

Large-scale self-supervised video foundation model for intelligent surgery

Received: 21 August 2025

Accepted: 22 January 2026

Cite this article as: Yang, S., Zhou, F., Mayer, L. *et al.* Large-scale self-supervised video foundation model for intelligent surgery. *npj Digit. Med.* (2026). <https://doi.org/10.1038/s41746-026-02403-0>

Shu Yang, Fengtao Zhou, Leon Mayer, Fuxiang Huang, Yiliang Chen, Yihui Wang, Sunan He, Yuxiang Nie, Xi Wang, Yueming Jin, Huihui Sun, Shuchang Xu, Alex Qinyang Liu, Zheng Li, Jing Qin, Jeremy YuenChun Teoh, Lena Maier-Hein & Hao Chen

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Large-scale Self-supervised Video Foundation Model for Intelligent Surgery

Shu Yang¹, Fengtao Zhou¹, Leon Mayer^{2,3}, Fuxiang Huang¹, Yiliang Chen⁴, Yihui Wang¹, Sunan He¹, Yuxiang Nie¹, Xi Wang¹, Yueming Jin^{5,6}, Huihui Sun⁷, Shuchang Xu⁷, Alex Qinyang Liu⁸, Zheng Li⁸, Jing Qin⁴, Jeremy YuenChun Teoh⁸, Lena Maier-Hein^{2,3,9,10,11,*}, and Hao Chen^{1,12,13,14,15,*}

¹Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

²Division of Intelligent Medical Systems, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany

³Faculty of Medicine, Heidelberg University Hospital, Heidelberg, Germany

⁴School of Nursing, The Hong Kong Polytechnic University, Hong Kong SAR, China

⁵Department of Biomedical Engineering, National University of Singapore, Singapore

⁶Department of Electrical and Computer Engineering, National University of Singapore, Singapore

⁷Department of Gastroenterology of Tongji Hospital, School of Medicine, Tongji University, Shanghai, China

⁸Department of Surgery, The Chinese University of Hong Kong, Hong Kong SAR, China

⁹HI Helmholtz Imaging, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany

¹⁰Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany

¹¹National Center for Tumor Diseases (NCT), NCT Heidelberg, Heidelberg, Germany

¹²Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

¹³Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China

¹⁴State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology, Hong Kong SAR, China

¹⁵Shenzhen-Hong Kong Collaborative Innovation Research Institute, The Hong Kong University of Science and Technology, Shenzhen, China

* e-mail: jhc@ust.hk

ABSTRACT

Computer-Assisted Intervention has the potential to revolutionize modern surgery, with surgical scene understanding serving as a critical component in supporting decision-making and improving procedural efficacy. While existing AI-driven approaches alleviate annotation burdens via self-supervised spatial representation learning, their lack of explicit temporal modeling during pre-training fundamentally restricts the capture of dynamic surgical contexts, resulting in incomplete spatiotemporal understanding. In this work, we introduce the first video-level surgical pre-training framework that enables joint spatiotemporal representation learning from large-scale surgical video data. To achieve this, we constructed a large-scale surgical video dataset comprising 3,650 videos and 3.55 million frames, spanning more than 20 surgical procedures and over 10 anatomical structures. Building upon this dataset, we propose **SurgVISTA** (**S**urgical **V**ideo-level **S**patial-**T**emporal **A**rchitecture), a reconstruction-based pre-training method that jointly captures intricate spatial structures and temporal dynamics. Additionally, SurgVISTA incorporates image-level knowledge distillation guided by a surgery-specific expert model to enhance the learning of fine-grained anatomical and semantic features. To validate its effectiveness, we established a comprehensive benchmark comprising 13 video-level datasets spanning six surgical procedures across four tasks. Extensive experiments demonstrate that SurgVISTA consistently outperforms both natural- and surgical-domain pre-trained models, demonstrating strong potential to advance intelligent surgical systems in clinically meaningful scenarios.

Keywords: Computer-Assisted Intervention, Surgical Scene Understanding, Video-level Surgical Foundation Model

1 Introduction

Computer-Assisted Intervention (CAI) has recently made significant strides in integrating sophisticated artificial intelligence (AI) technologies into clinical workflows to optimize preoperative planning, intraoperative execution, and postoperative assessment.

A fundamental component of CAI is surgical scene understanding, which involves the comprehensive analysis and interpretation of intricate surgical activities and tissue interactions to support informed decision-making, improve procedural effectiveness, and ensure operational safety¹⁻³. This interdisciplinary endeavor aims to provide critical insights through continuous monitoring of intraoperative workflows and procedural execution, thereby facilitating surgical process optimization and promoting improved patient outcomes. However, developing AI-driven approaches for comprehensive surgical scene understanding encounters a significant challenge: the heavy reliance on expert-annotated surgical datasets. The SNOMED-CT International Edition (April 2025) catalogues thousands of procedures that differ markedly in prevalence and technical complexity. Collecting adequately annotated video data for this long-tail distribution is prohibitively expensive and labour-intensive, resulting in surgical datasets that cover only a limited and incomplete spectrum of procedures. Consequently, the restricted procedural scope of existing surgical datasets hampers model generalizability, highlighting the inadequacy of task-specific and narrow AI in addressing the heterogeneous demands of surgical practice⁴, and emphasizing the need for procedure-agnostic methodologies capable of generalizing across diverse clinical contexts. In fact, the international Surgical Data Science Initiative^{5,6} has recently identified the methodological addressing of surgical data sparsity as a key next step in the field⁷.

Inspired by the rapid advancements in self-supervised learning (SSL) paradigm within the natural domain⁸⁻¹⁰ and its demonstrated effectiveness across various medical disciplines¹¹⁻¹⁵, leveraging large-scale surgical data for SSL-based pre-training has emerged as a promising direction to extract robust and transferable representations explicitly tailored to surgical contexts. Despite significant achievements of SSL in various medical fields such as radiology^{11,13} and pathology^{12,15}, its application within the surgical field remains nascent and underexplored. Recent works¹⁶⁻¹⁹ have begun to specifically explore the SSL paradigm in the surgical domain, demonstrating initial successes in developing surgical foundation models and substantially improving surgical feature representation learning for downstream tasks. However, these methods predominantly focus on image-level pre-training for effective spatial representation learning, while neglecting the learning of temporal dynamics essential for comprehensive spatiotemporal representations. Consequently, they follow a mainstream two-stage paradigm²⁰, where temporal dependencies are modeled during fine-tuning using external temporal modules applied to static and pre-extracted spatial features, resulting in a decoupled modeling process that hinders joint spatiotemporal representation learning. In practice, this decoupling is suboptimal for surgical video analysis, where seamless integration of spatial and temporal information is critical to address inherent complexities such as visual ambiguities from occlusions, motion blur, and procedure-specific variability⁴. Furthermore, a significant discrepancy exists between the upstream pre-training and downstream fine-tuning, as these methods adopt image-level pre-training combined with a two-stage fine-tuning framework. This fragmented pipeline hinders the unified learning of spatial and temporal representations: while spatial features benefit from a well pre-trained backbone, temporal modeling starts from a randomly initialized module. As a result, image-level pre-trained methods struggle to leverage the rich temporal dynamics embedded in large-scale surgical video data, instead depending on limited downstream annotations to learn temporal dependencies. This piecemeal optimization potentially compromises the robustness of spatiotemporal modeling, hindering generalizability across diverse downstream applications and resulting in suboptimal performance. These limitations highlight the urgent need for new approaches that holistically integrate spatial and temporal learning during pre-training to enhance the comprehensive surgical scene understanding.

This study advocates for explicit and unified spatiotemporal representation learning during the pre-training phase to fully exploit the potential of large-scale surgical video data, thereby initiating a paradigm shift from static image-based modeling to dynamic video-level representation learning. Such integration effectively facilitates the extraction of robust, transferable, and discriminative spatiotemporal feature representations, thereby advancing comprehensive surgical scene understanding. To support this objective, we constructed a publicly available, large-scale surgical video dataset comprising two essential components: SurgPub and SurgWeb, which collectively span varied anatomical structures and surgical procedures, providing the variability and complexity necessary for robust spatiotemporal representation learning. In Fig. 1a, we present an overview of the key anatomical structures and associated surgical procedures, accompanied by video-level statistical analyses and frame-level distribution analyses. SurgPub is constructed from eight publicly available surgical video datasets²¹⁻²⁷, encompassing 274 videos with a total of 1.20 million frames. Complementing SurgPub, SurgWeb comprises 3,376 surgical recordings sourced from online platforms, totaling 2.35 million frames, spanning diverse surgical procedures and anatomical regions encountered in extensive clinical scenarios. The resulting dataset encompasses over 20 surgical procedures across more than 10 anatomical structures, thereby promoting the learning of robust and generalizable spatiotemporal representations. Furthermore, we envision this curated resource as a catalyst for advancing more robust and generalized surgery AI models, while empowering researchers to tackle complex challenges in surgical scene understanding.

To transcend static image-based pre-training paradigms and establish video-level representation learning as the foundation for surgical AI, we pioneer **SurgVISTA** (**Surgical Video-level Spatial-Temporal Architecture**), the first surgical video foundation model explicitly designed to capture robust and discriminative spatiotemporal patterns across diverse surgical scenarios. Unlike conventional methods restricted to static frame-level representations, SurgVISTA integrates spatial and temporal modeling into a unified pre-training framework, simultaneously encoding fine-grained anatomical details and temporal

dynamics. As depicted in Fig. 1b, SurgVISTA employs an asymmetric encoder-decoder architecture, comprising a unified encoder for joint spatiotemporal dependencies modeling and two dedicated decoders. The reconstruction decoder learns to restore masked video regions, thereby compelling the model to extract spatial structures and temporal dynamics intrinsic to surgical activities. Concurrently, the refinement decoder leverages knowledge distillation from a powerful image-level expert model¹⁸, compensating for detail loss from temporal abstraction while simultaneously reinforcing the discriminative capacity of spatial features. Through joint spatiotemporal modeling on large-scale surgical video data, SurgVISTA acquires highly generalizable representations that effectively encapsulate the complexity and variability of surgical scenarios, enabling seamless adaptation to downstream tasks without additional temporal components, thus significantly bridging the gap between pre-training and fine-tuning phases. Overall, SurgVISTA establishes a scalable and pretraining-driven foundation for surgical scene understanding, significantly enhancing robust spatiotemporal modeling capabilities for complex intraoperative scenarios.

To verify the effectiveness of SurgVISTA, we conducted a comprehensive video-level evaluation across 13 datasets, covering six surgical procedures and four distinct downstream tasks. As illustrated in Fig. 1c, we provide a visual abstraction of diverse downstream tasks alongside the evaluated datasets, accompanied by comparative performance analyses of models pre-trained on natural-domain and surgical-domain data. Specifically, we began by comparing image-level and video-level pre-trained models within the natural domain, highlighting the superior performance of SurgVISTA and underscoring the necessity of domain-aligned video-level pre-training. Furthermore, we benchmarked SurgVISTA against the existing surgical foundation models, with a particular focus on surgical workflow recognition, demonstrating the advantages of joint spatiotemporal modeling empowered by large-scale surgical video pre-training in enhancing downstream performance. Despite the absence of task-specific data during the pre-training phase, SurgVISTA demonstrates exceptional generalization capabilities on anatomically and procedurally divergent surgical datasets, reinforcing its robustness and adaptability to a wide spectrum of surgical scenarios.

