

# A Surgical Foundation Model Reveals Task-Dependent Label Efficiency

Florian Philipp Stilz<sup>a,b,c,\*</sup>, Lorenzo Arboit<sup>a,b</sup>, Vinkle Srivastav<sup>a,b,d</sup>, CAMMA International Surgical Partners, Jacques Marescaux<sup>e</sup>, Sergio Alfieri<sup>f</sup>, Pietro Mascagni<sup>b,f</sup>, Nassir Navab<sup>c</sup>, Nicolas Padoy<sup>a,b</sup>

<sup>a</sup>University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France

<sup>b</sup>IHU Strasbourg, Strasbourg, France

<sup>c</sup>Technical University of Munich, Computer Aided Medical Procedures, Munich, Germany

<sup>d</sup>Indian Institute of Technology (IIT) Madras, Department of Data Science and AI, Wadhvani School of Data Science and AI (WSAI), Chennai, India

<sup>e</sup>IRCAD, Research Institute Against Digestive Cancer, Strasbourg, France

<sup>f</sup>Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

Developing label-efficient models is a central challenge in surgical AI due to the high cost and scarcity of expert annotation. While self-supervised foundation models adapt well to new tasks with minimal data, how label efficiency varies across different surgical tasks remains largely unexplored.

Here, we introduce SURGE, a surgical foundation model trained on SurgSpectrum-30M+, the largest pretraining dataset comprising over 30 million frames, with checkpoints released to enable further research. We systematically evaluate label efficiency across 5 task categories and 15 benchmarks. These range from temporal and spatial scene understanding to fine-grained reasoning tied to instrument–anatomy interactions and safety-critical maneuvers.

SURGE outperforms prior state-of-the-art on all benchmarks, even surpassing task-specific models on complex reasoning tasks. Crucially, we reveal a task-dependent scaling behavior: while scene understanding tasks saturate with minimal supervision, fine-grained reasoning tasks continue improving with substantially larger annotation budgets, providing a blueprint for allocating expert effort in complex domains.

**Code:** <https://github.com/CAMMA-public/SURGE>.

## 1. Introduction

Artificial intelligence is increasingly deployed to support expert decision-making in complex, high-stakes domains, ranging from healthcare and science to engineering and public safety [1]. In these settings, models are typically trained on annotated datasets that require substantial input from domain experts. However, generating such annotations is both time-consuming and costly, as expertise is scarce and tasks often demand deep contextual understanding. As a result, the development of new AI systems is frequently constrained not only by model capacity, but increasingly by the availability of high-quality labeled data [1]. Improving the efficiency with which new tasks can be learned is therefore critical. This can be approached from two complementary directions: developing models that achieve strong performance with fewer annotations and understanding how annotation requirements vary across tasks to allocate expert effort more effectively. Here, we address both challenges by developing a self-supervised foundation model for surgery and systematically characterizing how its performance scales with the amount of task-specific labeled data. We use this scaling behavior to quantify label efficiency across surgical perception and reasoning tasks, providing a principled basis for understanding and reducing the annotation requirements of AI systems in surgery.

Surgical procedures are complex, high-risk interventions that require continuous perception and timely reasoning to

support safe and effective decision-making. Automated analysis of endoscopic video acquired during minimally invasive surgery has the potential to improve intraoperative guidance, surgical training and postoperative assessment. Surgical AI [2, 3, 4, 5] is therefore a particularly high-stakes application, as errors and preventable complications during surgery contribute substantially to global morbidity and mortality [6, 7], highlighting the need for reliable intraoperative decision support. Importantly, surgical video analysis encompasses multiple levels of understanding, ranging from temporal perceptual tasks, such as recognizing surgical phases and steps [8, 9, 10, 11, 12] and spatial tasks such as segmenting instruments and anatomy [13, 14, 9, 11, 15, 16, 17], to higher-level reasoning about instrument–anatomy interactions [18, 19] and clinically grounded decision-making [13]. These diverse tasks vary substantially in complexity, raising the question of whether they also differ in the amount of supervision required to learn them effectively.

Recent advances in self-supervised and foundation model pretraining [20, 21, 22, 23] have enabled strong general-purpose visual representations by learning from large amounts of unlabeled data. When applied to surgical video, these approaches often yield competitive performance, particularly when sufficient labeled data is available [24, 25, 26]. Domain-adapted pretraining [24, 25, 27, 26, 28, 29] and multimodal approaches [30, 31] have further improved data efficiency. However, existing work has primarily focused on developing stronger representations, while the extent to which different categories of surgical tasks exhibit distinct label efficiency re-

\*Corresponding author: [florian.stilz@tum.de](mailto:florian.stilz@tum.de)

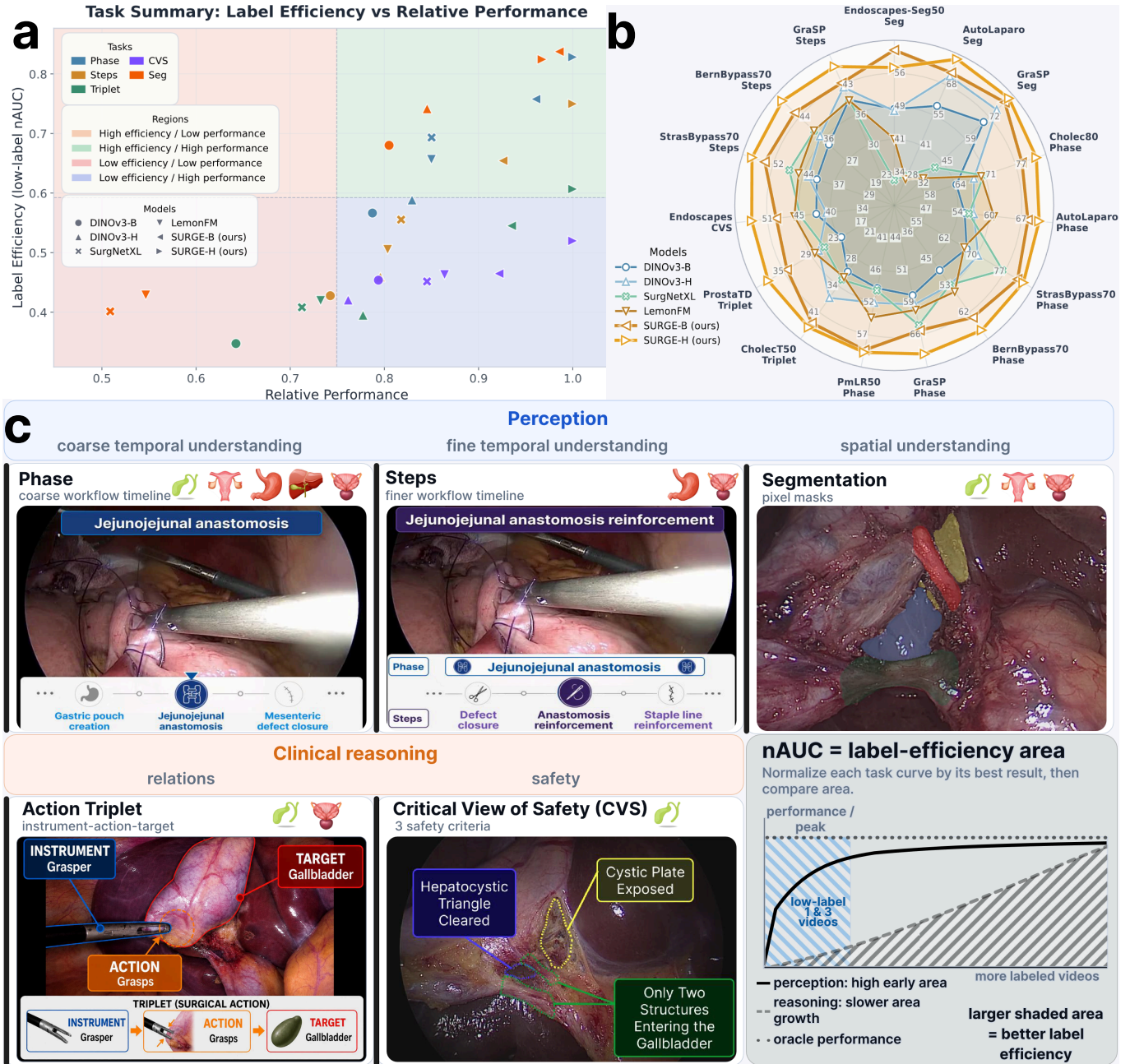


Fig. 1: **a**: Label efficiency across surgical tasks, shown as normalized area under the curve (nAUC) in low label regimes versus relative peak performance. Results are averaged across datasets per task. We observe a consistent ordering of tasks by data efficiency: perception tasks (segmentation (Seg), Phase, Steps) achieve strong performance in low-label regimes, while fine-grained and reasoning tasks (Action Triplet, CVS) require substantially more supervision. Notably, certain tasks exhibit low-data regimes with limited learning signal, indicating threshold effects that are not overcome by pretraining. **b**: Full-data performance comparison of our models with general-domain and surgical foundation baselines across 15 benchmarks. **c**: Five tasks are illustrated with representative examples and grouped into perception and clinical reasoning categories. Icons indicating anatomical structures denote the surgical procedures associated with each task (cholecystectomy, hysterectomy, liver resection, gastric bypass, prostatectomy). The nAUC metric is illustrated schematically.

mains largely unexplored.

