A Comparative Study in Surgical AI: Datasets, Foundation Models, and Barriers to Med-AGI

Kirill Skobelev[1*] Eric Fithian[1] Yegor Baranovski[1] Jack Cook[2]
Sandeep Angara[2] Shauna Otto[2] Zhuang-Fang Yi[2] John Zhu[2]
Daniel A. Donoho[2,3†] X.Y. Han[1,4†] Neeraj Mainkar[2†]
Margaux Masson-Forsythe[2†]

[1]Center for Applied AI, Chicago Booth, Chicago, IL, USA
[2]Surgical Data Science Collective, Washington D.C., USA
[3]Children's National Hospital, Washington D.C., USA
[4]Operations Management & Tolan Center for Healthcare, Chicago Booth, Chicago, IL, USA

March 28, 2026

Abstract

Recent Artificial Intelligence (AI) models have matched or exceeded human experts in several benchmarks of biomedical task performance, but have lagged behind on surgical image-analysis benchmarks. Since surgery requires integrating disparate tasks—including multimodal data integration, human interaction, and physical effects—generally-capable AI models could be particularly attractive as a collaborative tool if performance could be improved. On the one hand, the canonical approach of scaling architecture size and training data is attractive, especially since there are millions of hours of surgical video data generated per year. On the other hand, preparing surgical data for AI training requires significantly higher levels of professional expertise, and training on that data requires expensive computational resources. These trade-offs paint an uncertain picture of whether and to-what-extent modern AI could aid surgical practice. In this paper, we explore this question through a case study of surgical tool detection using state-of-the-art AI methods available in 2026. We demonstrate that even with multi-billion parameter models and extensive training, current Vision Language Models fall short in the seemingly simple task of tool detection in neurosurgery. Additionally, we show scaling experiments indicating that increasing model size and training time only leads to diminishing improvements in relevant performance metrics. Thus, our experiments suggest that current models could still face significant obstacles in surgical use cases. Moreover, some obstacles cannot be simply "scaled away" with additional compute and persist across diverse model architectures, raising the question of whether data and label availability are the only limiting factors. We discuss the main contributors to these constraints and advance potential solutions.

Results Summary. We present findings from six experiments. (1) We evaluate zero-shot surgical tool detection performance across 19 open-weight Vision Language Models (VLMs) from 2023 to early-2026 on SDSC-EEA, a large video dataset consisting of endoscopic endonasal approach (EEA) neurosurgical procedures. Despite dramatic increases in model scale and benchmark scores, only one model marginally exceeds

*Lead Author: `Kirill.Skobelev@chicagobooth.edu`
†Corresponding Authors: `{dan,neeraj.mainkar,margaux}@surgicalvideo.io`, `XY.Han@chicagobooth.edu`

the 13.4% majority class baseline on the validation set. (2) We fine-tune Gemma 3 27B with LoRA adapters to generate structured JSON predictions. The model achieves 47.63% exact match accuracy, surpassing the validation set baseline of 13.41%. (3) We replace off-the-shelf JSON generation with a specialized classification head. This approach achieves 51.08% exact match accuracy. (4) To assess the potential of increasing computational resources, we gradually increase trainable parameters (by increasing LoRA rank) by nearly three orders of magnitude. While training accuracy reaches 98.6%, validation accuracy remains below 40%, showing that scaling alone cannot overcome distribution shift. (5) We compare zero-shot and fine-tuned VLM performance against YOLOv12-m, a specialized 26M-parameter object detection model. YOLOv12-m achieves 54.73% exact match accuracy, outperforming all VLM-based methods while using 1,000× fewer parameters. (6) We demonstrate these findings generalize to CholecT50, an independent and public dataset of laparoscopic cholecystectomy procedures, with additional comparisons to five proprietary frontier VLMs. The fine-tuned open-weight model and YOLOv12-m outperform all zero-shot VLM methods including zero-shot methods using proprietary frontier VLMs.