2 Results

SurgVISTA outperforms natural-domain image- and video-level pre-trained methods on surgical tasks

In this section, we first evaluate the effectiveness of SurgVISTA on surgical workflow recognition, comparing it against both image-level^{8,9,28} and video-level²⁹⁻³² pre-trained methods within the natural domain. Experimental results are illustrated in Fig. 2, accompanied by 95% confidence intervals to quantify uncertainty and P-values to assess statistical significance. Specifically, Fig. 2a-d present results on in-domain datasets with related procedures in the pre-training data, while Fig. 2e-h illustrate results on out-of-domain datasets, involving surgical procedures excluded from the pre-training data. To provide a comprehensive understanding of the model's generalization capacity, we quantify the proportion of pre-training data associated with related surgical procedures and anatomical structures, thereby highlighting the extent of overlap between the pre-training data and downstream datasets. As illustrated in Fig. 2, natural-domain methods employing video-level pre-training consistently outperformed image-level counterparts across all evaluated datasets. On Cholec80, the average gains for video-level models over image-level ones reached 1.5%, 1.5%, and 2.8% in image-level accuracy, video-level accuracy, and phase-level Jaccard, respectively. This performance advantage became even more substantial on the out-of-domain Cataract-101 dataset, with average enhancements reaching 10.8% in image-level accuracy, 10.7% in video-level accuracy, and 19.6% in phase-level Jaccard, highlighting the pronounced benefits of video-level pre-training for surgical scene understanding. Furthermore, SurgVISTA consistently outperformed natural-domain pre-trained video-level methods across both in-domain and out-of-domain datasets. On representative in-domain benchmarks including Cholec80, M2CAI16-Workflow, and PmLR50, SurgVISTA achieved notable gains in image-level accuracy, video-level accuracy, and phase-level Jaccard over natural-domain methods. Despite minimal procedural overlap in the pre-training data (e.g., 1.13% for hysterectomy), SurgVISTA outperformed the best prior model²⁹ by 4.7%, 4.6%, and 7.3%, highlighting its robust generalization to underrepresented procedures. These in-domain results demonstrate SurgVISTA's capacity to acquire domain-relevant knowledge during pre-training, enabling improved downstream performance. In out-of-domain evaluations on ESD57, SurgVISTA surpassed VideoMAE²⁹ (pre-trained on UnlabeledHybrid³⁰) by 2.8%, 2.7%, and 5.4% on the three metrics, respectively. When evaluated on CATARACTS, Cataract-101, and Cataract-21 datasets, which are substantially different from the surgical procedures used during pre-training, SurgVISTA consistently outperformed all competing methods, reinforcing its generalizability across varied procedures and anatomical structures. We further evaluate SurgVISTA on additional critical surgical tasks beyond workflow recognition, providing a comprehensive assessment of its capacity for surgical scene understanding. As illustrated in Fig. 3, SurgVISTA consistently achieved superior performance compared to natural-domain pre-trained methods across multiple surgical tasks. For surgical action recognition on SurgicalActions160, it surpassed UMT³² by 1.8% in accuracy and 4.6% in mAP. For surgical triplet recognition, it outperformed MVD³¹ on CholecT50 by 5.8% (triplet mAP) and 5.9% (2-tuple mAP), and exceeded DINO⁸ on Prostate21 by 8.0% and 5.9%, suggesting effective modeling of instrument-target interactions. For surgical skill assessment, SurgVISTA achieved relative improvements of 8.3% in accuracy and 5.4% in mAP over MVD³¹ on Cholec80-CVS, and surpassed UMT³² with relative gains of 0.1% in accuracy and 1.9% in mAP on Endoscapes-CVS. These results highlight its capacity to capture intricate anatomical structures and motion patterns critical for evaluating surgical expertise.

Overall, these tasks constitute core components of comprehensive surgical scene understanding. Extensive evaluations across 13 diverse datasets demonstrate that SurgVISTA consistently surpasses natural-domain pre-trained methods across a wide range of tasks, surgical procedures, and anatomical structures.

SurgVISTA outperforms existing surgical-domain pre-trained methods

To further investigate the effectiveness of SurgVISTA in surgical scene understanding, we conducted a comprehensive comparison with existing surgical-domain pre-trained methods. Given that current surgical-domain pre-trained methods^{16–19} primarily emphasize the surgical workflow recognition task, we performed extensive evaluations on six surgical workflow recognition datasets encompassing diverse procedures and anatomical structures. Experimental results with 95% confidence intervals are presented in Fig. 4, providing a comprehensive comparison of model performance across evaluated datasets. Specifically, Fig. 4a-c represent results on in-domain datasets, while Fig. 4d-e illustrate the results on out-of-domain datasets. Overall, SurgVISTA significantly outperformed all comparative methods across all evaluation metrics and datasets. On in-domain datasets, SurgVISTA exhibited substantial improvements over the best-performing method EndoSSL¹⁸. Notably, EndoSSL is an image-level method pre-trained on 23.3 million private laparoscopic frames using the ViT-L architecture²⁸. Specifically, on the Cholec80 dataset, SurgVISTA achieved high performance, with image-level accuracy of 91.5%, video-level accuracy of 91.5%, phase-level precision of 87.3%, phase-level recall of 87.7%, and phase-level Jaccard of 78.1%, significantly outperforming other methods. On the M2CAI16-Workflow dataset, compared to EndoSSL¹⁸, SurgVISTA attained modest gains of 1.1% and 1.7% in image-level and video-level accuracies, while achieving significantly higher performance in phase-level metrics, with improvements of 11.3% in precision, 9.1% in recall, and 7.5% in Jaccard. On the AutoLaparo dataset, SurgVISTA exhibited remarkable improvements across multiple granularities compared to EndoSSL¹⁸, where it achieved an increase of 2.2% in image-level accuracy, 2.2% in video-level accuracy, 11.4% in phase-level precision, 4.8% in phase-level recall, and 5.1% in phase-level Jaccard. In out-of-domain evaluations involving surgical procedures not encountered during pre-training, SurgVISTA consistently achieved superior performance. For the CATARACTS dataset, containing numerous workflow categories and complex phase transitions, SurgVISTA surpassed all competing methods, and achieved improvements of 14.8%, 11.1% and 14.3% in phase-level metrics compared to SelfSupSurg¹⁶. Similarly, for the Cataract-101 and Cataract-21 datasets, SurgVISTA consistently surpassed existing models.

SurgVISTA benefits from large-scale and diverse surgical pre-training data

Data scaling laws describe how model performance varies as the amount of pre-training data increases, serving as a fundamental principle in developing robust foundation models. Extensive natural-domain studies have demonstrated that scaling pre-training data significantly enhances model generalization, robustness, and overall performance across various tasks. However, the investigation of data scaling laws within the surgical domain, particularly for video-level pre-training, remains relatively underexplored. To elucidate the relationship between data scale and performance within the surgical domain, we systematically examined data scaling laws by pre-training SurgVISTA on a series of sub-datasets of varying sizes. As depicted in Fig. 5a, we constructed five progressively expanding sub-datasets, denoted as Settings A through E, where each successive setting incorporated all data from the preceding settings, culminated in Setting E representing the full dataset. Sub-datasets from Settings A to D were exclusively sourced from SurgPub, with each subsequent setting incorporating additional public datasets to examine the impact of diverse surgical procedures on model performance and generalization capabilities. The detailed composition and distribution are illustrated in Supplementary Table 2. To assess model performance and generalization capabilities, extensive experiments were conducted across nine diverse video-level surgical datasets, spanning various surgical procedures and anatomical structures. The experimental results are shown in Fig. 5b-n, where line charts indicate the absolute data quantities of procedure-specific and full pre-training data, while the accompanying donut charts illustrate the corresponding proportions. Overall, as the scale of pre-training data progressively increases, all downstream datasets demonstrate a consistent and robust trend of performance enhancement. For Cholec80, M2CAI16-Workflow, CholecT50, Cholec80-CVS and Endoscapes-CVS which focus on laparoscopic cholecystectomy procedures, we observed an overall trend of performance improvement from Setting A to Setting C, with the proportion of procedure-specific data gradually decreasing from 100% to 47.3%, indicating the model's ability to effectively leverage relevant surgical content. Even in Settings D and E, where the proportion of procedure-specific data decreased to 17.4% and 15.1% respectively, consistent performance improvements were observed, highlighting the model's robust generalization across low-overlap surgical scenarios. For PmLR50, which has 0% procedural overlap with the pre-training data from Settings A to D, performance continued to improve, achieving gains of 3.4% in image-level accuracy and 10.1% in phase-level Jaccard. In Setting E, despite the procedure-specific data accounting for only 1.9% of the total, the model achieved further improvements of 1.3% in image-level accuracy and 7.1% in phase-level Jaccard, demonstrating the effectiveness of large-scale surgical pre-training in enhancing representation transferability. For gynecologic laparoscopic surgery datasets, including AutoLaparo and SurgicalActions160, we observed a slight performance decline from Settings A to B, followed by a substantial improvement in Setting C, where the amount of pre-training data increases and procedure-relevant surgical videos were introduced. Specifically, from Setting A to Setting C, AutoLaparo

demonstrated gains of 1.3% in image-level accuracy and 3.9% in phase-level Jaccard, while SurgicalActions160 showed improvements of 8.7% in accuracy and 8.2% in mAP. In subsequent settings, the continued inclusion of unrelated surgical procedures further enhanced performance, where AutoLaparo achieved an additional improvement of 4.5% in image-level accuracy and 3.2% in phase-level Jaccard, while SurgicalActions160 attained a further 18.7% and 9.4% increase in accuracy and mAP, respectively. For out-of-domain cataract surgery datasets, including CATARACTS, Cataract-21, and Cataract-101, we still observed significant performance improvements, demonstrating the strong generalization capability of large-scale surgical pre-training in leveraging diverse procedural contexts to learn robust and transferable spatiotemporal representations. For the privately collected benchmarks Prostate21 and ESD57, we similarly observed substantial performance improvements as the scale and diversity of the pre-training corpus increased. Although these datasets originate from distinct clinical centers and were not part of any public surgical video repository, the gains achieved across the scaling settings demonstrate that SurgVISTA effectively transfers to previously unseen intraoperative environments and institution-specific visual characteristics. Comprehensive analyses demonstrate that SurgVISTA exhibits substantial performance improvements as the scale of pre-training data expands, even when incorporating irrelevant surgical procedures, highlighting its robustness and generalization capacity. These performance improvements are particularly pronounced for datasets involving complex surgical procedures with numerous workflow categories and intricate phase transitions. The systematic investigation of data scaling laws in this study offers valuable insights into the relationship between data volume and model effectiveness, providing guidance for future research and the development of more advanced and generalizable models for surgical scene understanding.

Ablation studies confirm the necessity of knowledge distillation

Surgical scene understanding necessitates advanced spatiotemporal modeling to effectively capture the complex and dynamic interactions between anatomical structures and surgical instruments, wherein effective spatial modeling is essential for fine-grained intra-frame understanding and for enriching the overall spatiotemporal feature representation. To further improve spatial feature representations within the integrated spatiotemporal learning framework, SurgVISTA incorporates knowledge distillation, leveraging the superior spatial feature extraction capabilities of image-level foundation model. To assess the effectiveness of knowledge distillation, we implemented an ablated variant of SurgVISTA without knowledge distillation, pre-trained on Settings A through D, and conducted extensive experiments across two downstream datasets. The experimental results are illustrated in Fig. 6a–b. Under Setting A, knowledge distillation yielded marginal changes on Cholec80, with a slight increase of 0.2% in image-level accuracy and a slight decrease of 0.6% in phase-level Jaccard, possibly due to the inclusion of procedure-specific pre-training data aligned with Cholec80. In contrast, more substantial improvements were observed on AutoLaparo, with improvements of 2.9% and 3.0% in image-level accuracy and phase-level Jaccard, respectively. Furthermore, for Cholec80, the performance gains from knowledge distillation became increasingly evident as the scale and diversity of pre-training data increased. For AutoLaparo, although Settings A and B involved the same surgical procedure that differed from AutoLaparo's, knowledge distillation yielded gains in Setting A but slightly decreased phase-level Jaccard in Setting B. As procedurally relevant data and overall dataset diversity increased in subsequent settings, the benefits of knowledge distillation became progressively more pronounced. Overall, knowledge distillation consistently improves performance, yielding significant improvements in both accuracy and Jaccard across two datasets and all four pre-trained settings, compared to the non-distilled counterparts. Complementary qualitative results are presented in Fig. 6c–d, where reconstruction fidelity is evaluated using mean squared error (MSE) to verify the model's ability to preserve fine-grained visual details. These visualizations demonstrate that knowledge distillation facilitates the learning of robust spatial representations to reconstruct video clips with structural integrity and visual coherence, particularly in regions containing anatomical structures and instrument-tissue interactions.