In this work, we address this gap through a large-scale, systematic analysis of label efficiency in surgical AI. To this end, we introduce SURGE (Surgical Understanding through Representation and Generalization at Scale), a surgical foundation model trained via self-supervised learning on SurgSpectrum-30M+, a dataset of over 30 million surgical video frames. Using this unified representation, we evaluate performance across a spectrum of tasks, which we group into perceptual tasks (e.g., spatial and temporal understanding) and clinically grounded reasoning tasks (e.g., relational instrument–anatomy interactions and safety-critical maneuver recognition), which require higher-level judgment built upon perceptual understanding. Beyond overall performance, we study how these tasks scale with increasing supervision. While large-scale domain-adapted pretraining substantially reduces labeled training data requirements, it does not homogenize performance. Instead, tasks exhibit distinct label efficiencies that persist even under stronger supervision and higher-parameter adaptation, suggesting that the efficiency with which different aspects of surgical understanding can be learned depends not only on representation quality, but also on the structure of the task itself.

## 2. Results

Understanding label efficiency in surgical computer vision requires disentangling the interplay between representation quality, task structure, and supervision.

To this end we build on large-scale domain-adapted self-supervised pretraining, training ViT-Base and ViT-Huge+ on SurgSpectrum-30M+ (over 30 million surgical video frames), resulting in SURGE-B and SURGE-H, providing a strong and unified visual representation for surgical video analysis. As a result, SURGE provides a well-suited foundation for studying label efficiency, as it reduces performance bottlenecks associated with weaker or less robust representations. This unified representation enables us to study supervision scaling while reducing confounding effects from representation quality.

We evaluate these representations across a diverse set of surgical tasks, spanning two main categories: surgical perception and more fine-grained and clinically grounded reasoning tasks. Surgical perception tasks are separated into temporal workflow understanding at different granularities (phase and step recognition) and spatial perception (semantic segmentation), while reasoning tasks focus on analyzing surgical execution. These reasoning tasks include clinically grounded assessments (Critical View of Safety recognition) [32] as well as relational instrument-anatomy understanding (action triplet recognition), which has shown correlations with patient outcomes [33]. Action triplet recognition refers to predicting surgical actions as (instrument, verb, target) tuples of tool-tissue interactions.

We systematically vary the amount of supervision, from a single training video to the full training set for each task, and evaluate all models using a unified linear-probing protocol in which only a linear classifier is trained on frozen representations. This protocol isolates the contribution of representation

quality to downstream label efficiency. We compare SURGE with general-domain foundation models [23] and prior surgical self-supervised approaches [24, 28, 26]. Across tasks, performance improves consistently with increasing supervision, while stronger representations provide substantial gains, particularly in low-label regimes.

We quantify label efficiency using normalized area under the curve (nAUC; see Methods and see Fig. 1c). We compute nAUC in two settings: over the full training budget and in a low-label regime consisting of 1 and 3 training videos.

### 2.1. SURGE outperforms other foundation models across surgical AI tasks

SURGE achieves the best performance across all 15 benchmarks spanning diverse surgical tasks and supervision regimes (Fig. 1, Figs. 3, 4). For the subsequent label-efficiency analysis, we consider 14 benchmarks for which supervision can be systematically varied at the video level. AutoLaparo-Seg [9] is evaluated using its official train/validation/test splits only, as its data cannot be partitioned into individual videos for controlled supervision scaling. We illustrate this overall advantage with representative results from temporal workflow understanding and fine-grained relational reasoning.

For temporal workflow understanding, SURGE-B surpasses the strongest baseline on every dataset, achieving an average F1 score of 69.40 (95% CI: 63.30–75.85), compared with 63.63 (95% CI: 57.33–70.08) for the strongest competing foundation model. Scaling the representation from SURGE-B to SURGE-H further improves performance to 72.23 (95% CI: 66.18–78.51), demonstrating consistent gains with increased model capacity.

The advantage extends beyond temporal perception to more fine-grained relational reasoning. On instrument-anatomy interaction recognition, SURGE-B achieves 38.72 mAP (95% CI: 34.3–43.13), compared with 32.87 (95% CI: 28.6–37.13) for the strongest per-dataset baselines. SURGE-H further increases performance to 41.40 mAP (95% CI: 38.6–44.2). Thus, domain-adapted pretraining consistently improves surgical video representations across both temporal perception and fine-grained relational understanding, with further gains from scaling model capacity.

### 2.2. Consistently superior label efficiency across all annotation regimes

We next examine whether the performance gains of SURGE translate into greater label efficiency across the full range of annotation budgets (Fig. 2). Averaged across datasets, SURGE-H achieves an nAUC of 0.914 (95% CI: 0.884–0.942), substantially outperforming the strongest baselines on each dataset with 0.769, (95% CI: 0.724–0.809). This advantage persists in low-label settings (Fig. 2), where SURGE-H attains a low-label nAUC of 0.757 (95% CI: 0.683–0.835) versus 0.624 (95% CI: 0.549–0.692).

We further analyze the effect of model scale. SURGE-B and SURGE-H contain 85.67M and 840.63M parameters, respectively, and larger models consistently improve performance, with SURGE-H outperforming SURGE-B in 14 out of 15

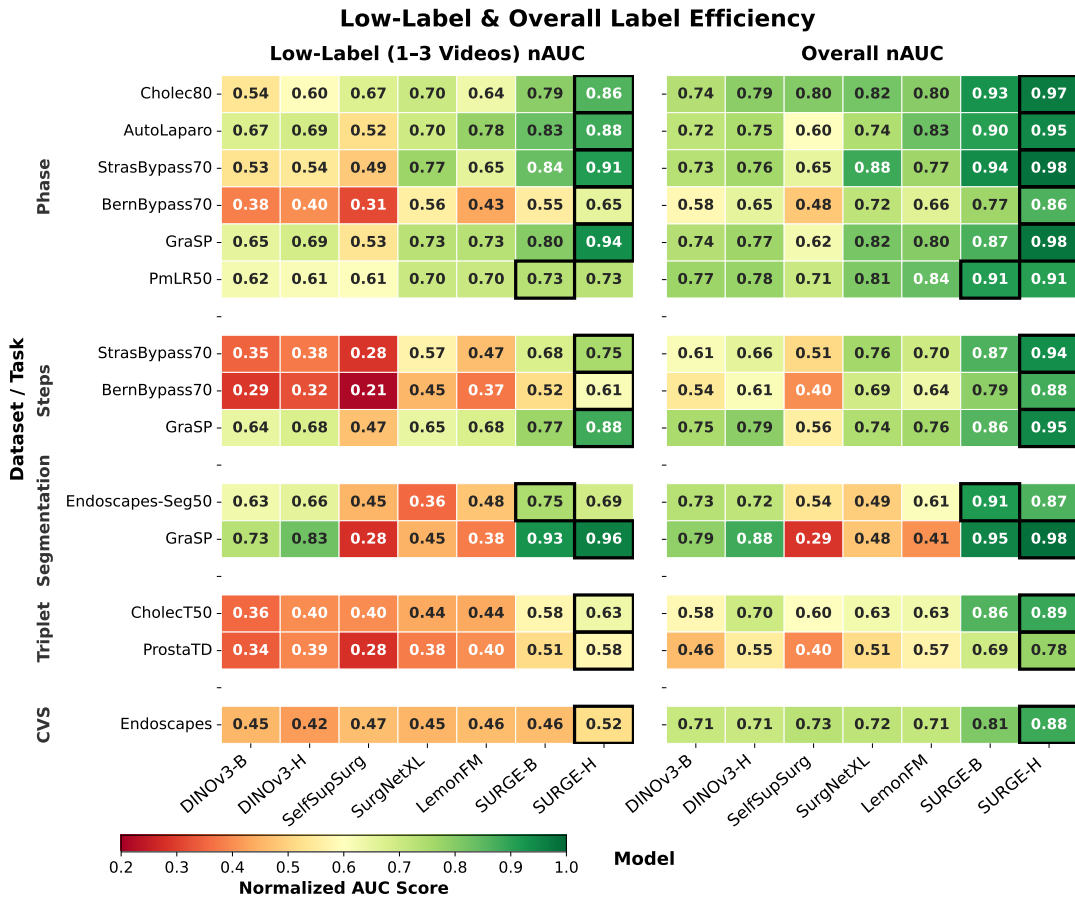


Fig. 2: **Overall and low-data regime label efficiency comparison:** measuring normalized AUC (nAUC) across tasks/datasets, highlighting most efficient model both in heatmap color coding and box highlighting.

benchmarks, as depicted in Fig. 1b. This improvement is especially pronounced in low-label settings; while task variance results in wide confidence intervals, paired analysis reveals a highly significant and substantial performance gain for the larger model (SURGE-H: 0.757 (95% CI: 0.683–0.835) and SURGE-B: 0.696 (95% CI: 0.619–0.769)); Wilcoxon signed-rank  $p < 0.01$ , Cohen’s  $d = 1.25$ ).