1 Introduction

The scaling hypothesis has become the dominant paradigm in AI research. Kaplan et al. [2020] documented that cross-entropy loss scales with model size, data, and compute as a power law. Wei et al. [2022] argued that certain capabilities emerge beyond critical model scales, while Chowdhery et al. [2022] demonstrated broad few-shot performance gains and emergent abilities in a 540B-parameter language model. These observations have led to increasingly bold claims: Bubeck et al. [2023] interpret GPT-4's behavior as indicative of emerging AGI, and Aschenbrenner [2024] explicitly argues that continued scaling alone is sufficient to reach AGI.

In medicine, similar optimism has taken hold. Saab et al. [2024] present Med-Gemini, a family of models achieving 91.1% on MedQA and large gains over GPT-4V on multimodal benchmarks, as evidence that large multimodal foundation models can deliver strong generalist capabilities across medical specialties. Such benchmark results have fueled speculation about the feasibility of a "Medical Artificial General Intelligence" (Med-AGI) through scaling. Yet, when tested in realistic clinical settings, the pictures is less optimistic. For example, Hager et al. [2024] find that state-of-the-art LLMs perform significantly worse than physicians across pathologies, often failing to follow instructions. Wu et al. [2025] further demonstrate that "generalist" radiology capability depends on large-scale in-domain pretraining and radiology-specific instruction tuning, suggesting progress toward Med-AGI may be bottlenecked by domain data coverage as much as by parameter count.

In surgery specifically, recently work has begun to apply vision–language models to surgical data across a range of tasks. Surgical-VQA Seenivasan et al. [2022] introduces visual question answering over laparoscopic scenes, while GP-VLS Schmidgall et al. [2024] demonstrates that large foundation models can be adapted to multiple surgical tasks, including instrument recognition, through extensive in-domain supervision. Related efforts fine-tune vision–language models for tool-related tasks such as keypoint estimation using low-rank adaptation, often relying on synthetic datasets to augment limited real annotations Duangprom et al. [2025]. This literature establishes VLMs as a viable modeling paradigm for surgical understanding and motivates their evaluation on fine-grained surgical perception tasks using real operative video.

Despite progress on visual tasks such as surgery, whether these models would lead to Med-AGI is an open question. The definition of AGI remains debated, but, in order to function in the operative setting, locating and classifying surgical instruments is the earliest (necessary, not sufficient) relevant task. Non-expert humans excel at this task: annotators in our study learned to label these tools with near-perfect accuracy after minimal training. In this paper, we evaluate

state-of-the-art AI models for tool detection on SDSC-EEA, a unique dataset of 67,634 annotated frames from neurosurgical videos from the Surgical Data Science Collective (SDSC) [2026].

The paper is organized as follows:

- Section 2 describes the dataset, models, and experimental methodology for five evaluations spanning zero-shot inference, fine-tuning, parameter scaling, specialized vision models, and cross-dataset validation.
- Section 3 presents five findings:
  - Zero-shot VLMs do not surpass a trivial baseline (Section 3.1). Across 19 models spanning 2B–235B parameters and two years of development, validation accuracy remains at or near the majority class baseline of 13.4%.
  - Fine-tuning helps but does not close the gap (Sections 3.2–3.3). LoRA fine-tuning of Gemma 3 27B raises validation exact match accuracy from 9.8% to 51.1%, but generalization to held-out procedures remains limited.
  - Scaling adapter capacity does not resolve generalization (Section 3.4). Increasing trainable parameters by nearly three orders of magnitude drives training accuracy to 98.6% while validation accuracy stays below 40%.
  - A small specialized model outperforms all VLMs (Section 3.5). YOLOv12-m (26M parameters) achieves 54.7% exact match accuracy with 1,000× fewer parameters than the best VLM.
  - These patterns replicate on a public, independent dataset (Section 3.6): CholecT50, a laparoscopic cholecystectomy benchmark. Results, which include comparisons with proprietary frontier VLMs, confirm the broad pattern across surgical domains.
- Section 4 argues that the bottleneck to surgical AI is specialized data, not model scale, and proposes hierarchical architectures where generalist VLMs delegate to specialized perception modules.
- Section 6 discusses limitations.
- Section 7 concludes the paper.