3 Discussion

To the best of our knowledge, this work is the first to explore video-level pre-training in the surgical domain using large-scale surgical video data. The following novel insights can be derived from our extensive experimental results.

Surgery-specific video-level pre-training outperforms natural-domain counterparts by enabling domain-aligned spatiotemporal representation learning. Experimental results reveal two critical insights: temporal modeling capabilities acquired during pre-training are essential for capturing dynamic surgical contexts, and domain-aligned pre-training surpasses domain-agnostic pre-training in addressing complex and variable surgical scenarios. While natural-domain video pre-training surpasses image-level methods by better modeling temporal dynamics, it lacks the surgery-specific knowledge required for precise interpretation of instruments, anatomy, and their interactions. In contrast, surgical-domain video pre-training yields more robust and discriminative spatiotemporal representations enriched with surgery-specific knowledge, reinforcing its importance for robust surgical scene understanding.

Video-level surgical pre-training outperforms image-level counterparts by enabling explicit spatiotemporal representation learning. Although image-level counterparts primarily focus on spatial representation learning from large-scale surgical images, the absence of explicit temporal modeling during pre-training limits their ability to capture dynamic surgical contexts.

Consequently, despite leveraging powerful surgery-specific spatial feature extractors, these methods inherently struggle to model temporal dependencies from purely spatial representations during fine-tuning, ultimately limiting spatiotemporal representation quality, particularly in complex surgical scenarios. SurgVISTA employs spatiotemporal representation learning during pre-training, effectively harnessing the intrinsic spatial and temporal cues embedded in video data. By integrating fine-grained spatial representations of instruments and anatomical structures with comprehensive temporal modeling encompassing action temporality, causality, and dynamic interactions, SurgVISTA provides a robust and generalizable foundation for surgical scene understanding. The demonstrated superiority highlights the critical necessity of video-level pre-training tailored to the surgical domain, facilitating the advancement of more sophisticated, reliable, and clinically applicable intelligent surgical systems.

Increasing the scale and diversity of surgical video data yields substantial improvements in both performance and generalization. Extensive experiments conclusively demonstrate that both factors are pivotal for improving accuracy and robustness across diverse surgical scenarios. First, scaling up pre-training data consistently enhances model performance and generalization, as reflected by SurgVISTA's significant gains across all evaluated datasets. In procedure-agnostic scenarios involving procedures not present in downstream datasets, SurgVISTA continues to benefit by learning more diverse visual and temporal patterns. In procedure-specific scenarios, incorporating additional relevant data reinforces domain-specific priors and facilitates discriminative feature learning, further improving task-specific performance. This highlights the importance of large-scale surgical video data for advancing comprehensive understanding of complex and diverse surgical scenarios. Second, increasing procedural diversity further boosts performance and generalization. As more diverse surgical procedures are incrementally integrated into the pre-training corpus, significant performance gains are observed across various surgical scenarios, indicating that learning from a wider procedural spectrum enables the model to capture shared spatial and temporal patterns that generalize beyond specific tasks. Enhancing procedural diversity facilitates the exploration of universal and transferable representations essential for understanding complex and diverse surgical scenarios. Nevertheless, our scaling analysis is based on cumulative pre-training settings whose composition is not perfectly distribution-matched across surgical procedures (e.g., cholecystectomy-heavy early settings), which may introduce procedure-specific biases. Despite this limitation, the highly consistent improvements observed across a broad range of in-domain and out-of-domain benchmarks indicate that the dominant effect comes from increasing both the scale and the procedural diversity of surgical pre-training data, rather than from procedure-specific bias.

Data splitting in surgical video analysis remains fundamentally constrained by the trade-off between maximizing training scale and preserving unbiased evaluation. Given the inherent complexity and procedural overlap in surgical video data, including shared instruments, anatomical structures, or activities, expanding the training set inevitably increases the risk of distributional leakage into the evaluation data. This presents a fundamental challenge: how to maximize training scale for generalization while rigorously maintaining evaluation integrity. In this study, we curated publicly available surgical video datasets with any form of annotation. To enhance the diversity and scale of pre-training, we selected several datasets specifically for representation learning. Among these, some were split according to their official training and testing partitions, with the testing portions held out to support downstream evaluation under strict isolation, while others were used exclusively during pre-training to expand the data scale. Additionally, a number of datasets were entirely excluded from the pre-training phase and reserved as unseen evaluation benchmarks to assess generalization across surgical domains. This strategy enables a rigorous assessment of model generalization while ensuring strict non-overlap between pre-training and evaluation sets. Although dynamic benchmarks like LiveBench have emerged in general machine learning, such adaptive evaluation remains challenging in surgical domains due to limited data access, high annotation costs, and strict privacy constraints.

SurgVISTA represents a robust and generalizable surgical video foundation model with clear advantages over both natural-domain and surgery-specific methods. Compared to natural-domain methods, SurgVISTA leverages large-scale surgical video data to acquire domain-specific knowledge that encompasses both static anatomical structures and surgical instruments, as well as dynamic intraoperative activities. This domain alignment results in more clinically relevant and effective performance in intelligent surgical applications. In contrast to existing surgery-specific methods, SurgVISTA pioneers a unified spatiotemporal representation learning paradigm, marking a transition from static image-based representation learning to dynamic video-level understanding. Through explicit modeling of spatiotemporal representation inherent in large-scale surgical video data, SurgVISTA demonstrates superior capacity to handle the complexity and variability of real-world surgical scenarios and exhibits improved generalization and transferability across heterogeneous procedural contexts. By delivering transferable and discriminative spatiotemporal features and enabling seamless adaptation to a broad range of downstream tasks, SurgVISTA serves as a solid foundation for the development of intelligent surgical systems with the potential to improve procedural safety and enhance clinical outcomes across diverse surgical scenarios. These results affirm that the generalized and robust spatiotemporal representations learned by SurgVISTA from large-scale surgical video data are pivotal for accurately interpreting complex surgical scenarios, significantly outperforming natural-domain methods lacking specialized surgical knowledge.

While SurgVISTA represents a substantial advancement in surgical video pre-training and spatiotemporal representation learning, several limitations remain, highlighting promising directions for future investigation. First, although we construct a large-scale video-level surgical dataset encompassing more than 20 surgical procedures and over 10 anatomical structures, it

remains insufficient to fully reflect the extensive procedural complexity and variability in real-world clinical practices. This limitation underscores the imperative for continuous expansion and diversification of the pre-training dataset to ensure broader coverage of surgical scenarios and enhance the generalizability of learned representations across diverse clinical environments. Beyond sheer data scale, future work should also explore data-centric research questions in surgical SSL. In particular, training foundation models under fixed dataset sizes with varying procedure distributions offers an intriguing direction for understanding how data composition affects model generalization. Second, while we establish the most comprehensive video-level benchmark encompassing 13 downstream datasets spanning six surgical procedures and four distinct tasks, the procedural diversity within the benchmark remains limited relative to the broader scope of the pre-training data. Future efforts should focus on extensive collaboration with clinical institutions to collect annotated datasets across various surgical procedures, focusing on the generation of downstream tasks of high clinical relevance. Finally, although SurgVISTA achieves superior performance across diverse benchmark tasks, thorough clinical validation is essential to determine its practical utility and impact in real-world surgical scenarios. Effective deployment of SurgVISTA-driven systems necessitates both the development of efficient and task-aligned downstream functionalities and close collaboration with clinical experts to ensure seamless integration into surgical workflows and alignment with clinical standards and practical requirements.

4 Method

Architecture Design

SurgVISTA is the first video-level self-supervision paradigm specifically targeted to surgical video analysis. The core innovation is a dual-decoder design designed for simultaneous holistic video understanding and fine-detail preservation. As illustrated in Fig. 1b, a video-level reconstruction branch enables the learning of motion dynamics and global contextual features while an image-level distillation branch ensures the retention of fine-grained spatial semantics and structural details. Specifically, given a video clip $X_v \in \mathbb{R}^{T \times H \times W \times 3}$, we employ a joint space-time cube embedding strategy²⁹ with cube size $P_T \times P_H \times P_W$ to partition the clip into multiple non-overlapping patches arranged in spatiotemporal order. Each patch is mapped to a visual token, yielding spatiotemporal representations $F_v \in \mathbb{R}^{\tilde{T} \times K \times C}$, where $\tilde{T} = \frac{T}{P_T}$ and $K = \frac{H \times W}{P_H \times P_W}$ quantify the numbers of temporal steps and spatial tokens, respectively. To enhance the effectiveness of masked video modeling and reduce temporal redundancy, we apply tube masking with a masking ratio to minimize information leakage while facilitating spatiotemporal representation learning. The encoder employs the vanilla ViT-B²⁸ architecture with embedding dimension $C = 768$, augmented with a joint spatiotemporal attention mechanism to effectively capture inter-frame dynamics and intra-frame spatial structures. Following the encoder, video-level reconstruction focuses on restoring low-level pixel information throughout the entire video clip, thereby facilitating the learning of holistic spatiotemporal representations. The enhanced visible tokens from the encoder are concatenated with learnable mask tokens and then fed into a lightweight spatiotemporal decoder, which infers the missing pixel-level details with structural and temporal consistency from partially visible inputs. The reconstruction loss is defined as the mean squared error (MSE) between the normalized ground-truth pixel values and their reconstructed counterparts, computed exclusively within the masked regions. By reconstructing masked portions from visible token representations, SurgVISTA is guided to attend to the temporal dynamics within surgical contexts, thereby facilitating the extraction of semantically meaningful spatiotemporal features. Furthermore, while the cube embedding strategy enables efficient temporal down-sampling and mitigates the computational overhead of video-level pre-training, it inevitably compromises fine-grained spatial details essential for comprehensive spatiotemporal interpretation. To address this limitation, image-level knowledge distillation is employed to optimize high-level spatial representations by leveraging supervision from a surgery-specific expert model¹⁸. Specifically, the encoder's enhanced visible tokens are concatenated with learnable mask tokens and fed into a lightweight spatiotemporal decoder to generate enhanced spatial representations, which are subsequently aligned with the frame-level spatial features extracted from the expert model. The distillation loss is defined as the Smooth L1 loss between the predicted features and the expert-derived spatial representations.

Data Splitting Strategy

Data leakage poses a significant challenge in the benchmarking of foundation models³³. To build a comprehensive and diverse corpus for both development and evaluation, we systematically collected a wide range of publicly available surgical video datasets. To avoid any contamination between the pre-training and evaluation phases, we strictly followed the principles below. First, no videos used during the pre-training phase were included in the test splits of any downstream datasets. For several datasets utilized in both pre-training and evaluation, all test samples were explicitly excluded from the pre-training data to prevent data leakage. Second, we conducted in-domain evaluations using two types of datasets: (1) the held-out test splits of benchmark datasets partially used in pre-training, and (2) additional datasets involving surgical procedures that are consistent with those seen during pre-training, but whose video samples were not included in the pre-training phase. These evaluations assessed the model's ability to generalize within familiar clinical domains while avoiding potential data leakage. Third, to assess generalization under distribution shifts, we conducted out-of-domain evaluations using datasets whose surgical procedures

were entirely absent from the pre-training corpus. Overall, these precautions ensured that no test split from any downstream evaluation contained videos seen during pre-training, allowing both in-domain and out-of-domain generalization to be assessed under rigorously controlled conditions.

