *et al.*, 2021]	Synthetic	Stomach	●	●	●	35k frames (320 × 320)
UCL [Rau *et al.*, 2019]	Synthetic	Colon	●	○	○	16k frames (256 × 256)
EndoMapper [Azagra *et al.*, 2023]	Synthetic	Colon	●	●	○	6 seq. (Dense GT)
Simulation platform [Zhang *et al.*, 2020]	Synthetic	Colon	○	●	○	15 cases (640 × 480)
Phantom Datasets (Physical Models)						
SCARED [Allan *et al.*, 2021]	Phantom	Abdomen	●	●	○	23k frames (1280 × 1024)
Endo-SLAM [Ozyoruk *et al.*, 2021]	Phantom	GI Tract	●	●	○	42k frames (640 × 480)
EndoAbs [Penza *et al.*, 2018]	Phantom	Spleen	●	●	●	120 frames (640 × 480)
SERV-CT [Edwards *et al.*, 2022]	Phantom	Abdomen	●	○	●	16 stereo pairs (256 × 256)
Colonoscopy Depth [Yang *et al.*, 2023]	Phantom	Colon	●	○	○	16k frames (256 × 256)
Surgical Datasets (Clinical Procedures)						
Gastric-X [Lu *et al.*, 2026]	Surgical	Stomach	●	●	●	7.1k CT scans (288 × 288)
EndoMapper [Azagra *et al.*, 2023]	Surgical	Colon	●	●	○	59 seq. (320 × 240)
C3VD [Bobrow *et al.*, 2023]	Surgical	Colon	●	●	○	10k frames (675 × 540)
Stereo surgical [Chen *et al.*, 2023]	Surgical	Lymph	●	○	○	128GB (1920 × 1080)
Colon10k [Posner *et al.*, 2023]	Surgical	Colon	○	○	●	10k frames (270 × 216)
CVC-ClinicDB [Mahmood *et al.*, 2018]	Surgical	Colon	○	○	●	612 frames (576 × 768)
Oblique/En-face [Wang *et al.*, 2024]	Surgical	Colon	●	○	○	94 seq. (270 × 216)
LDP-PolypVideo [Cheng *et al.*, 2021]	Surgical	Colon	○	○	●	4.2M frames (560 × 480)
Sinus Surgery [Qin *et al.*, 2020]	Surgical	Sinus	○	○	●	9k frames (256 × 256)

Table 1: Overview of datasets for endoscopic vision. The table categorizes datasets into Surgical (clinical procedures), Phantom (physical models), and Synthetic (computer-generated). Supported tasks are indicated based on ground truth availability: Rec. (Reconstruction/Depth), Loc. (Localization/Pose), and Seg. (Segmentation/Detection). ● indicates fully supported, while ○ denotes lack of support.

5.2 Surgical Vision-Language-Action Models

Vision-Language-Action (VLA) models bridge high-level clinical intent and low-level robotic execution, enabling agents to interpret complex instructions in unstructured surgical environments. Current research optimizes these systems across two primary dimensions: task execution quality and physical motion control.

Embodied Decision Making and Execution. Reliable surgical VLAs require strict adherence to clinical protocols and expert heuristics. Beyond traditional imitation learning, MedFlamingo-PPO [Muppidi and Radfar, 2025] and SER [Fan and Chen, 2025] embed procedural rules and virtual fixtures as explicit policy constraints to circumvent unsafe exploration. Recent advancements focus on expert-driven alignment to incorporate refined clinical logic. For instance, SurgWorld [He *et al.*, 2025] leverages world modeling to align surgical actions with expert text descriptions. This alignment improves the model's generalization ability across different anatomical structures and techniques, ensuring its robustness to intraoperative unpredictability.

Hierarchical Control. Hierarchical control decomposes surgical tasks into distinct layers to separate high-level planning from low-level execution. Precise surgical intervention requires continuous and responsive motion. Discrete action representations often introduce discontinuities, which are unacceptable in surgery due to the risk of tissue damage. To ensure smooth execution, EndoVLA [Ng *et al.*, 2025] utilizes reinforcement learning and continuous control formulations. Additionally, although SRT-H [Fan and Chen, 2025] splits planning and control to reduce delay, this separation causes the system to miss sudden events because it only observes the environment intermittently. To mitigate these risks, SAFe-Copilot [Nguyen *et al.*, 2025] introduces semantic arbitration. This framework utilizes safety contracts to dynamically reallocate control authority between the agent and the surgeon, maintaining real-time responsiveness during critical incidents like bleeding or tissue slippage.