2 Methods

This section describes the dataset and experimental methodology. Section 2.1 introduces the SDSC-EEA dataset. Section 2.2 describes zero-shot VLM evaluation. Section 2.3 describes LoRA fine-tuning of a VLM. Section 2.4 describes a specialized object baseline. Section 2.5 describes an validation on the external CholecT50 dataset. Section 2.6 defines the evaluation metrics used throughout. Corresponding results for each experiment are reported in Section 3.

2.1 SDSC-EEA Dataset

We evaluate surgical tool detection using a dataset of endoscopic endonasal approach (EEA) neurosurgical procedures. EEA is a minimally invasive technique used to access and treat lesions at the skull base through the nasal passages. The dataset is provided by the Surgical Data Science Collective (SDSC) and comprises of 67,634 annotated frames extracted from 66 unique surgical

procedures. Figure 1 exhibits frames from some videos sampled from this dataset. We will refer to it as SDSC-EEA in this paper.

The dataset was constructed from video recordings of surgical procedures donated to the SDSC by 10 surgeons across 7 institutions in the United States, France, and Spain. No exclusion criteria were applied. Ground truth annotations were produced by three annotators from a contracted labeling company, none of whom had clinical experience; annotators were provided with tool descriptions and representative example images prior to labeling. Labels were first reviewed by a senior annotator at the contracting company and subsequently by members of the SDSC. Fewer than 10% of frames required correction.

Each frame is annotated with multi-label ground truth indicating the presence or absence of 31 distinct surgical instrument classes. Annotations are provided in YOLO format with bounding box coordinates. The average number of tools per frame is 1.72 (median: 2), with the distribution showing 7.6% of frames containing no tools, 34.4% containing one tool, 38.2% containing two tools, and 19.8% containing three or more tools.

The tool class distribution exhibits significant imbalance. Suction is the most prevalent instrument, appearing in 63.3% of all frames. Cotton Patty (16.1%), Grasper (10.6%), Curette (8.6%), and Rhoton Dissector (8.0%) follow in frequency.

For all fine-tuning experiments (Section 2.3), we split the data by surgical procedure instances to prevent data leakage. Frames from the same surgical procedure appear exclusively in either the training or validation set, never both. This yields 47,618 training frames from 53 procedures and 20,016 validation frames from 13 procedures.