Implementation Details

For **pre-training**, we initialized SurgVISTA entirely from scratch, employing random initialization for all components. The model employed the ViT-B encoder configured with fixed spatiotemporal positional embeddings and without a class token, while tokenization relied on a spatiotemporal cube embedding^{29,34} sized $2 \times 16 \times 16$. The decoder designed for video-level reconstruction comprised four conventional transformer layers, while the decoder dedicated to image-level knowledge distillation employed a two-layer transformer architecture. The total number of parameters was 100.8 M, with the encoder accounting for 87.4 M. The pre-training dataset consisted of 3,650 videos, totaling approximately 3.55 million frames, spanning diverse surgical procedures across distinct anatomical structures. For each video, we first sampled frames at 1 frame per second (fps), and subsequently constructed input video clips by uniformly sampling 16 frames at a fixed interval of 4, starting sequentially from each frame. We selected the key hyperparameters following a principled strategy: the patch size and input sequence length were determined by combining established designs from video transformers^{29,31,34–36} with the spatiotemporal characteristics of surgical videos^{37,38}. The masking ratio was calibrated for the surgical domain via preliminary ablation experiments within our framework and was set to 0.85 to balance performance and efficiency. For each frame, we adopted an input resolution of 224×224 pixels, with height $H = 224$ and width $W = 224$, yielding $K = 196$ spatial tokens per frame. The model was pre-trained for 200 epochs with a batch size of 512, employing gradient accumulation over four iterations. We employed the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, an initial learning rate of $1.5e-4$, and a weight decay of 0.05. These hyperparameters were directly adopted from prior work²⁹ without additional tuning, except for the number of epochs, which was proportionally adjusted according to the size of the surgical video dataset to ensure sufficient convergence. To balance the dual objectives of spatiotemporal modeling and spatial representation refinement, the video-level reconstruction loss was assigned a weight of 1.0, while the image-level knowledge distillation loss was assigned a weight of 0.05. The pre-training of SurgVISTA was performed on 8 NVIDIA H800 GPUs for 10 days. Following standard practice in self-supervised learning^{29,31}, no downstream validation data were used to guide the pre-training process. The final model checkpoint was selected at the end of training, at which point the model had already converged. For **fine-tuning**, we retained the architectural configurations of each pre-trained method to ensure faithful reproduction and fair comparison. For natural-domain methods, the encoder was initialized with the corresponding pre-trained parameters, and a single-layer fully connected (FC) head was appended for prediction, with all task-specific layers initialized randomly. The primary distinction between image-level and video-level methods lay in the embedding mechanism: image-level models employed standard 2D convolutional embeddings, while video-level models adopted cube embeddings via 3D convolutions. All training configurations were kept consistent across SurgVISTA and natural-domain methods on each dataset, except for the input frame length, which was set to 16 frames for video-level methods and 8 frames for image-level ones, to maintain comparable computational overhead. For comparisons with surgical-domain methods, SurgVISTA retained its end-to-end framework, consisting of an encoder followed by a single-layer FC head. In contrast, other surgical-domain methods were implemented following the conventional two-stage paradigm²⁰, where the encoder was first fine-tuned using frame-level supervision, followed by a temporal convolutional module that aggregated the extracted spatial features for final prediction. To ensure training consistency and fair comparison, we followed the official training configurations provided in the original implementation²⁰. For rigorous **downstream evaluation**, we adopted a strict data partitioning protocol as summarized in Supplementary Table 2. For all in-domain downstream datasets, dedicated test sets were explicitly excluded from both the pre-training and fine-tuning phases. For out-of-domain evaluations, the entire dataset, including both training and testing splits, was completely disjoint from the pre-training data. This protocol ensured a fair assessment of both domain-specific performance and cross-domain generalization.

SurgWeb Dataset

SurgWeb is constructed from publicly accessible online platforms that host endoscopic surgical videos for public viewing. The dataset spans more than twenty surgical procedures across multiple surgical specialties, including hepatobiliary surgery (e.g., cholecystectomy), colorectal surgery (e.g., colectomy), upper gastrointestinal surgery (e.g., gastrectomy), gynecologic laparoscopy, and urologic laparoscopy. All videos included in SurgWeb are used solely for non-commercial academic research under the publicly offered viewing and access terms of the hosting platforms. To ensure privacy protection, we implement a strict de-identification pipeline before incorporating any video into SurgWeb. Only intraoperative endoscopic views are retained. Videos containing external camera views, patient faces, audio tracks, embedded textual identifiers, or any other potentially identifying content are removed. The screening process combines automated filtering rules with manual review to ensure that all remaining clips are fully de-identified. After screening process, the dataset consists of 3,376 videos with 2,349,618 frames, and contains exclusively intraoperative surgical video content without any identifiable patient information.

Evaluation Benchmark

To comprehensively assess the performance of SurgVISTA, we curated a video-level evaluation benchmark encompassing six distinct surgical procedures across five anatomical structures, including laparoscopic cholecystectomy, laparoscopic hysterectomy (gynecologic laparoscopic surgery), laparoscopic liver resection, endoscopic pituitary surgery, ophthalmic surgery, and endoscopic submucosal dissection. This benchmark contained 13 datasets, comprising more than 500 long-form videos and 600 short-form clips, totaling over 660,000 frames. Detailed statistical information of these downstream datasets was provided in Supplementary Table 3. To prevent potential data leakage, we strictly ensured that the test sets from in-domain downstream datasets overlapping with the SurgPub were entirely excluded from both the pre-training and fine-tuning phases. Furthermore, we conducted experiments across four clinically significant tasks, including surgical workflow recognition, surgical action recognition, surgical triplet recognition, and surgical skill assessment. Effective performance on these tasks is essential for comprehensive surgical scene understanding, facilitating more precise comprehension of complex and dynamic surgical scenarios.

Surgical workflow recognition serves as a fundamental component of surgical scene understanding, which aims to automatically assign predefined surgical phases to individual frames within surgical video. In clinical practice, surgical procedures are systematically partitioned into well-defined phases based on expert knowledge, with each phase representing a specific objective or set of actions performed during the operation. Precise phase recognition is crucial for understanding procedural context and tracking intraoperative progress, thereby enabling real-time surgical assistance, procedural monitoring, and intraoperative decision support. To comprehensively evaluate the effectiveness of our approach on surgical workflow recognition, we conducted extensive experiments across seven public datasets: Cholec80²¹, M2CAI16-Workflow^{21,22}, PmLR50³⁸, AutoLaparo²⁶, CATARACTS³⁹, Cataract-101⁴⁰, and Cataract-21⁴¹, as well as one private dataset, ESD57. These datasets collectively covered a wide range of surgical procedures and anatomical structures, enabling a comprehensive assessment of the model's generalizability across diverse clinical scenarios. Following standard practice in surgical workflow recognition^{21,22,37,42,43}, performance was evaluated across multiple granularities using five widely adopted metrics: image-level accuracy, video-level accuracy, phase-level precision, phase-level recall, and phase-level Jaccard. These metrics capture performance at three levels of granularity: image-level accuracy treats each frame independently to assess global frame-wise correctness; video-level accuracy measures average performance across individual videos, reflecting video-wise correctness; and phase-level metrics assess category-wise performance across the dataset, reflecting the model's ability to distinguish and localize surgical phases.

Surgical action recognition seeks to identify intraoperative actions at the frame or clip level, facilitating detailed analysis of surgical activities. In contrast to coarse-grained workflow recognition, it focuses on the temporal and semantic interpretation of surgical maneuvers, such as dissecting, retracting, or suturing, which are critical determinants of patient safety and postoperative outcomes. In clinical practice, accurate action recognition plays a pivotal role in enhancing intraoperative situational awareness, enabling context-specific assistance such as real-time procedural guidance, anticipatory error prevention, and adaptive decision-making. By providing fine-grained semantic interpretation of intraoperative activities, this task lays the foundation for improving surgical precision, reducing complication rates, and ensuring procedural safety. To evaluate our approach on surgical action recognition, we conducted experiments on the SurgicalActions160⁴⁴ dataset, which encompasses a diverse set of intraoperative actions involving various surgical instruments and anatomical targets, thereby enabling a comprehensive assessment of model performance across varied surgical activity patterns. For surgical action recognition, performance was evaluated using accuracy, which measures overall classification correctness, and mean Average Precision (mAP), which offers a class-aware assessment by integrating precision-recall curves across all action categories.

Surgical triplet recognition focuses on fine-grained activity analysis by modeling detailed interactions among surgical instruments, actions (verbs), and anatomical targets. This task involves recognizing individual components and systematically modeling their interrelationships to capture intricate instrument-tissue interactions, thereby enhancing real-time surgical assistance, procedural monitoring, and informed intraoperative decision support. To assess the effectiveness of our approach on surgical triplet recognition, we conducted experiments on one public dataset, CholecT50⁴⁵, and one private dataset, Prostate21. Both datasets contained annotated surgical videos with detailed triplet labels comprising $\langle \text{instrument}, \text{action}, \text{target} \rangle$, which served as benchmarks for evaluating the model's ability to understand instrument-tissue interactions. Consistent with prior work in surgical triplet recognition^{45,46}, we reported component average precision and triplet average precision. Component average precision evaluates the individual elements of the triplet, including instrument (AP_I), verb (AP_V), and target (AP_T), based on the area under the precision-recall curve for each class. Triplet average precision measures the model's ability to recognize interactions by considering combinations such as instrument-verb (AP_{IV}), instrument-target (AP_{IT}), and instrument-verb-target (AP_{IVT}). The primary metric in this study is AP_{IVT} , which captures the accuracy of full triplet recognition.

Surgical skill assessment aims to evaluate specific actions or activities at the frame or clip level, requiring a comprehensive understanding of complex anatomical structures and spatiotemporal relationships. Accurate skill assessment is crucial for reducing the risk of injury to critical anatomical structures and promoting procedural standardization, thereby facilitating intraoperative monitoring and improving postoperative evaluation. To comprehensively evaluate the effectiveness of our

approach in surgical skill assessment, we conducted evaluations on the Cholec80-CVS⁴⁷ and Endoscapes-CVS⁴⁸ datasets. Both datasets focused on assessing Strasberg's Critical View of Safety (CVS) criteria, serving as a benchmark for evaluating surgical proficiency and promoting procedural safety. For surgical skill assessment, we adopted accuracy and mean Average Precision (mAP) as evaluation metrics, where accuracy reflects overall prediction correctness, and mAP provides a class-wise evaluation of recognition performance.

Experimental settings

State-of-the-art pre-trained methods in the natural domain. To enable a comprehensive and structured comparison, we evaluate eight representative natural-domain pre-trained methods, categorized into two groups: **image-based pre-training** and **video-based pre-training**. All models are built upon the ViT-B architecture²⁸ and are pre-trained on large-scale general-purpose datasets, without any additional fine-tuning on task-specific or dataset-specific natural-domain benchmarks. Detailed configurations are provided in Supplementary Table 34.

Image-based Pre-trained Methods. These approaches exclusively focus on learning spatial representations and are pre-trained using large-scale static image datasets.

- **ImageNet-1K⁴⁹ (Supervised²⁸)** refers to the pre-trained parameters obtained through fully supervised training on the ImageNet-1K dataset, which contains over 1.2 million labeled images across 1,000 object categories.
- **ImageNet-1K⁴⁹ (DINO⁸)** refers to the pre-trained parameters obtained by applying the DINO framework to the ImageNet-1K dataset, which learns semantic feature representations through self-distilled contrastive learning with momentum encoders and clustering.
- **ImageNet-1K⁴⁹ (MAE⁹)** refers to the pre-trained parameters derived by applying the MAE framework to the ImageNet-1K dataset, which learns high-quality spatial representations by reconstructing pixel-level information from randomly masked image patches.

Video-based Pre-trained Methods. These methods are trained directly on large-scale video data to learn joint spatiotemporal representations by capturing both dynamic temporal patterns and spatial semantics.