Discussion. Current surgical VLA models rely on modular designs that separate perception, reasoning, and control. While this improves stability and reduces latency, it limits the system's ability to respond to unexpected events that unfold continuously during surgery. Motion smoothness and rule-based safety can be enforced locally, but they do not provide the global awareness needed to handle anatomical variability and procedural deviations. Even with the introduction of arbitration mechanisms, achieving reliable real-time responsiveness remains difficult. Future progress depends on the tighter integration of perception and action, enabling decisions to adapt in real time as the surgical scene evolves.

6 Datasets

Data quality and diversity determine the performance of endoscopic vision models. This section analyzes technical challenges, categorizes available datasets, and discusses regulatory constraints governing their clinical application.

6.1 Challenges in Dataset Construction

The development of endoscopic models is primarily hindered by a fundamental trilemma among label precision, visual realism, and data diversity. These constraints necessitate a multi-source data paradigm.

First, real surgical videos provide authentic visuals but lack precise labels. Synthetic data offers accurate labels but lacks realistic textures. This gap makes it difficult to train models that achieve both visual robustness and geometric accuracy. Second, surgical data labeling requires specialized clinical knowledge. Manual labeling by experts is expensive and slow. Differences in judgment among annotators introduce noise. Automated label generation in simulated environments is a practical way to acquire large-scale data. Third, most datasets focus on routine procedures, with rare events such as surgical complications being insufficiently represented. This distribution imbalance limits generalization in extreme cases.

6.2 Taxonomy of Endoscopic Datasets

To address these challenges, the research community leverages a diverse set of data resources, which can be broadly categorized into synthetic, phantom and clinical surgical datasets, as summarized in Table 1.

Synthetic datasets use computer graphics to generate large-scale data with precise labels, enabling efficient training and low annotation costs, as well as the simulation of rare clinical scenarios that are difficult to capture *in vivo*. Phantom datasets rely on physical organ models and external tracking to provide accurate labels in controlled settings, offering a more realistic appearance than synthetic data while retaining the precision needed for technical validation under realistic sensing conditions. Clinical datasets are collected during real procedures and provide authentic depictions of tissue deformation and surgical artifacts, but suffer from limited and noisy annotations due to practical constraints. Recently, benchmarks like Gastric-X [Lu *et al.*, 2026] have advanced this by aligning multi-phase CT and endoscopic imagery with biochemical indicators to reflect holistic clinical workflows.

7 Future Works

Physically Grounded Clinical Modeling. Most current models rely heavily on visual correlations, with limited incorporation of physical laws or medical principles. Yet safe surgical assistance depends on understanding how tissue deforms, how forces propagate, and how actions affect anatomy and risk. Future work should integrate physical modeling and medical knowledge directly into learning frameworks, enabling systems to reason about deformation, contact, and procedural constraints rather than inferring them implicitly from appearance.

Real-Time Closed-Loop Integration. Many existing approaches are developed without strict real-time requirements, assuming delayed inference or offline optimization. In surgery, however, models must operate under tight latency budgets. Future research should treat real-time operation as a core constraint, favoring models that support incremental updates, low-latency inference, and predictable execution while balancing model expressiveness.

Integrated System. Current systems often separate perception, understanding, and control into loosely connected stages, leading to delayed feedback and brittle behavior. A key future direction is the development of integrated architectures in which perception continuously informs reasoning and reasoning directly shapes action in a closed loop. Such integration would allow surgical systems to adapt online to evolving anatomy and unexpected events.

8 Ethical Considerations

The clinical deployment of AI systems for endoscopic surgery is constrained by regulatory, safety, and ethical requirements. Translating research advances into operating-room practice requires clear accountability, responsible data governance, and sustained safety assurances.

Data Privacy. Surgical data constitute sensitive patient information, and regulatory restrictions on data sharing often lead to fragmented datasets and institution-specific models. This limits generalizability and raises concerns across clinical settings. Clear standards for data anonymization, data use, and accountability are therefore essential for both ethical compliance and robust model development.