2.2 Zero-Shot Evaluation of Vision-Language Models

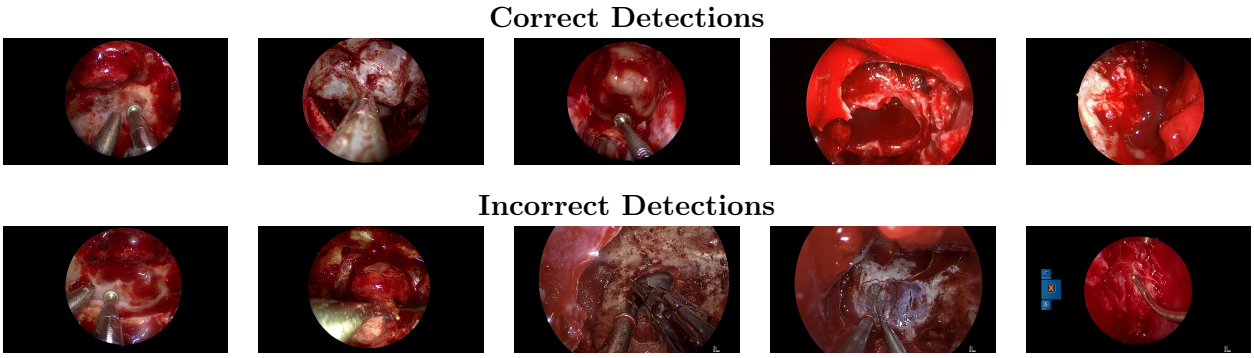

Figure 1: Example frames from SDSC-EEA with zero-shot predictions from Gemma 3 27B. Top row: correct detections (left to right: Drill + Suction; Suction; Drill + Suction; no tools; no tools). Bottom row: incorrect detections, left to right: $y =$ Drill, Suction; $\hat{y} =$ Curette, Grasper, Irrigation, Monopolar Electrocautery, Suction; $y =$ Cotton Patty, Rhoton Dissector, Suction; $\hat{y} =$ Grasper, Monopolar Electrocautery, Suction; $y =$ Bipolar Forceps, Suction; $\hat{y} =$ Curette, Drill, Suction, Tissue shaver; $y =$ Suction; $\hat{y} =$ Grasper, Monopolar Electrocautery, Suction; $y =$ Rhoton Dissector; $\hat{y} =$ Monopolar Electrocautery, Suction.

We evaluate zero-shot tool detection performance across 19 open-weight vision-language models spanning two years of development (September 2023–September 2025). The complete list of models is shown in Table 1.

Table 1: Vision-language models evaluated for zero-shot surgical tool detection.

Model	Params (B)	Release	MMBench
Qwen3-VL-235B-A22B-Thinking [Bai et al., 2025a]	235	Sep 2025	90.6
Qwen3-VL-32B-Instruct	32	Sep 2025	88.9
Qwen3-VL-8B-Instruct	8	Sep 2025	85.0
Qwen3-VL-4B-Instruct	4	Sep 2025	85.1
Qwen3-VL-2B-Instruct	2	Sep 2025	77.8
Qwen2.5-VL-72B-Instruct [Bai et al., 2025b]	72	Mar 2025	87.8
Qwen2.5-VL-32B-Instruct	32	Mar 2025	84.0
Qwen2.5-VL-7B-Instruct	7	Mar 2025	82.2
Qwen2.5-VL-3B-Instruct	3	Mar 2025	76.8
Qwen2-VL-72B-Instruct [Wang et al., 2024]	72	Sep 2024	85.9
Qwen2-VL-7B-Instruct	7	Sep 2024	81.0
Qwen2-VL-2B-Instruct	2	Sep 2024	72.2
Gemma 3 27B-it [Gemma-Team et al., 2025]	27	Mar 2025	78.9
Gemma 3 12B-it	12	Mar 2025	74.6
Gemma 3 4B-it	4	Mar 2025	66.4
MedGemma 3 27B-it [Sellergren et al., 2025]	27	July 2025	-
Llama-3.2-90B-Vision [Meta, 2024]	90	Sep 2024	79.5
Llama-3.2-11B-Vision	11	Sep 2024	67.5
LLaVA-1.5-13B [Liu et al., 2024a]	13	Sep 2023	65.8

Models span five families: Qwen (12 models across three generations), Gemma 3 (3 models), MedGemma 3 (1 model), Llama 3.2 Vision (2 models), and LLaVA 1.5 (1 model). Model sizes range from 2B to 235B parameters. MMBench [Liu et al., 2024b], a holistic benchmark evaluating multimodal models across perception, reasoning, and knowledge, scores range from 65.8 (LLaVA 1.5) to 90.6 (Qwen3-VL-235B).

For each model, we prompt the model to identify all visible surgical tools from a list of 31 valid tool names and return predictions as a JSON object. The complete prompt template is provided in Appendix B. Model outputs are validated against a strict schema; outputs that fail validation (malformed JSON, schema violations, or hallucinated tool names not in the ontology) are treated as empty predictions rather than silently excluded. The full output validation methodology is described in Appendix C.

Table 2 reports exact match accuracy separately on the training set ($n = 47{,}618$ frames from 53 procedures), validation set ($n = 20{,}016$ frames from 13 procedures), and the full dataset. Figure 1 shows representative examples from our dataset, illustrating both successful and unsuccessful tool detection cases.