Overall, SURGE substantially improves label efficiency in surgical AI, enabling strong performance with reduced annotation effort. These gains are consistent across five surgical procedures (Fig. 2), although the magnitude of improvement varies across tasks, reflecting differences in how performance scales with supervision.

### 2.3. Domain adaptation bridges general and surgical representation strengths

We next examine why SURGE performs consistently across diverse surgical tasks. Prior surgical self-supervised approaches are particularly effective on workflow and clinically grounded tasks (Critical View of Safety), especially in low-data regimes, but are less effective for dense spatial perception and fine-grained instrument-anatomy interaction semantics. In contrast, general-domain self-supervised models provide stronger representations for tasks such as semantic segmentation and instrument-anatomy interaction (Fig. 2).

SURGE combines these complementary strengths. It retains the strong performance of surgical self-supervised approaches on workflow and clinically grounded tasks while exceeding general-domain models on dense spatial and interaction-level tasks. For example, SURGE-H achieves

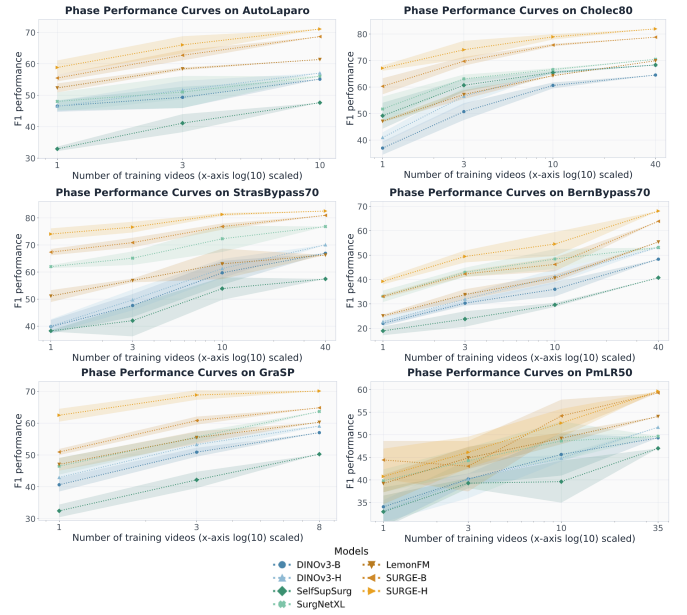


Fig. 3: **Coarse temporal perception (phase recognition) performance curves:** All subset runs are averaged over 3 runs with error bands highlighting the standard deviation.

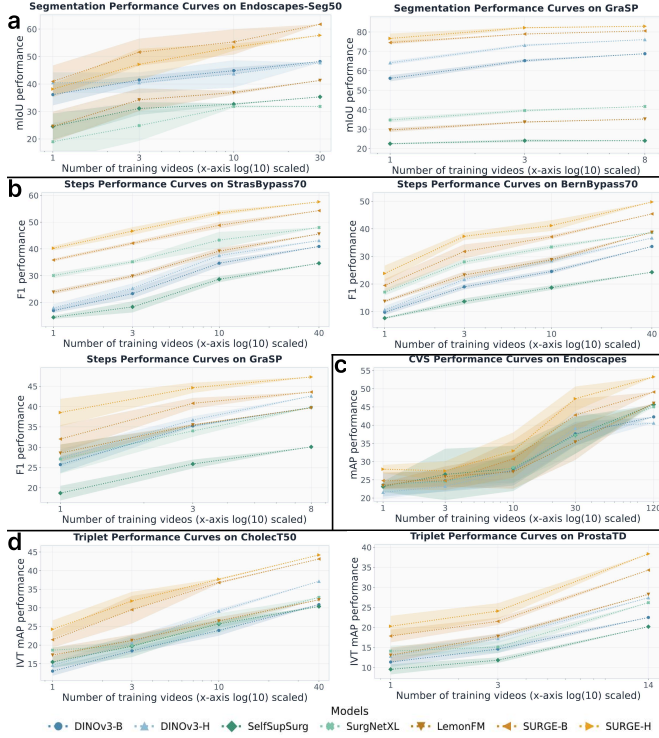


Fig. 4: **Additional perception & reasoning task performance curves:** **a** shows spatial understanding (semantic segmentation), **b**: fine temporal understanding (step recognition), **c**: safety maneuver analysis (CVS), and in **d** instrument-anatomy interaction (action triplet). All subset runs are averaged over 3 runs with error bands highlighting the standard deviation.

82.95 mIoU on GrASP, compared with 76.09 for DINOv3-H, and 44.20 mAP on CholecT50, compared with 37.13 for DINOv3-H. These results suggest that large-scale domain adaptation enables representations that capture both general visual structure and fine-grained surgical semantics, providing a potential basis for SURGE’s broad transferability and label efficiency.

#### 2.4. Perception tasks exhibit high label efficiency with task structure dependent variations

Our analysis reveals that surgical perception tasks exhibit high label efficiency, with performance saturating under minimal supervision. Tasks that rely primarily on visual cues or coarse temporal structure achieve strong performance with very limited annotated data, with diminishing gains as more data is added. For instance, in semantic segmentation on prostatectomy (GrASP dataset), SURGE-H attains an mIoU of  $76.68 \pm 2.53\%$  using only a single training video, with performance remaining nearly unchanged as the training set scales to eight videos as shown in Fig. 4a. Similarly, in phase recognition, SURGE-H reaches  $74.04 \pm 2.09\%$  F1 on gastric bypass (StrasBypass70) with one video and quickly plateaus with additional data, while achieving comparable behavior on cholecystectomy (Cholec80) depicted in Fig. 3. In these settings, a single labeled video often suffices to outperform competing pretrained surgical models trained on substantially larger datasets.

While this regime of high label efficiency is consistent across perception tasks, we find that structural properties of the task modulate the rate at which saturation is reached. Increasing the granularity of the prediction objective introduces a more pronounced dependence on labeled data. On StrasBypass70, while coarse phase recognition improves modestly with additional data (8.5% F1 gain from 1 to 40 videos), the more fine-grained step recognition task exhibits a substantially steeper scaling curve, improving by 17.3% F1 over the same range seen in Fig. 4b. Thus, increasing temporal granularity substantially increases the amount of supervision required to reach comparable performance.

Similarly, the complexity of the annotations influences data requirements even within this otherwise efficient regime. Comparing semantic segmentation across two datasets (GrASP and Endoscapes-Seg50) in Fig. 4a illustrates this effect. GrASP reaches strong performance quickly with limited data, whereas Endoscapes-Seg50 improves more gradually, starting from a lower baseline and continuing to benefit from additional annotations. This difference reflects the nature of the labels: GrASP focuses on a smaller set of well-defined structures, primarily surgical instruments, whereas Endoscapes-Seg50 includes a larger and more fine-grained label space, including small anatomical structures alongside instruments, resulting in a more complex prediction problem. As a result, the model must learn a more complex mapping from visual features to labels, which increases the amount of data required to achieve comparable performance.

Taken together, these results show that surgical perception tasks form a regime of high label efficiency characterized by rapid saturation with minimal supervision. Although factors such as task granularity, annotation density, and semantic complexity modulate the precise label efficiency, they do not alter the overall trend of efficient learning in these tasks. In the next section, we show that this regime breaks down for reasoning tasks, which exhibit distinctly different label efficiency and remain far from saturation even under substantially increased supervision.

#### 2.5. Reasoning tasks exhibit persistent data requirements

In contrast to the rapid saturation observed in perception tasks, reasoning tasks exhibit fundamentally different label efficiency, remaining far from saturation even under substantially increased supervision.