- **Kinetics-400⁵⁰ (VideoMAE²⁹)** refers to the pre-trained parameters obtained by applying the VideoMAE framework to the Kinetics-400 dataset, a large-scale human action dataset containing approximately 240,000 video clips across 400 action classes. VideoMAE leverages a masked autoencoder architecture tailored for video inputs, enabling the model to learn robust spatiotemporal representations by reconstructing masked video patches.
- **UnlabeledHybrid³⁰ (VideoMAE²⁹)** refers to the pre-trained parameters obtained by applying the VideoMAE framework to the UnlabeledHybrid dataset, which aggregates over 1.37 million video clips from Kinetics-710⁵¹, Something-Something V2⁵², AVA⁵³, WebVid2M⁵⁴, and other online sources.
- **Something-Something V2⁵² (VideoMAE²⁹)** refers to the pre-trained parameters obtained by applying the VideoMAE framework to the Something-Something V2 (SSV2) dataset, which contains approximately 169,000 video clips across 174 action categories with object-centric temporal interactions.
- **Kinetics-710⁵¹ (UMT³²)** refers to the pre-trained parameters obtained by applying the UMT framework to the Kinetics-710 dataset, which merges Kinetics-400⁵⁰, Kinetics-600⁵⁵, and Kinetics-700⁵⁶ while removing duplicates. Specifically, we adopt parameters from the first training stage of UMT, where masked video modeling is guided by CLIP-based supervision⁵⁷ in a single-modality setting.
- **Kinetics-400⁵⁰ (MVD³¹)** refers to the pre-trained parameters obtained by applying the MVD framework to the Kinetics-400 dataset. MVD employs masked feature modeling combined with a teacher-student distillation framework, utilizing both image-level and video-level teacher features to effectively guide the learning of robust and coherent spatiotemporal representations.

State-of-the-art pre-trained methods in the surgical domain. In this study, we select four state-of-the-art self-supervised surgical models as surgical foundation models to facilitate a comprehensive and complementary comparison. Detailed information is provided in Supplementary Table 35.

- **SelfSupSurg¹⁶** explores the application of four advanced image-level SSL methods, namely MoCo v2⁵⁸, SimCLR⁵⁹, DINO⁸, and SwAV⁶⁰, for surgical scene comprehension. For comparison, we adopt the DINO-based variant, which employs a ResNet-50⁶¹ backbone pre-trained on the Cholec80²¹ dataset containing a limited amount of surgical data. The study provides a performance evaluation across six datasets representing two distinct surgical procedures.

- **EndoViT¹⁷** compiles a large-scale publicly available image-level endoscopic dataset, Endo700k, sourced from nine public datasets and consisting of over 700,000 images. It adopts the MAE⁹ framework with a ViT-B backbone for domain-specific pre-training on the Endo700k dataset. The study evaluates model performance on three downstream datasets, all corresponding to the same surgical procedure.
- **EndoSSL¹⁸** constructs an extensive private dataset consisting of 7,877 laparoscopic procedure videos, encompassing 23.3 million frames and spanning over five distinct surgical procedures. It employs the MSN⁶² framework with a ViT-L backbone pre-trained on this private dataset, resulting in a strong image-level surgical foundation model. The study assesses model performance exclusively on the Cholec80 dataset, focusing on surgical workflow recognition task.
- **GSViT¹⁹** develops an efficient model that employs EfficientNet⁶³ as the backbone and utilizes next-frame prediction as the pretext task for pre-training, providing weak temporal modeling capabilities. To facilitate this, it open-sources a large-scale surgical video dataset comprising over 680 hours of surgical footage sourced from websites, totaling more than 2.5 million frames and spanning 28 distinct surgical procedures. The evaluation of model performance is conducted solely on the Cholec80 dataset, with a specific focus on surgical workflow recognition task.

Pre-training Datasets. In this study, we curate eight publicly available video-level surgical datasets, referred to as SurgPub, along with various surgical videos sourced from online platforms, designated as SurgWeb. Detailed information is provided in Supplementary Table 2.

- **Cholec80²¹** is a widely used laparoscopic surgery dataset focusing on the cholecystectomy procedure, comprising 80 videos in total. The dataset is officially divided into training and test sets, each consisting of 40 videos. For pre-training, we exclusively utilize the 40 videos from the training set, amounting to a total of 86,344 frames.
- **MI2CAI16-Workflow^{21,22}** is a dataset curated for surgical workflow recognition, containing 41 laparoscopic videos of cholecystectomy procedures. The dataset is split into 27 videos for training and 14 videos for testing. We incorporate the 27 videos from the training split into the pre-training dataset, contributing a total of 67,578 frames.
- **HeiChole²³** is an open benchmark for laparoscopic cholecystectomy, focusing on surgical workflow and skill analysis. The dataset comprises 24 videos allocated for training, while the remaining 9 videos are designated for testing. For pre-training, we utilize the publicly available 24 videos from the training set, providing 55,139 frames in total.
- **PitVis²⁴** is a dataset curated for step and instrument recognition in endoscopic pituitary surgery videos. It contains 25 publicly accessible videos comprising 120,018 frames. The entire dataset is incorporated into the pre-training phase to enhance the procedural diversity of the pre-training data.
- **PSI-AVA²⁵** is a dataset focused on robot-assisted radical prostatectomy procedures, designed to enhance holistic surgical scene understanding. It encompasses approximately 20.45 hours of surgical footage, comprising 8 videos with a total of 72,318 frames. The entire dataset is incorporated into the pre-training process.
- **AutoLaparo²⁶** is a widely used laparoscopic surgery dataset focusing on the hysterectomy procedure. It contains 21 videos, with 10 videos allocated for training and the remaining 11 designated for testing. We incorporate the 10 videos from the training set into the pre-training dataset, contributing a total of 40,211 frames.
- **BernBypass70²⁷** consists of 70 laparoscopic Roux-en-Y gastric bypass videos, encompassing 303,764 frames. This dataset is fully incorporated into the pre-training stage to increase the procedural and anatomical diversity.
- **StrasBypass70²⁷** is a dataset dedicated to laparoscopic Roux-en-Y gastric bypass procedures, containing 70 surgical videos and a total of 457,787 frames. The entire dataset, comprising all available videos and frames, is used for pre-training to expand the diversity of surgical contexts.
- **SurgWeb** consists of 3,376 videos with 2,349,618 frames, covering more than 20 distinct surgical procedures and related anatomical structures. The entire dataset is incorporated into the pre-training phase to enhance data diversity and strengthen the model's spatiotemporal modeling capability.

Benchmark Datasets. In this study, we define datasets involving surgical procedures included in the pre-training data as in-domain, while those comprising previously unseen surgical procedures are referred to as out-of-domain. All evaluation datasets are sampled at 1 frame per second (1 FPS). Detailed information is provided in Supplementary Table 3.

The in-domain datasets include the following:

- **Cholec80**²¹ is a benchmark dataset designed for surgical workflow recognition in laparoscopic cholecystectomy. Each frame is annotated with one of seven surgical workflow categories. Following the official data split protocol, our pre-training stage includes videos from both the Cholec80 and M2CAI16-Workflow training sets. However, we identify a possible data leakage issue due to five overlapping videos between M2CAI16-Workflow and the Cholec80 test set. To mitigate any risk of data leakage and ensure fair evaluation, we conservatively excluded five potentially overlapping videos (Videos 73, 77, 78, 79, and 80) from the Cholec80 test set, as these may have been included in the M2CAI16-Workflow training set. After excluding the five overlapping videos, we use 40 videos containing 86,344 frames for training, and the remaining 35 videos with 88,494 frames for testing.
- **Cholec80-CVS**⁴⁷ is a dataset for surgical skill assessment, providing video-level annotations of Strasberg's Critical View of Safety (CVS) criteria for Cholec80. Each clip is labeled with three scores corresponding to the CVS criteria, with each score ranging from 0 to 2, resulting in a total score ranging from 0 to 6. The dataset is partitioned into training, validation and testing sets, while manually excluding samples labeled with a score of 6 that only appear in the test set. The training set consists of 263 clips totaling 17,201 frames, the validation set contains 43 clips totaling 3,327 frames, and the test set comprises 137 clips totaling 10,469 frames.
- **CholecT50**⁴⁵ is a dataset designed for surgical triplet recognition in laparoscopic cholecystectomy. The dataset comprises 6 instruments, 10 verbs, and 15 targets, yielding a total of 100 triplet classes in the format of (instrument, verb, target). Following the common partition^{45,64}, 45 videos containing 90,489 frames are designated for training, while the remaining 5 videos with 10,374 frames are reserved for testing.
- **M2CAI16-Workflow**^{21,22} is a benchmark dataset used for surgical workflow recognition in laparoscopic cholecystectomy. Each frame is labeled with one of eight distinct surgical workflow categories. Consistent with the standard partition, 27 videos containing 67,578 frames are allocated for training, and the remaining 14 videos comprising 26,961 frames are set aside for testing.
- **Endoscapes-CVS**⁴⁸ is a dataset designed for surgical skill assessment in laparoscopic cholecystectomy. It comprises 201 clips, with each frame annotated using three averaged expert-assigned scores corresponding to CVS criteria. Each score is thresholded into a binary value (0 or 1), resulting in a total discrete score ranging from 0 to 3. A total of 120 videos, comprising 36,694 frames, are allocated for training, 41 videos containing 12,372 frames are used for validation, and the remaining 40 videos, with 9,747 frames, are designated for testing.
- **PmLR50**³⁸ is a dataset for surgical workflow recognition in laparoscopic liver resection, containing 50 videos with totally 23,037 frames. Each frame is annotated with one of five distinct surgical workflow categories. Following the official partition, 35 videos totaling 17,378 frames are used for training, 5 videos with 2,309 frames are designated for validation, and 10 videos with 5,350 frames are reserved for testing.
- **AutoLaparo**²⁶ is a dataset for image-guided surgical automation in laparoscopic hysterectomy. It contains 21 videos of hysterectomy procedures, comprising 83,243 frames. Each frame is annotated with one of seven distinct surgical workflow categories. Following the official partition, 10 videos totaling 40,211 frames are used for training, 4 videos with 14,972 frames are designated for validation, and 7 videos with 28,060 frames are reserved for testing.
- **SurgicalActions160**⁴⁴ is a dataset containing 160 short-form clips, each depicting typical surgical actions in gynecologic laparoscopy. It consists of 16 distinct actions, with each action represented by exactly 10 examples. A total of 112 videos, comprising 13,238 frames, are allocated for training, while 48 videos with 5,783 frames are designated for testing.
- **Prostate21** is a dataset consisting of 12 publicly available videos obtained from^{25,65} and 9 privately collected recordings, all capturing radical prostatectomy procedures. Based on manual annotations, the dataset includes 7 instruments, 10 verbs, and 10 targets, resulting in 89 unique triplet classes structured as (instrument, verb, target). To prevent potential data leakage, all videos originating from the PSI-AVA dataset²⁵ are allocated exclusively to the training set. In total, the training set consists of 14 videos with 38,730 frames, the validation set includes 3 videos with 9,571 frames, and the test set comprises 4 videos with 12,228 frames.

The out-of-domain datasets include the following:

- **CATARACTS**³⁹ is a dataset designed for surgical workflow recognition in cataract surgery, consisting of 50 videos and 31,955 frames. In this dataset, the cataract surgery is divided into 19 phases and each frame is annotated with a phase label. Following the official partition, 25 videos totaling 16,507 frames are allocated for training, while 5 videos and 20 videos with 15,448 frames are designated for validation and testing.

- **Cataract-101**⁴⁰ is a dataset designed for surgical workflow recognition in cataract surgery, comprising 101 untrimmed videos, of which 100 are employed for downstream task. Each frame is assigned to one of ten distinct surgical phases. The dataset is partitioned into 50 videos with 25,288 frames for training, 10 videos with 4,148 frames for validation, and 40 videos with 18,215 frames for testing.
- **Cataract-21**⁴¹ is a dataset designed for surgical workflow recognition in cataract surgery. It consists of 21 video recordings, with the surgical procedure divided into 11 distinct phases, each frame annotated with the corresponding phase. The dataset includes 18 videos with 7,716 frames for training and 3 videos with 937 frames for testing.
- **ESD57** is a private dataset designed for surgical workflow recognition in endoscopic submucosal dissection (ESD) procedures, containing 57 video recordings spanning the stomach and lower digestive tract. Each frame is labeled with one of eight distinct workflow phases: Lesion Detection, Lesion Marking, Submucosal Injection, Mucosal Incision, Submucosal Dissection, Wound Management, Wound Suturing, and Others. A total of 39 videos with 121,225 frames are used for training, 6 videos totaling 24,187 frames for validation, and the remaining 12 videos totaling 38,934 frames are allocated for testing.