For the zero-shot results reported in Table 2, Figure 2, and Figure 3, we use exact match accuracy and Jaccard similarity as primary metrics, with per-tool precision, recall, and F1 reported in Appendix L. All evaluation metrics are defined in Section 2.6. These results are analyzed in Section 3.1.

2.3 LoRA Fine-Tuning

We fine-tune Gemma 3 27B using Low-Rank Adaptation (LoRA) [Hu et al., 2021] with adapters applied to attention projection matrices in both the language model and vision encoder. We evaluate

three configurations:

JSON generation (Figure 4, Section 3.2): The model learns to produce structured JSON outputs in the format `{"detected_tools": ["Tool1", "Tool2"]}` via supervised fine-tuning.

Classification head (Figure 5, Section 3.3): We replace JSON generation with a single-layer linear classification head that maps mean-pooled hidden states to 31 output logits, trained with binary cross-entropy loss. At inference, predictions are obtained by thresholding sigmoid outputs at 0.5. This approach enables continuous prediction scores for ROC-AUC and AUPRC metrics and requires only a single forward pass rather than autoregressive generation.

Rank sweep (Figure 6, Table 3, Section 3.4): To investigate whether increasing model capacity improves generalization, we sweep LoRA rank from $r = 2$ to $r = 1024$, varying trainable parameters by nearly three orders of magnitude (4.7M to 2.4B parameters).

All three configurations use the same procedure-level train/validation split described in Section 2.1. Full configuration details (ranks, learning rates, batch sizes, and compute requirements) are provided in Appendix D.

2.4 Specialized Supervised Model

As a supervised baseline, we train YOLOv12-m [Tian et al., 2025], a state-of-the-art object detection model with 26M parameters. Unlike VLMs, which perform set-based multi-label classification, YOLO directly predicts bounding boxes with associated class labels and confidence scores. We train using default YOLO hyperparameters; the full configuration is provided in Appendix H.

To enable direct comparison with VLMs, we convert YOLO's per-frame bounding box predictions into tool sets: for each frame, we collect the unique set of tool classes with confidence ≥ 0.25 and compare against the ground truth tool set. This allows us to compute exact match accuracy, Jaccard similarity, top-1 accuracy, and per-tool precision/recall/F1 on the same basis as VLM-based classifiers. Results, including a per-tool comparison with Gemma (Table 4), are reported in Section 3.5.

2.5 External Dataset: CholecT50

To evaluate generalization to an independent surgical domain, we use CholecT50 [Nwoye et al., 2022], a publicly available dataset of laparoscopic cholecystectomy procedures. CholecT50 comprises 50 videos with frame-level annotations for 6 surgical instruments (grasper, bipolar, hook, scissors, clipper, irrigator), 10 surgical verbs, 15 anatomical targets, and 100 instrument-verb-target triplets. We focus exclusively on instrument detection to maintain consistency with our primary evaluation.

The dataset contains 100,863 annotated frames. We perform an 80/20 train/validation split at the video level to prevent data leakage, yielding 80,940 training frames (40 videos) and 19,923 validation frames (10 videos). The majority class baseline—predicting the most common tool set (grasper, hook) for every frame—achieves 34.76% exact match accuracy on the validation set.

We evaluate zero-shot performance using Gemma 3 27B, fine-tune with LoRA and a classification head using the same configuration as Section 2.3, conduct a LoRA rank sweep ($r \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$) using the same protocol as Section 2.3, and train YOLOv12-m using the same setup as Section 2.4. Results, including Table 6 and Figure 7, are reported in Section 3.6.