In the instrument-anatomy interaction tasks (action triplet recognition) on the CholecT50 and ProstaTD benchmarks shown in Fig. 4d, SURGE-H maintains a consistent lead over all baselines, yet displays a remarkably persistent log-linear scaled curve that still allows for sustained performance gains under constant increase in labeled data. For instance, on ProstaTD, SURGE-H begins at  $20.30 \pm 2.53$  mAP and requires 14 training videos to reach 38.36 mAP, a trajectory that lacks the ‘jump-start’ characteristic, where 1-video performance is already a significant fraction of the ceiling, observed in temporal (phase) or spatial (semantic segmentation) perception tasks. This is also supported by the nAUC metric, where perception tasks achieve average scores of 0.925,

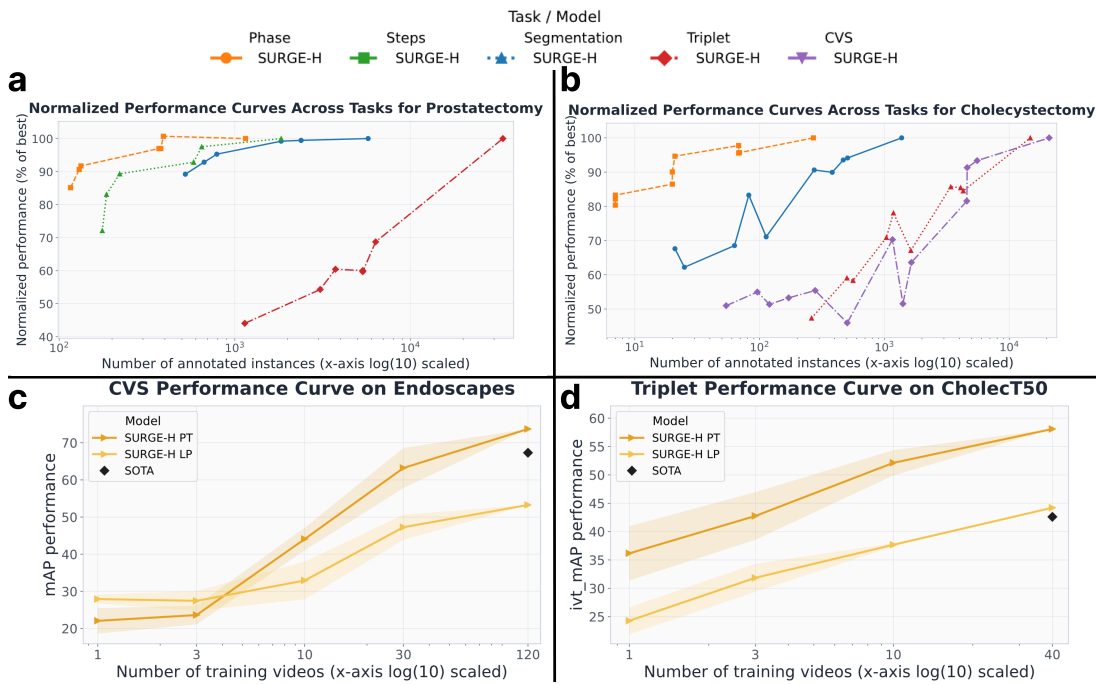


Fig. 5: **Label instances and label efficiency comparison across tasks & higher parameter adaptation results:** **a** shows task comparisons for prostatectomy and **b** for cholecystectomy. Highlighting severe annotation density differences across tasks. **c&d:** We select the best model for this analysis (SURGE-H) and compare linear probing (SURGE-H LP) and partial tuning of the backbone (SURGE-H PT). All subset runs are averaged over 3 runs with error bands highlighting the standard deviation. SOTA on Endoscopes is LG-CVS [34] and on CholecT50 is TripletGCN [35].

0.940, and 0.919 for semantic segmentation, phase recognition, and step recognition, respectively, compared to 0.832 for action triplet recognition. This difference is also present in the low-label regime, where semantic segmentation, phase recognition, and step recognition achieve average low-label nAUC values of 0.825, 0.828, and 0.750, respectively, versus 0.607 for instrument-anatomy interaction (action triplet recognition).

This structural bottleneck is most pronounced in the clinical quality assessment (Critical View of Safety (CVS)), depicted in Fig. 4c, on the Endoscopes dataset. Despite leveraging the same large-scale pretrained representation, performance for SURGE-H starts at  $27.92 \pm 1.17$  mAP and requires substantially more labeled data, reaching 53.26 mAP only when scaling to 120 training videos. In contrast to other tasks, we observe an apparent threshold-like behavior, where all models exhibit limited learning signal when trained with 10 or fewer videos, and only begin to improve meaningfully beyond this regime. This pattern is further reflected in the label-efficiency metric (nAUC), which reaches only 0.882 when evaluated across all training budgets and drops to 0.520 in the low-label regime. Notably, the low-label nAUC is substantially lower than that observed for any other task, indicating that this task benefits less from limited supervision. This suggests that certain reasoning tasks require a minimum amount of supervision before meaningful generalization can occur.

## 2.6. Label instances as an indicator of supervision requirements

To further ground these differences in data requirements, we next quantify the underlying supervision effort associated with each task, providing a practical proxy for label efficiency. While true annotation cost is a complex composite of mechanical time (e.g., pixel-wise segmentation) and cognitive

load (e.g., clinical consensus for safety criteria), direct measurement of human hours across retrospective, multi-center datasets is very difficult and annotator specific. To overcome this limitation and provide a standardized baseline, we introduce label instances, the absolute count of all explicitly annotated instances across a dataset, as an objective, macro-level proxy for total annotation effort. By systematically quantifying the volume of human-generated labels required per task category, we offer a first-of-its-kind approximation of comparative supervision demands. As shown in Fig. 5a-b, this resulting distribution closely mirrors the previously observed scaling patterns, demonstrating that total label instances serves as a powerful indicator of the intrinsic human effort required to solve different surgical tasks.

Notably, more clinically relevant tasks requiring clinical reasoning and relational instrument-anatomy understanding (e.g., CVS and action triplet recognition) exhibit substantially higher label counts, reflecting significantly greater annotation effort. Despite this increased supervision, their performance continues to improve with additional annotations without clear signs of saturation. This indicates that these tasks remain data-hungry even under substantially higher annotation volumes.

In contrast, temporal and spatial perception tasks rely on comparatively fewer annotations and exhibit more favorable learning efficiency. Among these, semantic segmentation stands out with the highest number of labels per video due to its dense annotation structure, yet still benefits from comparatively efficient learning dynamics.

Overall, these findings reveal a pronounced imbalance in supervision demands across tasks and indicate that label efficiency is strongly task-dependent.

### 2.7. Higher parameter adaptation surpasses supervised SOTA without altering label efficiency

Given the strong task-dependent differences in label efficiency observed above, we next examine whether these dynamics can be altered through higher parameter adaptation, in which more parameters of the pretrained backbone are updated using task-specific supervision rather than training only a linear classifier. While higher parameter adaptation substantially improves absolute performance, enabling data-intensive tasks such as relational instrument-anatomy recognition and clinical reasoning to surpass task-specific supervised state-of-the-art baselines, the fundamental label efficiency with respect to supervision remains unchanged.

As shown in Fig. 5c-d, partial adaptation consistently shifts performance upwards, yet preserves the relative gains obtained from additional supervision. Notably, this upward shift yields absolute state-of-the-art gains of 6.4% mAP on Endoscapes and 15.5% mAP on CholecT50 action triplet recognition. These improvements are achieved using our generalized approach, without relying on the complex, task-specific modifications used by previous baselines, such as segmentation-based supervision [34] or ensemble strategies [35].

## 3. Discussion

Our results provide a systematic view of how surgical AI learns from expert supervision across tasks with fundamentally different demands. Large-scale domain-adapted self-supervised pretraining produces a broadly transferable representation that consistently outperforms both general-domain and prior surgical foundation models across 15 benchmarks, while providing substantial gains in low-label regimes. However, our results also reveal that representation quality alone does not determine label efficiency. Surgical perception tasks typically reach high performance with minimal supervision and rapidly saturate, whereas clinically grounded and relational reasoning tasks continue to benefit from substantially more supervision. Increasing the number of adapted parameters raises absolute performance and enables state-of-the-art results on these more demanding tasks, but does not fundamentally alter their scaling with supervision. Together, these findings suggest that representation quality determines how effectively models can exploit available supervision, whereas the structure of the downstream task determines how much supervision is ultimately required.

### 3.1. The saturation of perception tasks

The rapid saturation observed in spatial understanding tasks (semantic segmentation) and temporal understanding (phase recognition) suggests that the field may be approaching a milestone in perceptual learning. Our results show that with sufficiently large domain-adapted pretraining, the model’s internal representation of surgical instruments, tissue textures, and broad procedural states becomes “task-ready” with very limited label supervision. In these regimes, a single training video acts not as a source of feature learning, but as a “Rosetta

Stone” that simply maps existing high-quality representations to class names.