Liability and Explainability. As AI systems increasingly influence surgical decisions and actions, opaque model behavior complicates responsibility assignment in the event of failure. Without transparent and auditable decision processes, regulatory approval is difficult to obtain and clinical trust remains limited, restricting the level of autonomy that can be safely delegated to AI systems.

Safe Deployment. Model deployment requires ongoing oversight beyond initial approval. Surgical environments evolve over time, and models must remain reliable under distribution shifts and rare events. Ethical deployment therefore demands supervision, uncertainty awareness, and post-deployment monitoring to ensure sustained safety.

9 Conclusion

Endoscopic surgical intelligence is moving beyond passive video analysis toward systems that actively support intraoperative decision-making. In this survey, we have framed this evolution through a perception-to-action pipeline. Although notable progress has been made at individual stages, existing methods are often developed in isolation, limiting real-time performance and overall reliability. Because surgery is dynamic and safety-critical, effective systems must be physically grounded, temporally responsive, and tightly integrated. Future advances will therefore depend on unifying perception, reasoning, and control into coherent, real-time frameworks that respect both physical and clinical constraints. We hope this survey draws greater attention to these challenges and highlights new research directions that could stimulate the further development of endoscopic surgical intelligence.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62471306), the Shenzhen Medical Research Fund (D2402010), and the Shenzhen Major Scientific and Technological Project (KJZD20230923114615031).

Contribution Statement

Juyan Ba and Hao Chen contributed equally.

References

[Allan *et al.*, 2021] Max Allan, Jonathan Mcleod, Congcong Wang, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *TMI*, 2021.

[Almalioglu *et al.*, 2020] Yasin Almalioglu, Kutsev Bengisu Ozyoruk, Abdulkadir Gokce, et al. Endol2h: deep super-resolution for capsule endoscopy. *IEEE Transactions on Medical Imaging*, 39(12):4297–4309, 2020.

[Azagra *et al.*, 2023] Pablo Azagra, Carlos Sostres, Ángel Ferrández, et al. Endomapper dataset of complete calibrated endoscopy procedures. *Sci. Data*, 2023.

[Bobrow *et al.*, 2023] Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, et al. Colonoscopy 3D video dataset with paired depth from 2D-3D registration. *MedIA*, 2023.

[Chen *et al.*, 2023] Ziyang Chen, Aldo Marzullo, Davide Alberti, et al. FRSR: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery. *Comput. Biol. Med.*, 163, 2023.

[Chen *et al.*, 2024] Tong Chen, Qingcheng Lyu, Long Bai, et al. LighTDiff: Surgical endoscopic image low-light enhancement with T-diffusion. In *MICCAI*, 2024.

[Chen *et al.*, 2025a] Hao Chen, Rui Yin, Yifan Chen, et al. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. In *Proceedings of the International Conference on Learning Representations (ICLR) 2026*, 2025.

[Chen *et al.*, 2025b] Jialei Chen, Xin Zhang, Mobarak I Hoque, et al. SurgicalGS: Dynamic 3d gaussian splatting for accurate robotic-assisted surgical scene reconstruction. In *MICCAI*, 2025.

[Cheng *et al.*, 2021] Kai Cheng, Yiting Ma, Bin Sun, et al. Depth estimation for colonoscopy images with self-supervised learning from videos. In *MICCAI*, 2021.

[Edwards *et al.*, 2022] PJ Eddie Edwards, Dimitris Psychogyios, Stefanie Speidel, et al. SERV-CT: A disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction. *MedIA*, 2022.

[Fan and Chen, 2025] Ke Fan, Ziyang Chen. Integrating reinforcement learning and virtual fixtures for safer automatic robotic surgery. *IEEE Robotics and Automation Letters*, 2025.

[He *et al.*, 2025] Yufan He, Pengfei Guo, Mengya Xu, et al. Surgworld: Learning surgical robot policies from videos via world modeling. *preprint arXiv:2512.23162*, 2025.

[Huang *et al.*, 2025] Yiming Huang, Long Bai, Beilei Cui, et al. SurgTPGS: Semantic 3D surgical scene understanding with text promptable gaussian splatting. In *MICCAI*, 2025.

[Jaspers *et al.*, 2025] Tim JM Jaspers, Ronald LPD de Jong, Yiping Li, et al. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, 2025.

[Kerbl *et al.*, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023.