2.6 Evaluation Metrics

We report the following metrics throughout. Exact match accuracy is the percentage of frames where the predicted tool set exactly matches the ground truth; this is a strict metric that penalizes any

false positive or false negative. Jaccard similarity is computed for each frame as $J = |P \cap G|/|P \cup G|$ where P is the predicted set and G is the ground truth set, and we report the mean across all frames. We also compute per-tool precision, recall, and F1 scores as standard binary classification metrics independently for each tool class. For models with continuous prediction scores (classification head), we additionally report ROC-AUC (area under the receiver operating characteristic curve) and AUPRC (area under the precision-recall curve) per tool class, as well as macro-averaged values across tools present in the validation set. Per-class accuracy for zero-prevalence classes is meaningless (a model predicting all negatives achieves 100% accuracy) and is excluded from macro-averaged metrics.

To enable direct comparison between YOLO and VLM-based classifiers, we additionally report top-1 accuracy: the fraction of frames where the tool with the highest predicted probability is present in the ground truth set. Both YOLO (via class confidence scores) and the Gemma classifier (via sigmoid outputs) produce explicit per-tool probabilities, making this metric computable for both. However, top-1 accuracy cannot be computed for generative VLM outputs, which produce unordered tool lists without per-tool probability scores. This metric isolates the model's ability to identify the single most salient tool in each frame, a prerequisite for reliable surgical assistance.

For 95% confidence intervals on exact match accuracy, we use bootstrap resampling with $B =$ 1,000 iterations. For a dataset of N frames, we resample N observations with replacement from the binary correct/incorrect results and compute the mean for each of the B bootstrap samples; the 2.5th and 97.5th percentiles form the confidence interval.

3 Results

We present results in five parts. Section 3.1 establishes the baseline: zero-shot VLMs fail to exceed a trivial majority class baseline despite two years of scaling. Given this failure, the next three sections ask whether adaptation can close the gap. Sections 3.2 and 3.3 explore two parallel fine-tuning strategies—JSON generation and a classification head—that both improve substantially over zero-shot but plateau well below human-level accuracy. Section 3.4 then tests whether this plateau is due to insufficient capacity by scaling LoRA rank by nearly three orders of magnitude; training accuracy saturates near 99% while validation accuracy remains below 40%, indicating that the bottleneck is not model capacity. Section 3.5 compares against YOLOv12-m, a specialized 26M-parameter object detection model that outperforms all VLM-based approaches with 1,000× fewer parameters. Section 3.6 replicates the key experiments on CholecT50, a laparoscopic cholecystectomy dataset, and finds the same broad patterns across both surgical domains.

3.1 Zero-shot accuracy of open-weight models does not surpass the majority class baseline

Takeaways Even for larger VLMs, in the zero-shot setting, performance stays at or near the majority-class baseline. Progress on general multimodal benchmarks and parameter scale does not transfer reliably to this surgical perception task.

Detailed Results. We evaluate zero-shot tool detection performance across 19 open-weight vision-language models (Section 2.2) released between September 2023 and September 2025. Despite dramatic increases in model scale, from LLaVA 1.5 13B (2023) to Qwen3-VL-235B (2025), and substantial improvements on general vision benchmarks, no model meaningfully surpasses the majority class baseline on the validation set.