This finding suggests an immediate pivot in how we allocate annotation resources. For standard procedures and instrument-centric tasks, the pursuit of larger datasets may exhibit diminishing returns. Instead, these “solved” perceptual components should be leveraged as automated foundations to assist in the more difficult task of mapping complex medical/surgical logic.

### 3.2. The persistent reasoning gap and the need for hybrid modeling

In contrast to the saturation of perception, fine-grained and reasoning tasks such as action triplet recognition and Critical View of Safety (CVS) assessment exhibit a “Reasoning Gap” that is not fully addressed by scale alone. These tasks require the model to move beyond *identification* toward *compositional logic* and *clinical judgment*. The persistent growth of the scaling curves observed in these domains indicates that the model is actively learning the underlying rules of surgery frame-by-frame, rather than solely relying on pretrained visual archetypes.

This resistance to pretraining gains does not suggest a limitation of the underlying representations, adaptation strategy, or model capacity, but rather may reflect the current frontier of deep learning. To cross this Reasoning Gap, we appear to require substantially more annotated data efforts compared to prior perception tasks or a transition toward hybrid methodologies. Such approaches might include a combination of largely-pretrained foundation models with task-specific architectural modeling, the integration of multi-modal clinical inputs, or the development of pretraining objectives that explicitly target causal and relational structures to learn those complicated problems under limited supervision efficiently.

### 3.3. A strategic roadmap to jumpstart Surgical AI

By characterizing how data requirements vary across surgical tasks, our work provides a practical framework for understanding their label efficiency and guiding annotation strategies. This perspective enables a more principled allocation of labeling effort toward clinically impactful objectives:

- **Low-shot transfer for perception:** Tasks such as instrument segmentation and coarse video understanding (phase recognition) exhibit high label efficiency and benefit strongly from domain-adapted pretraining. For these tasks, annotation requirements can be substantially reduced, opening the door to few-shot or weakly supervised approaches.
- **Targeted annotation for reasoning:** In contrast, tasks involving relational instrument-anatomy understanding and clinical assessment remain significantly more data-intensive. Expert annotation effort, being scarce and costly, should therefore be prioritized toward these settings.

- **Toward hybrid surgical AI:** Bridging the gap between perception and higher-level understanding is likely to require approaches that go beyond representation learning, combining strong visual foundations with task-aware modeling, temporal reasoning, and integration of clinically relevant context. Such approaches may also be important for achieving the robustness required for deployment in real-world clinical settings.

Ultimately, our findings suggest that while foundation models have substantially advanced the visual perception capabilities of AI, translating these representations into data-efficient reasoning remains a key challenge. By grounding this perspective in task-dependent data requirements, we outline a path toward more efficient, scalable, and clinically meaningful AI systems.

#### 3.4. Limitations & Future Work

While our study provides a controlled analysis of representation quality through linear probing and partial fine-tuning, this protocol inherently underestimates the peak performance achievable through task-specific fine-tuning. To realize its full potential, we release this foundation model as a shared tool for the research community. Future optimization, incorporating this model into specialized pipelines via architectural adapters or proxy losses, will enable practitioners to deeply adapt these baseline representations for their own specialized clinical applications.

A more fundamental limitation lies in the ‘‘Reasoning Gap’’ identified in our scaling analysis. The persistent log-linear data requirements for CVS and action triplet recognition suggest that distillation-based pretext tasks may not fully capture the causal and logical dependencies of surgery. To resolve this, future pretraining should move toward embedding a deeper clinical reasoning behind observed videos and multimodal foundations, specifically targeting vision-text alignment (e.g., surgical reports). Such models could enable zero-shot transfer and may facilitate semi-automated annotation in label-sparse environments, effectively ‘‘jump-starting’’ the development of AI for rare or specialized medical procedures.

## 4. Methods

### 4.1. SurgSpectrum-30M+: pretraining dataset generation

In order to unlock a truly label-efficient and multi-procedure foundation model, we construct **SurgSpectrum-30M+**, the largest surgical pretraining corpus to date. The dataset is assembled from a combination of public and proprietary sources across multiple centers and is designed to capture broad variability across surgical practice, spanning 132 procedures and 47 anatomical targets; more details can be seen in Fig. 6. We curated this dataset from multiple public sources [8, 36, 11, 19, 37, 38, 39, 40, 41, 42, 43, 10, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 30] as well as in-house surgical recordings and sampled frames from each video at 1 FPS which ensures a certain degree of visual diversity. To avoid any data leakage for potential downstream tasks we ensured to

remove all videos/frames associated to validation or test sets of the public sources. As a critical preprocessing step, we train an out-of-body detector to remove non-informative content, including preparation phases and external camera views, thereby retaining only intraoperative endoscopic footage. This filtering step reduces visual noise and enforces domain specificity, resulting in a dataset that emphasizes clinically meaningful patterns while preserving diversity. The final corpus comprises **30,381,943 frames from 38,753 videos**.

### 4.2. Model architecture

The network employed in this work is based on the transformer architecture [54], which leverages self-attention to model long-range dependencies and capture global contextual interactions. Originally developed for natural language processing, transformer-based models have since demonstrated strong performance in visual representation learning by operating on images as sequences of patches [55]. This formulation enables the encoding of complex spatial and semantic relationships within visual data.

Specifically, input images are partitioned into non-overlapping patches of size  $16 \times 16$ , which are flattened and projected into a learnable embedding space. Positional information is incorporated using rotary positional embeddings (RoPE) [56], allowing the model to capture relative spatial relationships between patches. The resulting sequence of patch embeddings is processed by a stack of transformer layers, yielding both global image-level representations and localized patch features.

We consider two model scales based on the Vision Transformer (ViT) design: ViT-Base and ViT-Huge+. Both models are initialized from pretrained DINOv3 checkpoints and subsequently adapted to our surgical pretraining dataset. The base model comprises 12 transformer layers with a hidden dimension of 768 and 12 attention heads, whereas the huge+ variant consists of 32 transformer layers with a hidden dimension of 1280 and 20 attention heads, and uses SwiGLU as its feed-forward network.

### 4.3. Pretraining objectives

We adopt a discriminative self-supervised objective combining image-level and patch-level supervision. Specifically, the overall loss consists of a DINO loss [57] applied at the image level and an iBOT loss [22] applied at the patch level, complemented by additional regularization terms.

At the image level, the DINO objective aligns the outputs of a student and a teacher network operating on different views of the same image. Let  $p_s$  and  $p_t$  denote the softmax-normalized outputs of the student and teacher for a given view, respectively.  $G$  and  $L$  correspond to the sets of global and local views of an input image, respectively. The loss is defined as shown in Eq. 1:

$$\mathcal{L}_{\text{DINO}} = - \sum_{g \in G} \sum_{l \in L} p_t(g) \log p_s(l). \quad (1)$$

At the patch level, the iBOT objective, shown in Eq. 2, extends this formulation to masked image modeling. A subset of patches is masked in the student input, while the teacher