[Kim *et al.*, 2025] Ji Woong Kim, Juo-Tung Chen, Pascal Hansen, et al. SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. *Science Robotics*, 2025.

[Li *et al.*, 2024] Chenxin Li, Hengyu Liu, Yifan Liu, et al. Endora: Video generation models as endoscopy simulators. In *MICCAI*, 2024.

[Li *et al.*, 2025a] Hao Li, Daiwei Lu, Jiacheng Wang, et al. Endostreamdepth: Temporally consistent monocular depth estimation for endoscopic video streams. *preprint arXiv:2512.18159*, 2025.

[Li *et al.*, 2025b] Ruocheng Li, Lei Wu, Jingjun Gu, et al. MoE-SAM: Enhancing SAM for medical image segmentation with mixture-of-experts. In *MICCAI*, 2025.

[Lin *et al.*, 2024] Hongbin Lin, Bin Li, Chun Wai Wong, et al. World models for general surgical grasping. In *RSS*, 2024.

[Liu *et al.*, 2025] Yifan Liu, Chenxin Li, Hengyu Li, et al. Foundation model-guided gaussian splatting for 4D reconstruction of deformable tissues. *TMI*, 2025.

[Low *et al.*, 2025] Chang Han Low, Zhu Zhuo, Ziyue Wang, et al. Cares: Collaborative agentic reasoning for error detection in surgery. *preprint arXiv:2508.08764*, 2025.

[Lu *et al.*, 2026] Sheng Lu, Hao Chen, Rui Yin, et al. Gastric-x: A multimodal multi-phase benchmark dataset for advancing vision-language models in gastric cancer analysis. In *CVPR*, 2026.

[Mahmood *et al.*, 2018] Faisal Mahmood, Richard Chen, Nicholas J Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *TMI*, 37, 2018.

[Mildenhall *et al.*, 2020] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, et al. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[Muppidi and Radfar, 2025] Akshay Muppidi, Martin Radfar. Medical vision language models as policies for robotic surgery. In *IEEE CAI*, 2025.

[Ng *et al.*, 2025] Chi Kit Ng, Long Bai, Guankun Wang, et al. Endovla: Dual-phase vision-language-action for precise autonomous tracking in endoscopy. In *CoRL*, 2025.

[Nguyen *et al.*, 2025] Phat Nguyen, Erfan Aasi, Shiva Sreeram, et al. Safe-copilot: Unified shared autonomy framework. *preprint arXiv:2511.04664*, 2025.

[Ozyoruk *et al.*, 2021] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, et al. Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *MedIA*, 2021.

[Penza *et al.*, 2018] Veronica Penza, Andrea S Ciullo, Sara Moccia, et al. Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms. *Int. J. Med. Robot.*, 2018.

[Posner *et al.*, 2023] Erez Posner, Adi Zholkover, Netanel Frank, Moshe Bouhnik. C^3Fusion: Consistent contrastive colon fusion, towards deep SLAM in colonoscopy. In *Proc. ShapeMI (MICCAI)*, 2023.

[Qin *et al.*, 2020] Fangbo Qin, Shan Lin, Yangming Li, et al. Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision. *RA-L*, 5, 2020.

[Rau *et al.*, 2019] Anita Rau, PJ Eddie Edwards, Omer F Ahmad, et al. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *IJCARS*, 2019.

[Shen *et al.*, 2025] Yiqing Shen, Chenjia Li, Bohan Liu, others. Operating room workflow analysis via reasoning segmentation over digital twins. In *MICCAI*, 2025.

[Sommersperger *et al.*, 2023] Michael Sommersperger, Shervin Dehghani, Philipp Matten, et al. Semantic Virtual Shadows (svs) for improved perception in 4d oct guided surgery. In *MICCAI*, 2023.

[Trinh *et al.*, 2024] Quoc-Huy Trinh, Hai-Dang Nguyen, Ngoc, others. SAM-EG: Segment anything model with edge guidance framework for efficient polyp segmentation. *preprint arXiv:2406.14819*, 2024.

[Walimbe *et al.*, 2025] Soham Walimbe, Britty Baby, Vinkle Srivastav, Nicolas Padoy. Adaptation of multi-modal representation models for multi-task surgical computer vision. In *MICCAI*, 2025.

[Wang *et al.*, 2024] Shuxian Wang, Akshay Paruchuri, Zhaoxi Zhang, et al. Structure-preserving image translation for depth estimation in colonoscopy. In *MICCAI*, 2024.