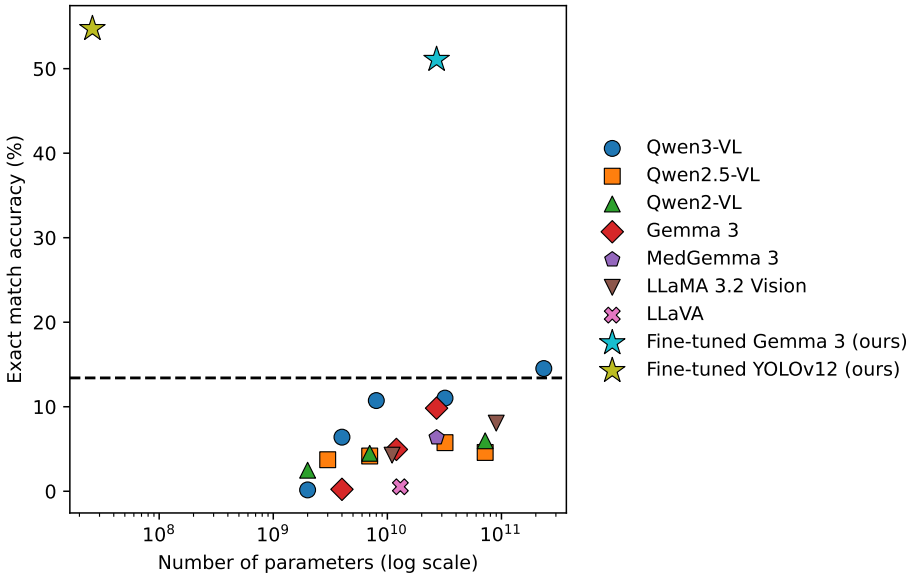

Figure 2: Exact-match accuracy on the SDSC-EEA validation set (n = 20,016) as a function of model parameter count. Colors and marker shapes denote model families. The black dashed line indicates the majority-class baseline (13.4%). Accuracy exhibits a positive but strongly sublinear relationship with parameter count; the relationship is family-dependent, with Qwen models consistently outperforming similarly-sized Gemma and Llama models.

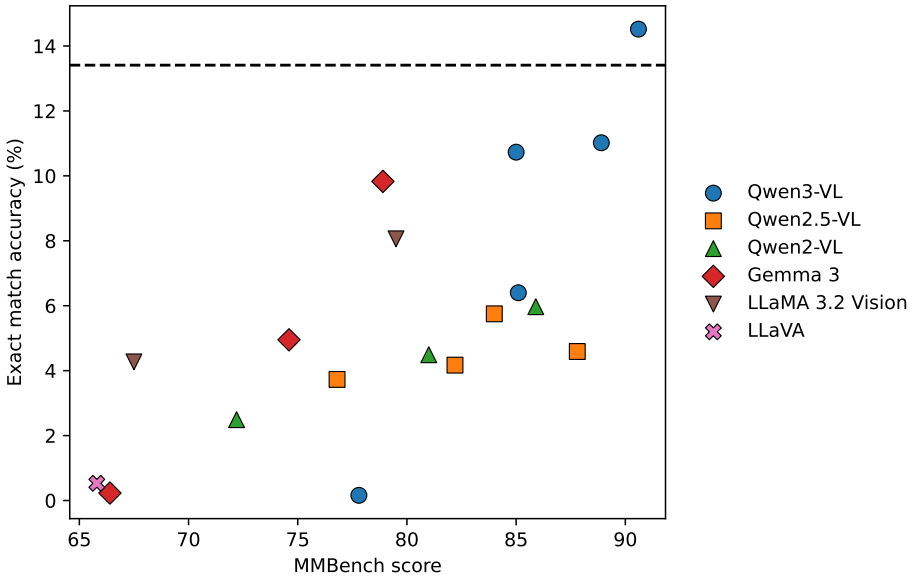

Figure 3: Zero-shot exact-match accuracy on the SDSC-EEA validation set (n = 20,016) plotted against MMBench score. Colors and marker shapes denote model families. The black dashed line indicates the majority-class baseline (13.4%). Higher MMBench scores correlate with higher tool detection accuracy, but even the best model (Qwen3-VL-235B, MMBench 90.6) achieves only 14.52%—far below fine-tuned models (51.08%, Section 3.3).

Table 2 reports exact match accuracy for all models; no model meaningfully surpasses the majority class baseline. As shown in Figure 3, higher MMBench scores are correlated with higher performance on the tool detection benchmark in our dataset, and the relationship appears to be linear. However, even the best performing model, Qwen3-VL-235B, which achieves a 90.6 out of 100 score on MMBench, significantly underperforms the fine-tuned Gemma 3 27B in Section 3.3 (14.52% vs 51.08% validation exact match accuracy). This suggests that there are surgical visual capabilities that go beyond what can be measured by multi-purpose benchmarks like MMBench.