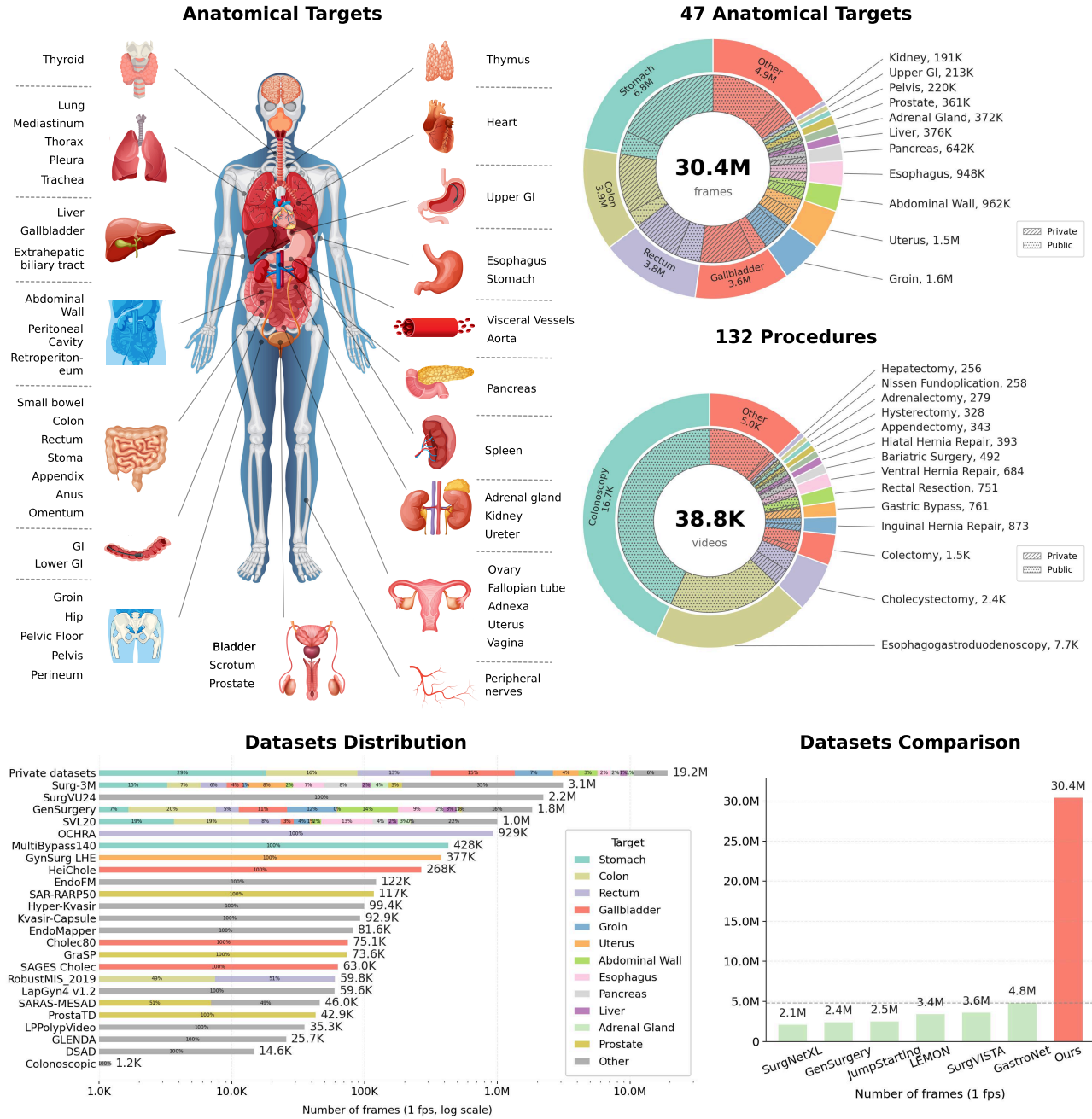


Fig. 6: **SurgSpectrum-30M+**. Overview of the proposed large-scale surgical pretraining dataset. Dataset composition is visualized on an absolute scale for total size, while source distributions are reported as percentages. Compared to existing resources, SurgSpectrum-30M+ increases data scale by more than 6× relative to the next largest dataset, comprising over 30 million frames from approximately 39k videos. The dataset spans more than 130 surgical procedures, substantially increasing procedural diversity and enabling large-scale representation learning in the surgical domain.

processes the full image. The student predictions for masked tokens are trained to match the corresponding teacher outputs, where  $\mathcal{M}$  denotes the set of masked patches:

$$\mathcal{L}_{\text{IBOT}} = - \sum_{g \in G} \sum_{l \in L} \sum_{i \in \mathcal{M}} p_t^{(i)}(g) \log p_s^{(i)}(l). \quad (2)$$

The final objective combines both terms, encouraging consistency at both global and local levels. Additional regularization such as feature decorrelation [58] is also applied denoted as  $\mathcal{L}_{\text{reg}}$ , depicted in Eq. 3:

$$\mathcal{L} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{IBOT}} + 0.1 \cdot \mathcal{L}_{\text{reg}}. \quad (3)$$

#### 4.4. Post-Training refinement

Following the established DINOv3 pretraining paradigm [23], we apply a post-training refinement stage consisting of resolution scaling and Gram anchoring. Prolonged self-supervised training can sometimes degrade patch-level features, reducing performance on dense prediction tasks. As demonstrated by DINOv3, this degradation is mitigated by a high-resolution adaptation step, in which the model is exposed to inputs at varying and higher resolutions. This enables the model to better capture fine-grained spatial details and improves robustness to resolution changes at inference time [23].

In conjunction, we apply Gram anchoring, a regularization objective that enforces consistency between feature correlations of the student and a teacher model. Concretely, it aligns the Gram matrices of patch features, thereby preserving spatial structure and preventing the collapse of dense representations. This is particularly important at high resolutions, where maintaining coherent local features becomes more challenging [23].

Together, these refinements restore patch-level consistency and significantly enhance performance on dense tasks, while preserving strong global representations.

#### 4.5. Downstream task evaluation

To comprehensively evaluate the learned representations, we consider five distinct surgical task categories across a total of 15 benchmarks. These tasks span complementary aspects of surgical understanding, including temporal workflow analysis, spatial perception, relational reasoning, and clinically grounded assessment. Importantly, they cover multiple procedures with diverse visual characteristics and workflow structures, enabling a systematic study of generalization across both tasks and domains.

Evaluating across multiple tasks and procedures is critical for two reasons. First, it allows us to assess whether learned representations capture transferable surgical knowledge beyond a single task or dataset. Second, it provides a controlled setting to analyze how label efficiency varies as a function of task structure, rather than being confounded by dataset-specific properties. To measure label efficiency in training-data subset experiments, we use three independent random video splits and report the mean  $\pm$  standard deviation across splits.

*Surgical Workflow Recognition.* We evaluate temporal understanding at two levels of granularity: coarse-grained surgical phases and fine-grained procedural steps. Phase recognition captures high-level stages of a procedure (e.g., preparation, dissection, closure), whereas step recognition focuses on more detailed sub-actions within each phase, requiring finer temporal discrimination.

For phase recognition, we consider Cholec80 [8] (cholecystectomy), GraSP [11] (prostatectomy), MultiBypass140 [10], comprising StrasBypass70 (gastric bypass, Strasbourg, France) and BernBypass70 (gastric bypass, Bern, Switzerland), AutoLaparo [9] (hysterectomy), and PmLR50 [59] (liver resection). For step recognition, we evaluate on the respective steps annotations of StrasBypass70, BernBypass70, and GraSP.

This distinction between phases and steps enables us to study how label efficiency varies with temporal granularity, as step recognition typically requires more precise temporal localization and finer semantic understanding.

*Surgical Semantic Segmentation.* To assess spatial perception, we evaluate semantic segmentation on Endoscapes-Seg50 [13], consisting of 50 videos with annotations for six semantic classes, and GraSP [11], which provides pixel-wise annotations for seven surgical tools. These tasks require dense spatial predictions and provide strong supervision signals at the pixel level, enabling analysis of label efficiency under highly localized supervision.

*Triplet Recognition.* We evaluate relational reasoning via triplet recognition, capturing interactions between surgical instruments, actions, and anatomical structures. For this we consider CholecT50 [18], which defines 100 unique instrument-verb-tissue combinations for cholecystectomy, and ProstaTD [19] for prostatectomy with 89 interaction classes. These tasks require joint reasoning over multiple entities and represent a higher level of semantic abstraction compared to purely spatial or temporal tasks.

*Critical View of Safety (CVS).* We evaluate clinically grounded understanding using the Critical View of Safety (CVS) task on the Endoscapes dataset [13]. CVS assessment is based on three criteria defined by Strasberg [32], which must be satisfied to ensure safe dissection during cholecystectomy. Each frame is annotated independently by three experts, and final labels are obtained via majority voting. This task requires high-level clinical reasoning and is particularly challenging due to its abstract and safety-critical nature.

*Evaluation metrics.* We use task-appropriate evaluation metrics for each category. For phase and step recognition, we report the F1 score, which captures the balance between precision and recall in temporally structured predictions. For semantic segmentation, we use mean Intersection over Union (mIoU), a standard metric for evaluating pixel-wise overlap between predicted and ground-truth masks. For triplet recognition and CVS assessment, we report mean Average Precision (mAP), which reflects performance in multi-label and

detection-style settings where multiple interactions or criteria may be present simultaneously and are threshold independent.

*Normalized area under the curve (nAUC).* To compactly quantify label efficiency across tasks and datasets, we use the normalized area under the curve (nAUC). The metric summarizes model performance as a function of annotation budget. For each task, performance is normalized by the maximum value achieved across all models. We then compute the area under the resulting performance curve using trapezoidal integration and normalize it by the range of the input domain, yielding a value in  $[0, 1]$ . Higher nAUC indicates higher performance relative to the task-specific best across the annotation budget, reflecting greater overall label efficiency.