Data Availability

Publicly available datasets used to construct the pre-training corpus and evaluation benchmarks are summarized in Supplementary Table 37. The remaining clinical data cannot be shared publicly due to institutional and patient privacy restrictions.

Code Availability

The implementations of SurgVISTA framework will be released in GitHub: <https://github.com/isyangshu/SurgVISTA>. The pre-trained natural-domain parameters used in this study are listed in **Supplementary Table 34**, while the pre-trained surgical-domain parameters are listed in **Supplementary Table 35**. The other public codes used in this study are listed in **Supplementary Table 36**.

Acknowledgements

The work described in this paper was supported by a grant from the Germany/Hong Kong Joint Research Scheme sponsored by the Research Grants Council of Hong Kong and the Germany Academic Exchange Service of Germany (Reference No. G-HKUST605/24) and Hong Kong Innovation and Technology Commission (Project No. GHP/006/22GD and ITCPD/17-9).

Author Contributions

S.Y., L.M.-H., and H.C. conceived and designed the work. S.Y. contributed to the technical implementation and conducted experiments. F.Z. participated in discussions regarding the design of the self-supervised learning framework and were responsible for reproducing the natural-domain models. L.M. participated in discussions regarding the design of the self-supervised learning framework and contributed to part of the experimental evaluations. F.H., Y.W., S.H., Y.N., and Y.C. collected the data for self-supervised learning and downstream task evaluation. Xi.W., Y.J. and J.Q. offered insightful suggestions for the experimental design and thoughtfully directing the research trajectory. H.S., S.X., A.Q.L, Z.L., and J.Y.T. provided clinical expertise and facilitated access to proprietary datasets. All authors contributed to the drafting and revising of the manuscript. L.M.-H. and H.C. supervised the research.

Competing Interests

S.Y. and H.C. are inventors on a patent application related to this work that is currently being prepared for filing via the Patent Cooperation Treaty (PCT) route, with The Hong Kong University of Science and Technology as the applicant. The application will cover the pre-training framework, model architecture, and pre-trained parameters presented in this manuscript. All other authors declare no competing interests.

References

1. Kiyasseh, D. *et al.* A vision transformer for decoding surgeon activity from surgical videos. *Nat. biomedical engineering* **7**, 780–796 (2023).
2. Ma, R. *et al.* Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *NPJ Digit. Medicine* **5**, 187 (2022).
3. Kiyasseh, D. *et al.* Human visual explanations mitigate bias in ai-based assessment of surgeon skills. *NPJ Digit. Medicine* **6**, 54 (2023).

4. Demir, K. C. *et al.* Deep learning in surgical workflow analysis: a review of phase and step recognition. *IEEE J. Biomed. Heal. Informatics* **27**, 5405–5417 (2023).
5. Maier-Hein, L. *et al.* Surgical data science for next-generation interventions. *Nat. Biomed. Eng.* **1**, 691–696 (2017).
6. Surgical Data Science Initiative. Surgical data science. <http://www.surgical-data-science.org/> (2025). Accessed: April 12, 2025.
7. Maier-Hein, L. *et al.* Surgical data science—from concepts toward clinical translation. *Med. image analysis* **76**, 102306 (2022).
8. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660 (2021).
9. He, K. *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009 (2022).
10. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
11. Azizi, S. *et al.* Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat. Biomed. Eng.* **7**, 756–779 (2023).
12. Ma, J. *et al.* Towards a generalizable pathology foundation model via unified knowledge distillation. *arXiv preprint arXiv:2407.18449* (2024).
13. Huang, S.-C. *et al.* Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit. Medicine* **6**, 74 (2023).
14. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **6**, 1346–1352 (2022).
15. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Medicine* **30**, 850–862 (2024).
16. Ramesh, S. *et al.* Dissecting self-supervised learning methods for surgical computer vision. *Med. Image Analysis* **88**, 102844 (2023).
17. Batić, D., Holm, F., Özsoy, E., Czempel, T. & Navab, N. Endovit: pretraining vision transformers on a large collection of endoscopic images. *Int. J. Comput. Assist. Radiol. Surg.* **19**, 1085–1091 (2024).
18. Hirsch, R. *et al.* Self-supervised learning for endoscopic video analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 569–578 (Springer, 2023).
19. Schmidgall, S., Kim, J. W., Jopling, J. & Krieger, A. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949* (2024).
20. Czempel, T. *et al.* Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* **23**, 343–352 (Springer, 2020).
21. Twinanda, A. P. *et al.* Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* **36**, 86–97 (2016).
22. Stauder, R. *et al.* The tum lapchole dataset for the m2cai 2016 workflow challenge. *arXiv preprint arXiv:1610.09278* (2016).
23. Wagner, M. *et al.* Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Med. Image Analysis* **86**, 102770 (2023).
24. Das, A. *et al.* Pitvis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery. *arXiv preprint arXiv:2409.01184* (2024).
25. Valderrama, N. *et al.* Towards holistic surgical scene understanding. In *International conference on medical image computing and computer-assisted intervention*, 442–452 (Springer, 2022).
26. Wang, Z. *et al.* Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 486–496 (Springer, 2022).
27. Lavanchy, J. L. *et al.* Challenges in multi-centric generalization: Phase and step recognition in roux-en-y gastric bypass surgery. <https://arxiv.org/abs/2312.11250> (2023). ArXiv:2312.11250.

28. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
29. Tong, Z., Song, Y., Wang, J. & Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. neural information processing systems* **35**, 10078–10093 (2022).
30. Wang, Y. *et al.* Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191* (2022).
31. Wang, R. *et al.* Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6312–6322 (2023).
32. Li, K. *et al.* Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19948–19960 (2023).
33. White, C. *et al.* Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314* **4** (2024).
34. Wang, L. *et al.* Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14549–14560 (2023).
35. Bertasius, G., Wang, H. & Torresani, L. Is space-time attention all you need for video understanding? In *ICML*, vol. 2, 4 (2021).
36. Feichtenhofer, C., Li, Y., He, K. *et al.* Masked autoencoders as spatiotemporal learners. *Adv. neural information processing systems* **35**, 35946–35958 (2022).
37. Yang, S., Luo, L., Wang, Q. & Chen, H. Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 606–616 (Springer, 2024).
38. Guo, D., Si, W., Li, Z., Pei, J. & Heng, P.-A. Surgical workflow recognition and blocking effectiveness detection in laparoscopic liver resections with pringle maneuver. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)* (2025).
39. Al Hajj, H. *et al.* Cataracts: Challenge on automatic tool annotation for cataract surgery. *Med. image analysis* **52**, 24–41 (2019).
40. Schoeffmann, K. *et al.* Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM multimedia systems conference*, 421–425 (2018).
41. Primus, M. J. *et al.* Frame-based classification of operation phases in cataract surgery videos. In *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I* **24**, 241–253 (Springer, 2018).
42. Jin, Y. *et al.* Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Med. Imaging* **40**, 1911–1923 (2021).
43. Yang, S. *et al.* Surgpetl: Parameter-efficient image-to-surgical-video transfer learning for surgical phase recognition. *arXiv preprint arXiv:2409.20083* (2024).
44. Schoeffmann, K. *et al.* Video retrieval in laparoscopic video recordings with dynamic content descriptors. *Multim. Tools Appl.* **77**, 16813–16832, DOI: [10.1007/s11042-017-5252-2](https://doi.org/10.1007/s11042-017-5252-2) (2018).
45. Nwoye, C. I. *et al.* Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Med. Image Analysis* **78**, 102433 (2022).
46. Nwoye, C. I. *et al.* Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* **23**, 364–374 (Springer, 2020).
47. Ríos, M. S. *et al.* Cholec80-cvs: An open dataset with an evaluation of strasberg’s critical view of safety for ai. *Sci. Data* **10**, 194 (2023).
48. Murali, A. *et al.* Latent graph representations for critical view of safety assessment. *IEEE Transactions on Med. Imaging* (2023).
49. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).
50. Kay, W. *et al.* The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

51. Li, K. *et al.* Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552* (2022).
52. Goyal, R. *et al.* The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, 5842–5850 (2017).
53. Gu, C. *et al.* Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6047–6056 (2018).
54. Bain, M., Nagrani, A., Varol, G. & Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738 (2021).
55. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C. & Zisserman, A. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340* (2018).
56. Carreira, J., Noland, E., Hillier, C. & Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).
57. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
58. Chen, X., Fan, H., Girshick, R. & He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).
59. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).
60. Caron, M. *et al.* Unsupervised learning of visual features by contrasting cluster assignments. *Adv. neural information processing systems* **33**, 9912–9924 (2020).
61. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
62. Assran, M. *et al.* Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, 456–473 (Springer, 2022).
63. Liu, X. *et al.* Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14420–14430 (2023).
64. Nwoye, C. I. & Padoy, N. Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235* (2022).
65. Bawa, V. S. *et al.* The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. *arXiv preprint arXiv:2104.03178* (2021).

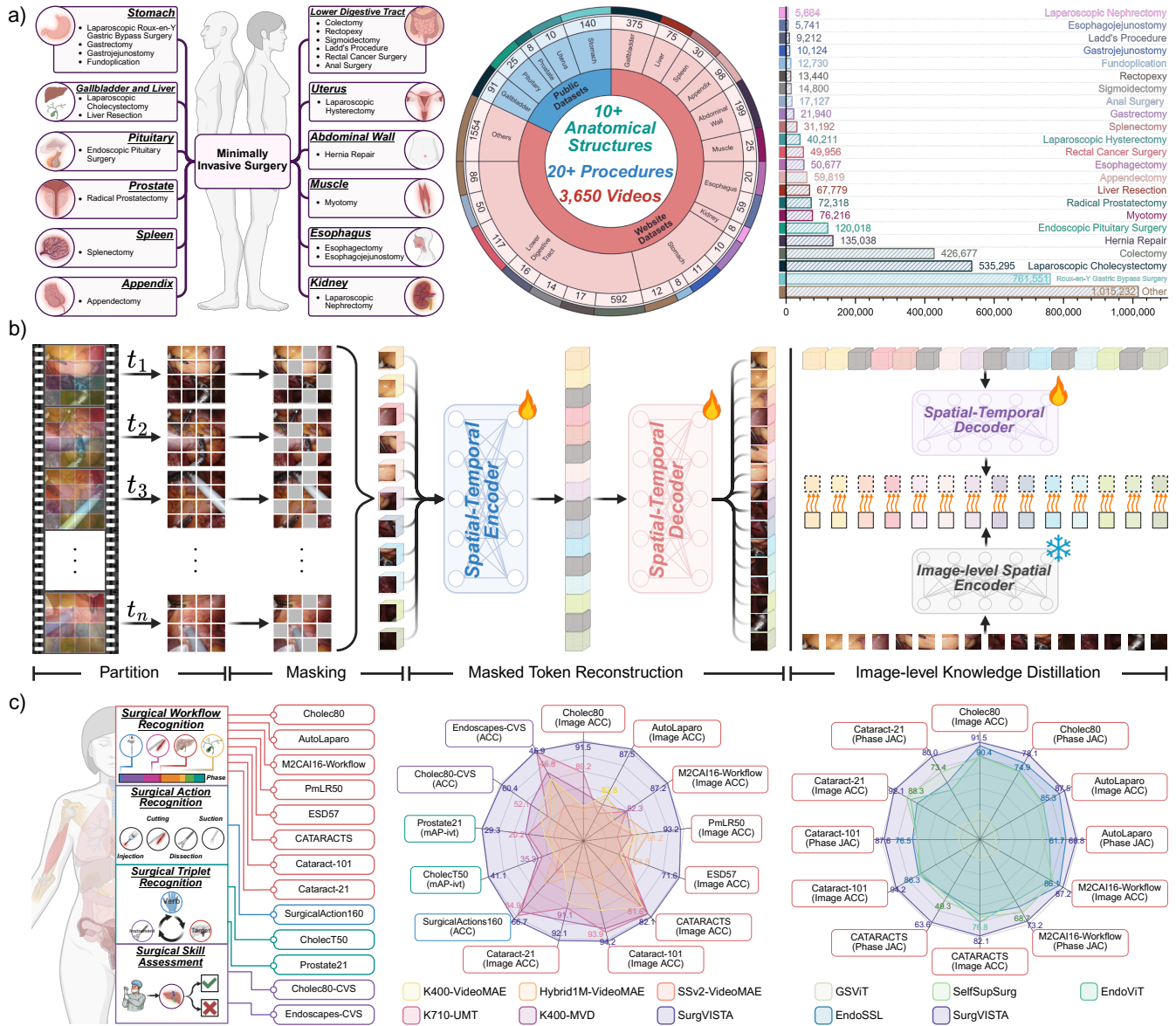


Figure 1. Overview of the study. (a) **Pre-training Dataset:** Illustration of the primary anatomical structures and their associated surgical procedures, complemented by video-level statistics and detailed frame-level distribution analysis. (b) **SurgVISTA Framework:** An asymmetric encoder-decoder architecture featuring a unified encoder for comprehensive spatiotemporal modeling, and two decoders: one dedicated to video-level reconstruction for capturing temporal dynamics and structural consistency, and another for image-level knowledge distillation guided by an expert model to enhance spatial representations. (c) **Downstream Evaluation:** Comprehensive visualization of the four downstream tasks along with 13 evaluated datasets, accompanied by comparative performance analyses of models pre-trained on natural-domain and surgical-domain data. Created with BioRender.com under an academic license.