[Wang *et al.*, 2025] Zipei Wang, Sitian Pan, Mengjie Fang, others. Cholecmamba: A mamba-based multimodal reasoning model for cholecystectomy surgery. In *MICCAI*, 2025.

[Xie *et al.*, 2024] Tianyi Xie, Zeshun Zong, Yuxing Qiu, et al. PhysGaussian: Physics-integrated 3D gaussians for generative dynamics. In *CVPR*, 2024.

[Xu *et al.*, 2025a] Handing Xu, Zhenguo Nie, Tairan Peng, et al. DGGAN: Degradation guided generative adversarial network for real-time endoscopic video enhancement. *preprint arXiv:2512.07253*, 2025.

[Xu *et al.*, 2025b] Jinjing Xu, Chenyang Li, Peng Liu, et al. T^2GS: Comprehensive reconstruction of dynamic surgical scenes with gaussian splatting. In *MICCAI*, 2025.

[Xu *et al.*, 2025c] Miao Xu, Xiangyu Zhu, Jinlin Wu, et al. Reconstructing 3D hand-instrument interaction from a single 2D image in medical scenes. In *MICCAI*, 2025.

[Yang *et al.*, 2023] Yongming Yang, Shuwei Shao, Tao Yang, et al. A geometry-aware deep network for depth estimation in monocular endoscopy. *Eng. Appl. Artif. Intell.*, 2023.

[Yang *et al.*, 2024a] Shu Yang, Luyang Luo, Qiong Wang, Hao Chen. Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In *MICCAI*, 2024.

[Yang *et al.*, 2024b] Zhenya Yang, Kai Chen, Yonghao Long, Qi Dou. Simendogs: Efficient data-driven scene simulation using robotic surgery videos via physics-embedded 3d gaussians. *preprint arXiv:2405.00956*, 2024.

[Yao *et al.*, 2025a] Jieru Yao, Xueran Li, Qiang Xie, et al. LLaVA-Endo: A large language-and-vision assistant for gastrointestinal endoscopy. *Front. Comput. Sci.*, 2025.

[Yao *et al.*, 2025b] Xinning Yao, Bo Liu, Bojian Li, others. Endoufm: Utilizing foundation models for monocular depth estimation of endoscopic images. *preprint arXiv:2508.17916*, 2025.

[Yin *et al.*, 2025] Ming Yin, Fu Wang, Xujiong Ye, et al. MA-SAM2: Memory-augmented SAM 2 for training-free surgical video segmentation. In *MICCAI*, 2025.

[Yuan *et al.*, 2024] Kun Yuan, Vinkle Srivastav, Nassir Navab, Nicolas Padoy. HecVL: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *MICCAI*, 2024.

[Yue *et al.*, 2024] Wenxi Yue, Jing Zhang, Kun Hu, et al. SurgicalSAM: Efficient class promptable surgical instrument segmentation. In *AAAI*, 2024.

[Zeinoddin *et al.*, 2024] Mona Sheikh Zeinoddin, Chiara Lena, Jiongqi Qu, et al. DARES: Depth anything in robotic endoscopic surgery with self-supervised vector-lora of the foundation model. In *ECCV*, 2024.

[Zeng *et al.*, 2025] Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *preprint arXiv:2506.02555*, 2025.

[Zhang *et al.*, 2020] Shuai Zhang, Liang Zhao, Shoudong Huang, et al. A template-based 3D reconstruction of colon structures and textures from stereo colonoscopic images. *IEEE Trans. Med. Robot. Bionics*, 2020.

[Zhang *et al.*, 2025] Tiyao Zhang, Xue Yuan, Hongze Xu. Surgical instrument segmentation via Segment-Then-Classify framework with instance-level spatiotemporal consistency modeling. *Journal of Imaging*, 2025.

[Zheng *et al.*, 2023] Qingyuan Zheng, Rui Yang, Xinmiao Ni, others. Development and validation of a deep learning-based laparoscopic system for improving video quality. *IJCARS*, 2023.

[Zhu *et al.*, 2025] Yinna Zhu, Zijing Zhang, Zhenyang Li, et al. RVDNet: A reference-guided video desmoking network for endoscopic thyroid surgery. *IEEE Access*, 2025.