Let  $\{(x_i, y_i)\}_{i=1}^n$  denote performance  $y_i$  measured at training set sizes  $x_i$ , sorted such that  $x_1 < \dots < x_n$ . We normalize performance as shown in Eq. 4:

$$\tilde{y}_i = \frac{y_i}{y_{\max}}, \quad \text{where } y_{\max} = \max_{1 \leq j \leq n} y_j. \quad (4)$$

The normalized area under the curve (nAUC) is computed using trapezoidal integration displayed in Eq. 5:

$$\text{nAUC} = \frac{1}{x_n - x_1} \sum_{i=1}^{n-1} \frac{\tilde{y}_i + \tilde{y}_{i+1}}{2} (x_{i+1} - x_i). \quad (5)$$

To emphasize low-label regime learning behaviour, we additionally compute a low-label regime nAUC over the first two points  $\tilde{y}_1$  and  $\tilde{y}_2$  representing the performance at 1 and 3 training videos respectively, shown in Eq. 6. Higher low-label regime nAUC indicates stronger performance when only a small number of labeled videos are available:

$$\text{nAUC}_{\text{low-label}} = \frac{1}{x_2 - x_1} \cdot \frac{\tilde{y}_1 + \tilde{y}_2}{2} (x_2 - x_1) = \frac{\tilde{y}_1 + \tilde{y}_2}{2}. \quad (6)$$

#### 4.6. Implementation details

*Pretraining.* We perform large-scale self-supervised pretraining for both model variants, ViT-Base (SURGE-B) and ViT-Huge+ (SURGE-H), on the SurgSpectrum-30M+. Pretraining is conducted for 180k iterations with a batch size of 2048, corresponding to approximately 12 epochs over the dataset.

Following this stage, we apply an additional post-refinement phase for 30k iterations with a reduced batch size of 64. This stage incorporates multi-resolution adaptation and Gram anchoring to further improve representation quality.

During multi-resolution training, we employ two global crops at resolutions of  $512 \times 512$  and  $768 \times 768$ , and three local crops for the student network at  $112 \times 112$ ,  $168 \times 168$ , and  $336 \times 336$ . Optimization is performed using a base learning rate of  $4 \times 10^{-4}$  and a weight decay of 0.04.

*Linear probing evaluation.* To evaluate the quality of the learned representations, we adopt a linear probing protocol across all downstream tasks. For image-level prediction tasks (i.e., all except semantic segmentation), we follow the protocol used in [23] by constructing a global representation via

concatenating the [CLS] token with the average pooled patch tokens, which is then passed to a single-layer MLP for task-specific prediction.

For semantic segmentation, we directly use the patch-level features and apply a linear classification head to obtain dense predictions, which are subsequently upsampled via bilinear interpolation to the original image resolution. For convolutional architectures, we apply an equivalent strategy by operating on the final feature maps.

During linear probing, the pretrained backbone remains frozen and only the task-specific head is optimized. To ensure fair comparison, we perform a systematic hyperparameter search over learning rates for each task and dataset over the respective validation sets, and apply the same protocol consistently across all evaluated methods, including both our models and the considered baselines. Learning rates were selected from the following set based on validation performance:  $[10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}, 10^{-1}]$ .

This protocol isolates the quality of the learned representations and enables a controlled comparison of label efficiency across tasks without confounding effects from task-specific fine-tuning.

*Higher parameter evaluation.* In addition, we consider a higher-capacity adaptation setting in which a subset of the pretrained backbone layers is fine-tuned together with the task-specific head. We perform this experiment only for SURGE-H, which achieved the best performance in the linear probing experiments, and conduct a hyperparameter search over the number of final layers to fine-tune ( $k \in \{1, 3, 6, 9, 12, 15\}$ ) and the learning rate separately for each of the two datasets described in section 2.7. The remaining backbone layers are kept frozen, and the best configuration on the respective validation set is used for the final evaluation on the corresponding test set.

#### Data availability

The pretraining corpus used in this study, **SurgSpectrum-30M+**, is a large-scale aggregation of both public and private surgical video sources. Publicly available components, including datasets such as Cholec80 [8], ProstaTD [19], SurgVu24 [49], OCHRA rectal cancer surgery [45], and many more (see section 4.1) can be accessed through their respective official repositories. The private portions of the corpus consist of institutional data that cannot be publicly released due to patient privacy regulations and institutional data-sharing agreements. For downstream evaluation, all datasets (Cholec80, CholecT50, GraSP, Endoscapes, etc.) are publicly available through their original providers.

#### Code availability

To support the reproducibility of our findings and facilitate further research in surgical foundation models, we provide the pretrained model weights for **SURGE-B** and **SURGE-H**,

along with the necessary loading-, inference-, and evaluation scripts on all downstream tasks. The implementation is based on PyTorch and is available at <https://github.com/CAMMA-public/SURGE>.

## Acknowledgements

This work was partially supported by French state funds managed by the ANR under Grants ANR-22-FAI1-0001 (project DAIOR), ANR-10-IAHU-02 (IHU Strasbourg), ANR-23-IACL-0004 (AI Cluster Grand Est ENACT) and by the Interdisciplinary Thematic Institute HealthTech (ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm) via the IdEx Unistra (ANR-10-IDEX-0002) and SFRI (STRATUS project, ANR-20-SFRI-0012). This work has also received funding from the European Union (ERC, Comp-SURG, 101088553). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## Author contributions

F.S., N.N. and N.P. conceived and designed the work. F.S. contributed to the technical implementation and conducted experiments. L.A. and P.M. provided clinical expertise and collected the data for the self-supervised pretraining. The members of the CAMMA International Surgical Partners provided clinical expertise and access to proprietary datasets. F.S., L.A., V.S., P.M., and N.P. contributed to the drafting and revising of the manuscript. All authors have received and reviewed the manuscript. N.N. and N.P. supervised the research.

## Competing interests

P.M. and N.P. are co-founders and shareholders of Scialytics. The remaining authors declare no competing interests.

## CAMMA International Surgical Partners

### Members

Luigi Boni<sup>1</sup>, Marta Goglia<sup>2</sup>, Gianfranco Silecchia<sup>2</sup>, Andrea Balla<sup>3</sup>, Salvador Morales-Conde<sup>3</sup>, Giovanni Guglielmo Laracca<sup>4</sup>, Claudio Fiorillo<sup>5</sup>, Giuseppe Quero<sup>5</sup>, Vincenzo Tondolo<sup>5</sup>, Ludovica Baldari<sup>1</sup>, Elisa Cassinotti<sup>1</sup>, Ludovica Guerriero<sup>6</sup>, Isacco Montroni<sup>7</sup>, Anna Fagotti<sup>8</sup>, Alice Zampolini<sup>8</sup>, Riccardo Oliva<sup>8</sup>, Xin Wang<sup>9</sup>, Matteo Pavone<sup>8</sup>, Nicolò Bizzarri<sup>8</sup>, Lise Lecointre<sup>10</sup>

### Affiliations

1. Department of General & Minimally Invasive Surgery Fondazione IRCCS - Ca' Granda - Ospedale Maggiore Policlinico di Milano, Milan, Italy
2. Department of Medical-Surgical Sciences and Translational Medicine, Faculty of Medicine and Psychology, Sapienza University of Rome, Italy

3. University Hospital Virgen Macarena, University of Sevilla, Seville, Spain
4. Sant'Andrea Hospital, Sapienza University, Rome, Italy
5. Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy
6. Monaldi Hospital, AORN dei Colli, Naples, Italy
7. IRCCS Fondazione Istituto Nazionale dei Tumori, Division of Colon and Rectal Cancer Surgery, Milan, Italy
8. Gynecologic Oncology Unit, Fondazione Policlinico Universitario A. Gemelli IRCCS, Università Cattolica del Sacro Cuore, Rome, Italy
9. West China Hospital of Sichuan University, Chengdu, China
10. Department of Gynecologic Surgery, University Hospitals of Strasbourg, Strasbourg, France; ICube, Laboratory of Engineering, Computer Science and Imaging, Department of Robotics, Imaging, Teledetection and Healthcare Technologies, University of Strasbourg, CNRS, UMR 7357, Strasbourg, France

## References

- [1] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.
- [2] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science—from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022.
- [3] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.
- [4] S Swaroop Vedula and Gregory D Hager. Surgical data science: the new knowledge domain. *Innovative surgical sciences*, 2(3):109–121, 2017.
- [5] Pietro Mascagni, Deepak Alapatt, Luca Sestini, Maria S Altieri, Amin Madani, Yusuke Watanabe, Adnan Alseidi, Jay A Redan, Sergio Alfieri, Guido Costamagna, et al. Computer vision in surgery: from potential to clinical value. *npj Digital Medicine*, 5(1):163, 2022.
- [6] Dmitri Nepogodiev, Janet Martin, Bruce Biccard, Alex Makupe, Aneel Bhangu, D Nepogodiev, et al. Global burden of postoperative death. *Lancet*, 393(10170):401, 2019.
- [7] Marieke Zegers, Martine C de Bruijne, Bertus de Keizer, Hanneke Merten, Peter P Groenewegen, Gerrit van der Wal, and Cordula Wagner. The incidence, root-causes, and outcomes of adverse events in surgical units: implication for potential prevention strategies. *Patient safety in surgery*, 5(1):13, 2011.
- [8] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.
- [9] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer, 2022.
- [10] Joël L Lavanchy, Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Paolo Fiorini, Beat P Müller-Stich, Philipp C Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. *International journal of computer assisted radiology and surgery*, pages 1–9, 2024.