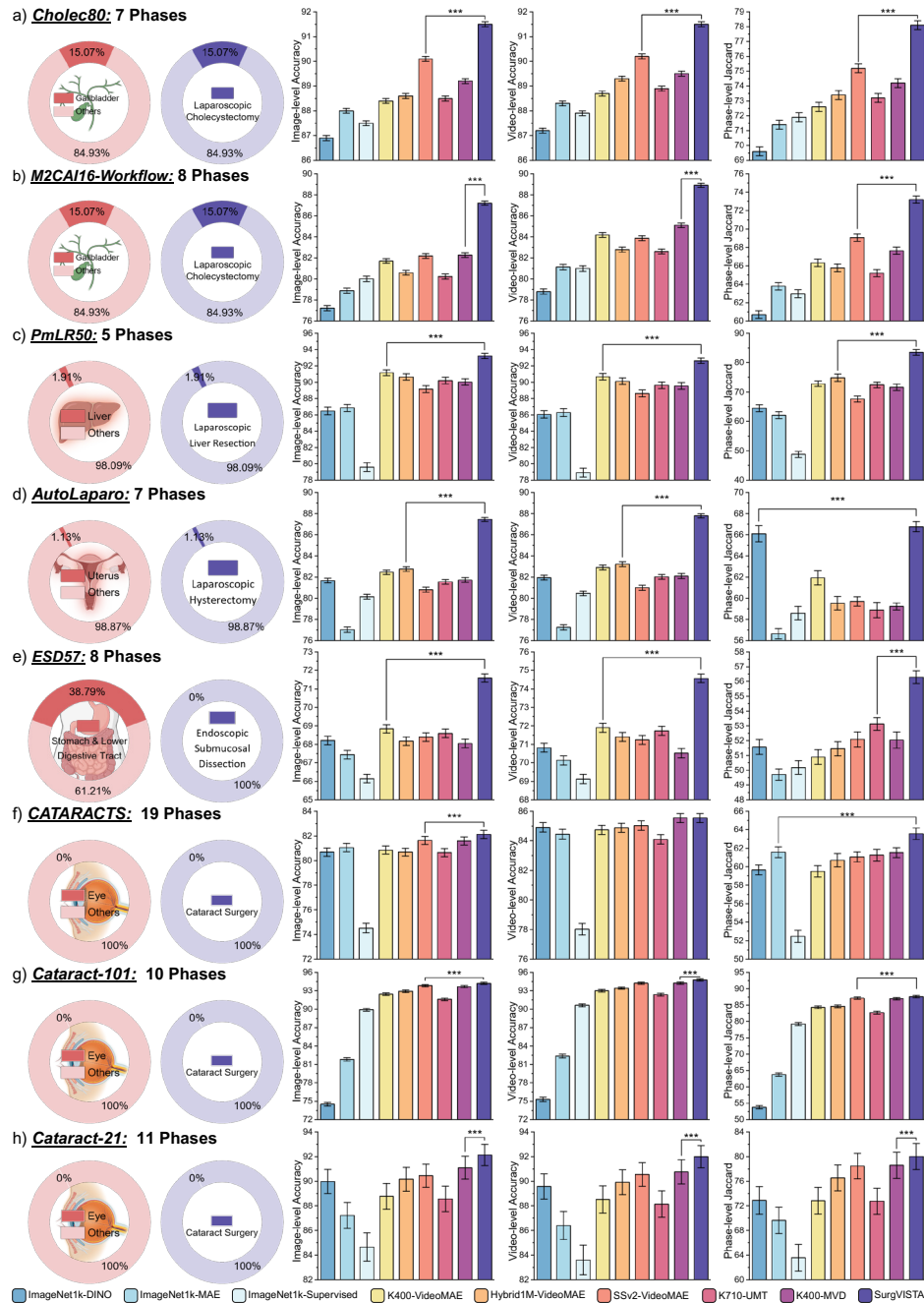


Figure 2. Experimental results of various natural-domain pre-trained methods and SurgVISTA on surgical workflow recognition datasets. Subfigures (a)-(d) correspond to in-domain datasets, while (e)-(h) represent out-of-domain datasets. To provide a comprehensive understanding of the model's generalization capacity, we quantify the proportion of pre-training data associated with relevant anatomical structures and surgical procedures. These proportions are visualized using donut charts, with anatomical overlap depicted in red and procedural overlap shown in blue, effectively illustrating the degree of alignment between the pre-training data and downstream datasets. Evaluation metrics include image-level accuracy, video-level accuracy and phase-level Jaccard, providing a comprehensive assessment of model performance across multiple granularities. Statistical significance (P-value) is reported whenever SurgVISTA demonstrates superior performance compared to other methods. Detailed results are presented in Supplementary Tables 4, 5, 6, 7, 8, 9, 10 and 11. Created with BioRender.com under an academic license.

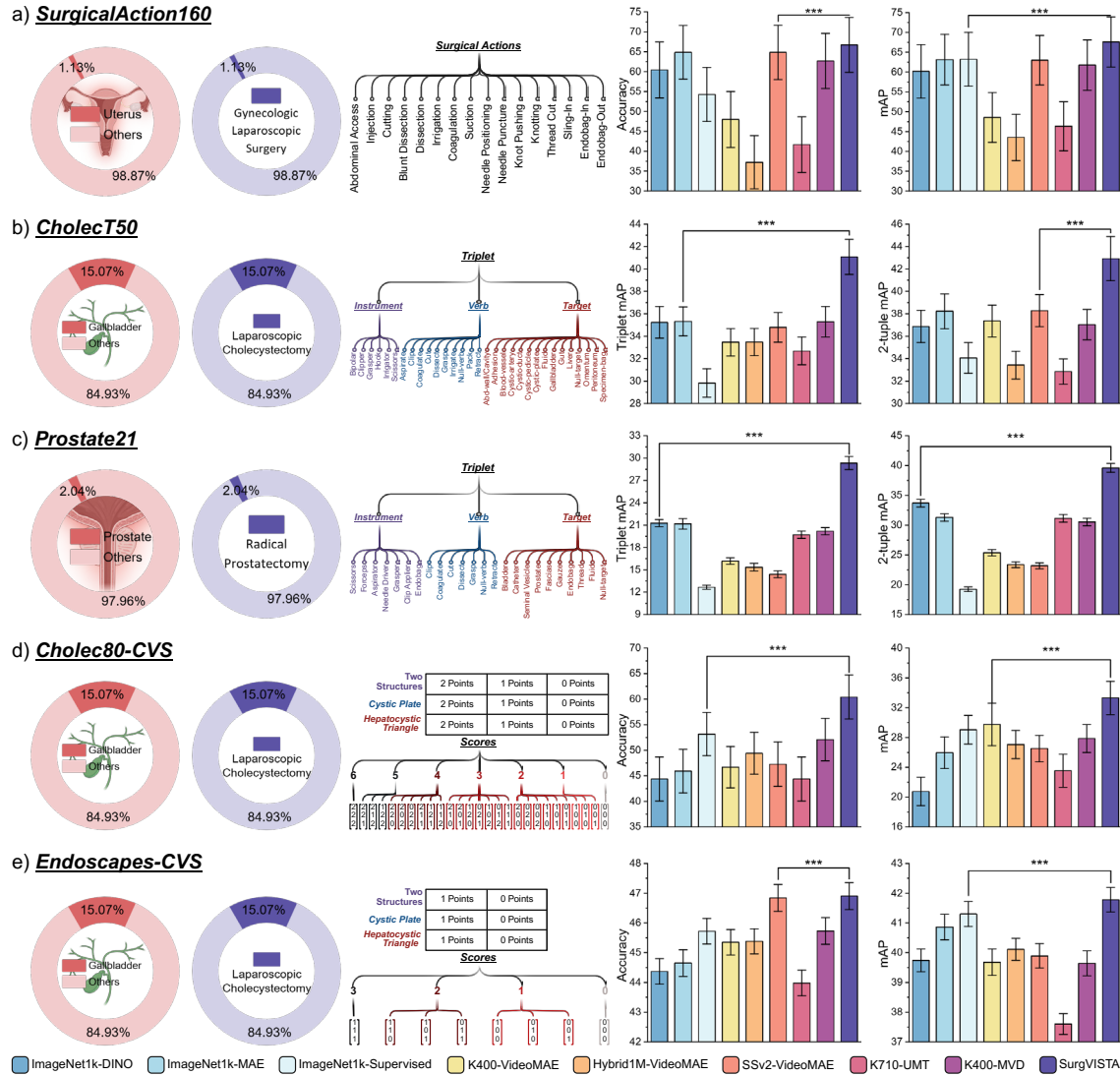


Figure 3. Experimental results of various natural-domain pre-trained methods and SurgVISTA on surgical tasks beyond phase recognition. Subfigure (a) presents the results for the surgical action recognition task, with evaluation metrics reported as accuracy and mAP. Subfigures (b)-(c) illustrate the surgical triplet recognition task, with performance reported in terms of triplet mAP and 2-tuple mAP. Subfigures (d)-(e) display the results for the surgical skill assessment task, evaluated in terms of accuracy and mAP. Statistical significance (P-value) is reported whenever SurgVISTA demonstrates superior performance compared to other methods. Detailed results are presented in Supplementary Tables 12, 13, and 14. Created with BioRender.com under an academic license.