- [11] Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, et al. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, page 103726, 2025.
- [12] Yiru Ye, Wenlong Wang, Yonghao Long, Chi-Fai Ng, QingYin Zhou, Dongren Yang, Qi Dou, Mengya Xu, and Zhifang Pan. Self-supervised adaptive transformer for surgical step recognition in robotic-assisted radical prostatectomy. *IEEE Robotics and Automation Letters*, 11(6):6568–6575, 2026.
- [13] Aditya Murali, Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, et al. The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *arXiv preprint arXiv:2312.12429*, 2023.
- [14] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.
- [15] Devanish N Kamtam, Joseph B Shrager, Satya Deepya Malla, Xiaohan Wang, Nicole Lin, Juan J Cardona, Serena Yeung-Levy, and Clarence Hu. A fine-tuned foundational model surgisam2 for surgical video anatomy segmentation and detection. *Scientific Reports*, 15(1):35961, 2025.
- [16] Devanish N Kamtam, Joseph B Shrager, Satya Deepya Malla, Xiaohan Wang, Nicole Lin, Juan J Cardona, Serena Yeung-Levy, and Clarence Hu. Surgisam2: fine-tuning a foundational model for surgical video anatomy segmentation and detection. *arXiv preprint arXiv:2503.03942*, 2025.
- [17] Ronald LPD de Jong, Tim JM Jaspers, Raf AH Vervoort, Aron FHA Bakker, Yiping Li, Jip L Tolenaar, Jelle P Ruurda, Willem M Brinkman, Josien PW Pluim, Marcel Breeuwer, et al. Surgical anatomy recognition with context learning using foundation representations. *arXiv preprint arXiv:2606.22124*, 2026.
- [18] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.
- [19] Yiliang Chen, Zhixi Li, Cheng Xu, Alex Qinyang Liu, Xuemiao Xu, Jeremy Yuen-Chun Teoh, Shengfeng He, and Jing Qin. Prostatd: A large-scale multi-source dataset for structured surgical triplet detection. *arXiv preprint arXiv:2506.01130*, 2025.
- [20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.
- [21] Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *arXiv preprint arXiv:2203.14415*, 2022.
- [22] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.
- [23] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Badasar, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.
- [24] Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, et al. Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis*, 88:102844, 2023.
- [25] Dominik Batić, Felix Holm, Ege Özsoy, Tobias Czempel, and Nassir Navab. Endovit: pretraining vision transformers on a large collection of endoscopic images. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1085–1091, 2024.
- [26] Ronald de Jong, HJ Carolus, HA Franciscus, Romy C van Jaarsveld, Richard van Hillegersberg, PW Josien, Peter HN de With, Yasmina al Khalil, Fons van Der Sommen, et al. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, page 103873, 2025.
- [27] Martijn R Jong, Tim GW Boers, Kiki N Fockens, Jelmer B Jukema, Carolus HJ Kusters, Tim JM Jaspers, RAH van Eijck van Heslinga, Floor C Slooter, Maarten R Struyvenberg, Raf Bisschops, et al. Gastronet-5m: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. *Gastroenterology*, 2025.
- [28] Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. Lemon: A large endoscopic monocular dataset and foundation model for perception in surgical settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 42659–42669, 2026.
- [29] Ronald LPD de Jong, Yiping Li, Tim JM Jaspers, Romy C van Jaarsveld, Gino M Kuiper, Franco Badaloni, Richard van Hillegersberg, Jelle P Ruurda, Fons van der Sommen, Josien PW Pluim, et al. Towards effective surgical representation learning with dino models. In *Medical Imaging with Deep Learning*, 2026.
- [30] Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L Lavanchy, Jacques Marescaux, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, 105:103644, 2025.
- [31] Florian Stilz, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Clipper: Contextual video-language pretraining on long-form intra-operative surgical procedures for event recognition. *arXiv preprint arXiv:2603.24539*, 2026.
- [32] Steven M Strasberg and Michael L Brunt. Rationale and use of the critical view of safety in laparoscopic cholecystectomy. *Journal of the American College of Surgeons*, 211(1):132–138, 2010.
- [33] John R Heard, Atharva Deo, Umar Ghaffar, Runzhuo Ma, Cherine Yang, Ujjwal Pasupulety, Jasmine Lin, Carter M Prentice, Hyung L Kim, Melissa Assel, et al. Ai model predicts patient outcomes from surgical gestures and provides insights into explainability. *npj Digital Surgery*, 1(1):4, 2026.
- [34] Aditya Murali, Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Didier Mutter, and Nicolas Padoy. Latent graph representations for critical view of safety assessment. *IEEE Transactions on Medical Imaging*, 43(3):1247–1258, 2023.
- [35] Xiaoyang Zou, Derong Yu, and Guoyan Zheng. Capturing action triplet correlations for accurate surgical activity recognition. *Computerized Medical Imaging and Graphics*, page 102604, 2025.
- [36] Samuel Schmidgall, Ji Woong Kim, Jeffrey Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*, 2024.
- [37] Pablo Azagra, Carlos Sostres, Ángel Ferrández, Luis Riazuelo, Clara Tomasini, O León Barbed, Javier Morlana, David Recasens, Victor M Batlle, Juan J Gómez-Rodríguez, et al. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data*, 10(1):671, 2023.
- [38] Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):3, 2023.
- [39] Andreas Leibetseder, Sabrina Kletz, Klaus Schoeffmann, Simon Keckstein, and Jörg Keckstein. GLENDa: gynecologic laparoscopy endometriosis dataset. In *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II*, volume 11962 of *Lecture Notes in Computer Science*, pages 439–450. Springer, 2020.
- [40] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical image analysis*, 86:102770, 2023.
- [41] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.
- [42] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142, 2021.
- [43] Andreas Leibetseder, Stefan Petscharnig, Manfred Jürgen Primus, Sabrina Kletz, Bernd Münzer, Klaus Schoeffmann, and Jörg Keckstein. Lapgyn4: a dataset for 4 automatic content analysis problems in the do-

- main of laparoscopic gynecology. In *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*, pages 357–362. ACM, 2018.
- [44] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International conference on medical image computing and computer-assisted intervention*, pages 101–111. Springer, 2023.
  - [45] Walaa Ghamrawi, Nathan Curtis, Jialang Xu, Matt Boal, Evangelos Mazomenos, Eddie Edwards, Danail Stoyanov, and Nader Francis. A rectal cancer surgery dataset: Use of artificial intelligence to aid automation of error identification. *Scientific Data*, 12(1):180, 2025.
  - [46] Tobias Ross, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299*, 2020.
  - [47] Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. Surg-3m: A dataset and foundation model for perception in surgical settings. *arXiv e-prints*, pages arXiv–2503, 2025.
  - [48] Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023.
  - [49] Aneeq Zia, Max Berniker, Rogerio Nespolo, Conor Perreault, Ziheng Wang, Benjamin Mueller, Ryan Schmidt, Kiran Bhattacharyya, Xi Liu, and Anthony Jarc. Surgical visual understanding (surgvu) dataset. *arXiv preprint arXiv:2501.09209*, 2025.
  - [50] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ld-polypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *International conference on medical image computing and computer-assisted intervention*, pages 387–396. Springer, 2021.
  - [51] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Inna Skarga-Bandurova, Elettra Oleari, Alice Leporini, Carmela Landolfo, Pengfei Zhao, Xi Xiang, Gongning Luo, et al. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. *arXiv preprint arXiv:2104.03178*, 2021.
  - [52] Deepak Alapatt, Jennifer Eckhoff, Zhiliang Lyu, Yutong Ban, Jean-Paul Mazellier, Sarah Choksi, Kunyi Yang, Po-Hsing Chiang, Noemi Zorzetti, Samuele Cannas, et al. The sages critical view of safety challenge: A global benchmark for ai-assisted surgical quality assessment. *arXiv preprint arXiv:2509.17100*, 2025.
  - [53] Sahar Nasirihaghghi, Negin Ghamsarian, Leonie Peschek, Matteo Munari, Heinrich Husslein, Raphael Sznitman, and Klaus Schoeffmann. Gynsurg: A comprehensive gynecology laparoscopic surgery dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13141–13147, 2025.
  - [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
  - [55] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
  - [56] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.
  - [57] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
  - [58] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.
  - [59] Diandian Guo, Weixin Si, Zhixi Li, Jialun Pei, and Pheng-Ann Heng. Surgical workflow recognition and blocking effectiveness detection in laparoscopic liver resection with pringle maneuver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3220–3228, 2025.