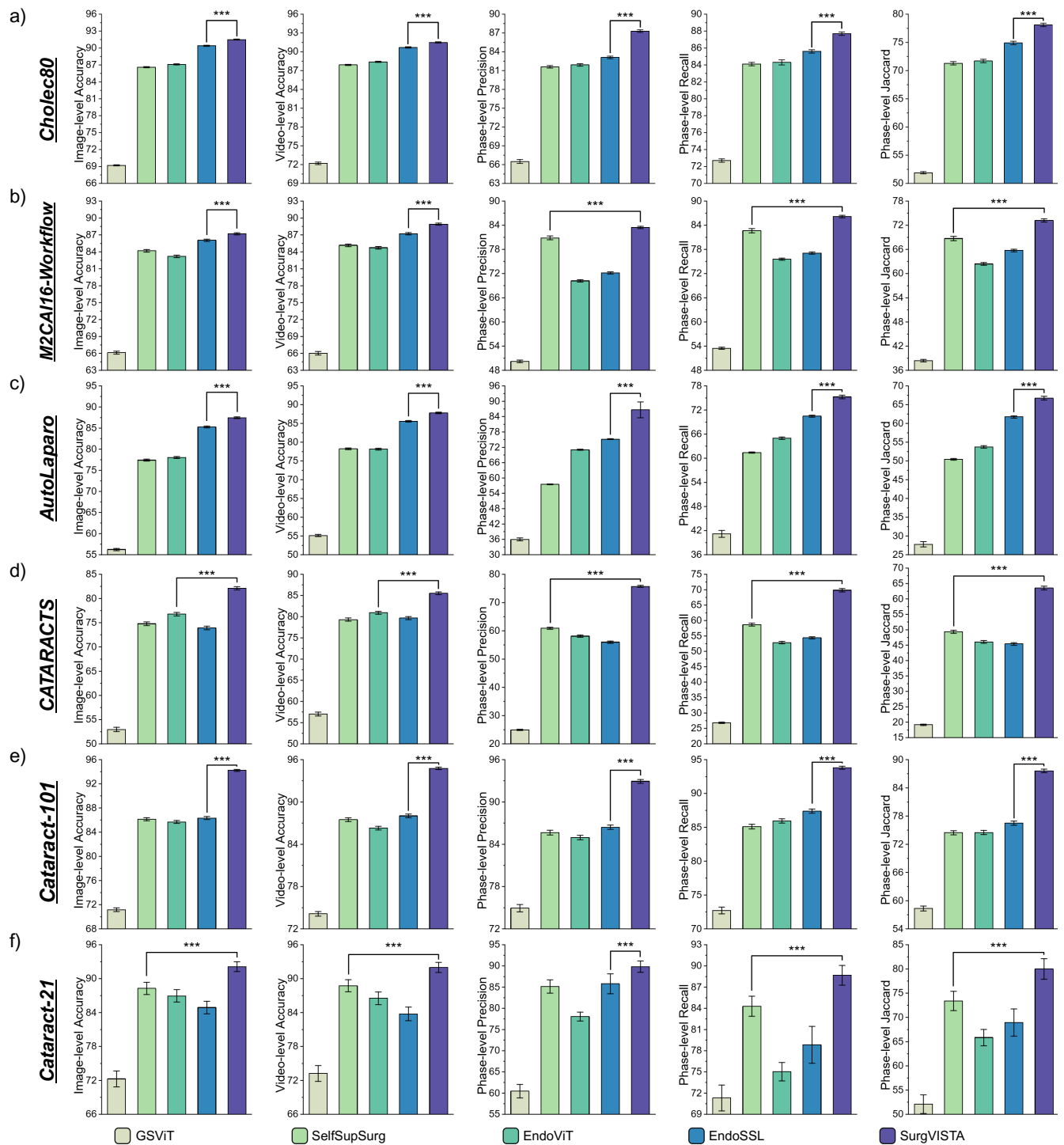


Figure 4. Experimental results comparing SurgVISTA and various surgical-domain pre-trained methods on surgical phase recognition datasets. Subfigures (a)-(c) correspond to in-domain datasets, while (d)-(f) represent out-of-domain datasets. Evaluation metrics include image-level accuracy, video-level accuracy, phase-level precision, phase-level recall and phase-level Jaccard, providing a comprehensive assessment across multiple granularities. Statistical significance (P-value) is reported whenever SurgVISTA demonstrates superior performance compared to competing methods. Detailed results are presented in Supplementary Tables 15, 16, 17, 18, 19 and 20. Created with BioRender.com under an academic license.

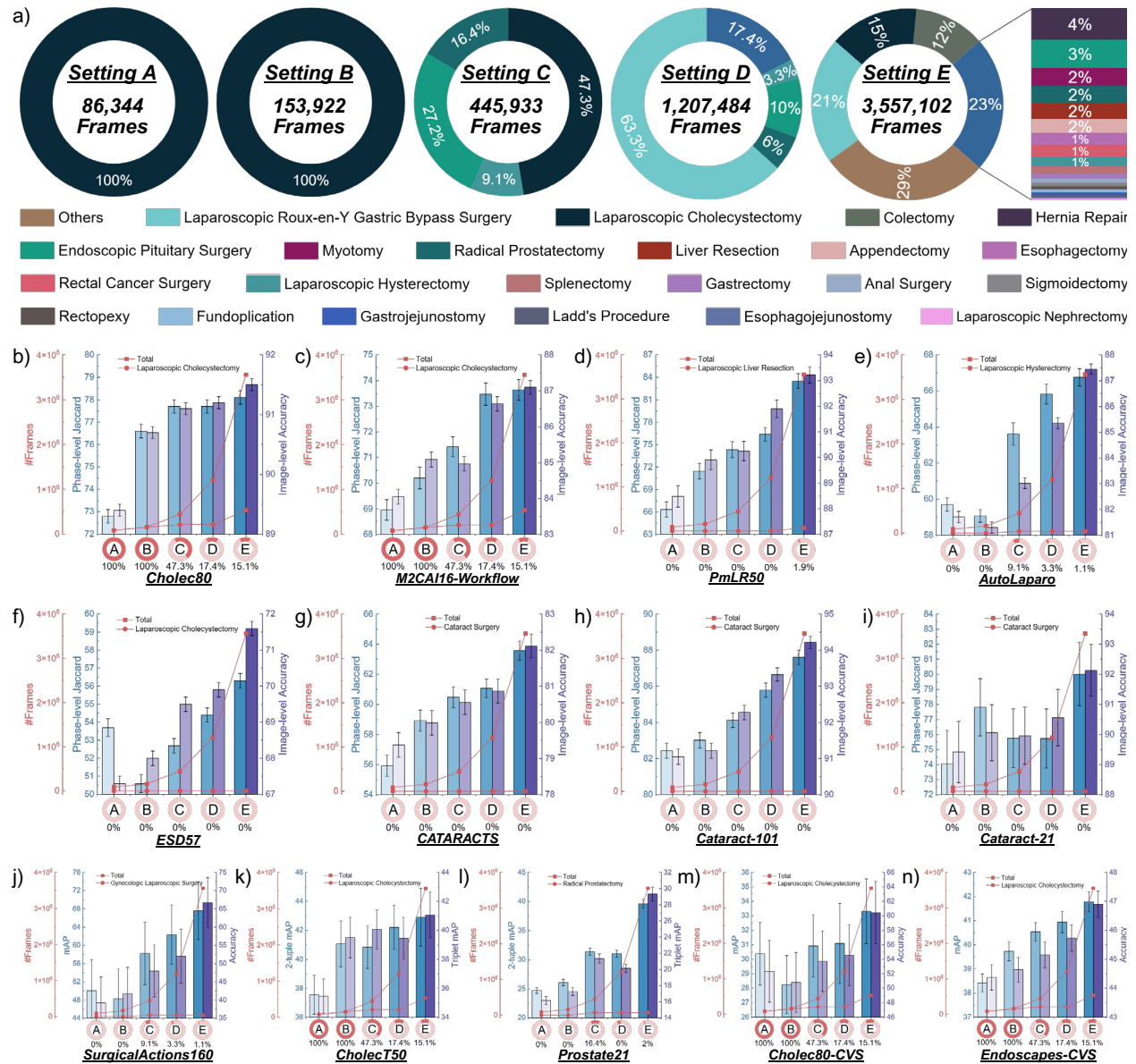


Figure 5. Experimental overview of how varying pre-training data scales affect the performance and generalization. Subfigure (a) illustrates the distribution of surgical procedures across progressively constructed sub-datasets. Subfigures (b)-(n) present performance variations across 13 surgical datasets under different data scales, along with data volumes and proportional distribution. Phase-level Jaccard and image-level accuracy are visualized using dual-axis bar plots: blue bars (left y-axis) denote phase-level Jaccard, while purple bars (right y-axis) represent image-level accuracy. Detailed results are presented in Supplementary Tables 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32 and 33. Created with BioRender.com under an academic license.

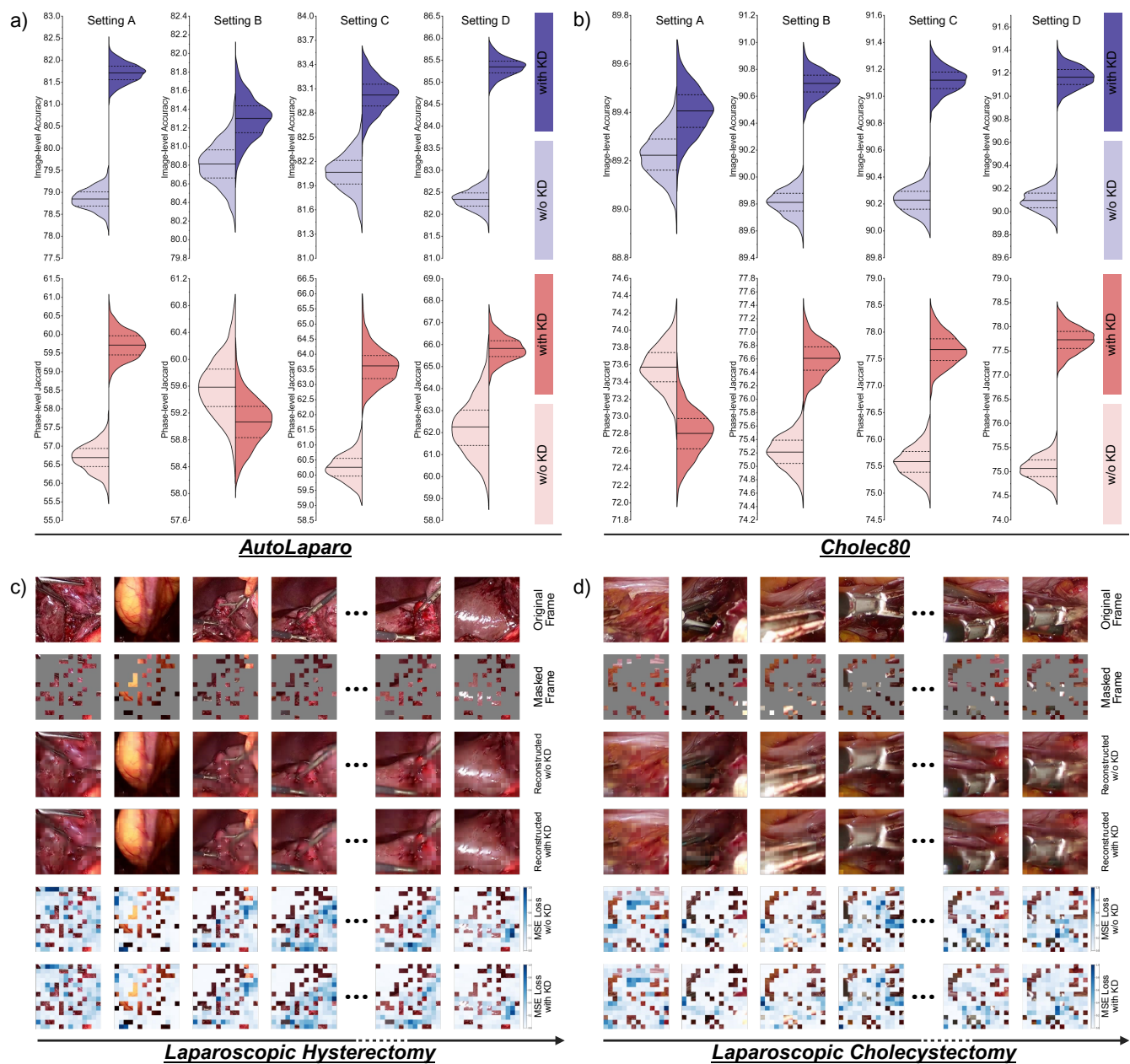


Figure 6. The effectiveness of knowledge distillation (KD). Subfigures (a)-(b) present the performance difference between SurgVISTA with and without knowledge distillation. Significance testing was conducted using the Wilcoxon signed-rank one-sided test, demonstrating that expert knowledge distillation consistently improves performance across different pre-trained sub-datasets and downstream datasets. Subfigures (c)-(d) qualitatively illustrate the reconstruction performance based on Setting D by showing the original frame, masked input, reconstructions with and without knowledge distillation, and corresponding MSE maps. The results demonstrate that knowledge distillation improves reconstruction fidelity by better preserving fine-grained anatomical details, as supported by reduced reconstruction error in related regions. Created with BioRender.com under an academic license.