Notably, MedGemma 3 27B-it, which is described as a model optimized for medicine, underperforms Gemma 3 27B-it—a sibling that MedGemma is based on—on the validation set (6.36% vs. 9.83%). Per-tool classification metrics (precision, recall, F1) for all 19 evaluated zero-shot models are provided in Appendix L. Appendix G shows representative failed outputs, which are dominated by hallucinated tool names rather than formatting errors.

3.2 LoRA fine-tuning improves tool detection modestly but remains below human-level

Takeaways Task-specific fine-tuning improves performance relative to zero-shot evaluation, but it does not close the generalization gap on held-out procedures.

Detailed Results Given that zero-shot models fail at surgical tool detection regardless of scale, we next ask whether task-specific fine-tuning can bridge the gap. We fine-tune Gemma 3 27B with LoRA adapters to generate structured JSON predictions (Section 2.3). Figure 4 shows training and validation loss curves over 10 epochs.

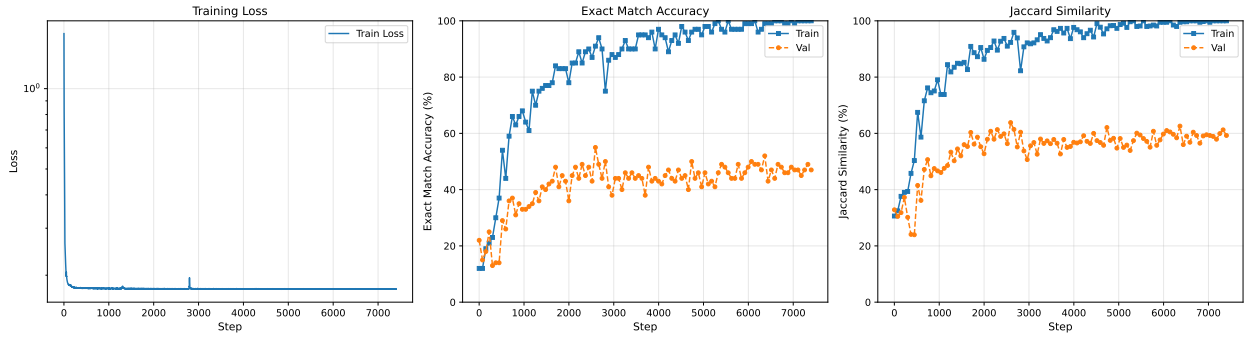

Figure 4: Training dynamics for LoRA fine-tuning with JSON output on SDSC-EEA ($r = 1024$). Left: Training loss (log scale) decreases steadily, confirming the model learns the structured output format. Center: Exact match accuracy. Right: Jaccard similarity. Both accuracy and Jaccard show a persistent gap between training and validation performance, indicating limited generalization to held-out procedures. Metrics are computed on fixed random subsets of 100 frames from each set, evaluated 100 times throughout training.

After 10 epochs, the fine-tuned model achieves 47.63% exact match accuracy (95% CI: 46.97%–48.34%) and 57.34% Jaccard similarity on the validation set ($n = 20{,}016$). This represents a substantial improvement over both the majority class baseline (13.41% exact match, 31.91% Jaccard) and the pre-training baseline (9.83% exact match, 25.98% Jaccard).

Table 2: Zero-shot tool detection exact match accuracy (%) on SDSC-EEA for all evaluated VLMs with 95% bootstrap confidence intervals. Train (n = 47,618 frames, 53 procedures), validation (n = 20,016 frames, 13 procedures), and full dataset. Output validation failures are counted as incorrect predictions. The majority class baseline, which predicts the most common tool set for every frame, achieves 13.41% exact match accuracy on the validation set. The pre-training baseline (Gemma 3 27B) achieves 9.83% validation exact match accuracy (95% CI: 9.43%–10.21%). Validation accuracy ranges from 0.11% (Qwen3-VL-2B) to 14.52% (Qwen3-VL-235B-A22B-Thinking); only Qwen3-VL-235B marginally surpasses the majority class baseline. Output validation failure rates range from 0.8% (Gemma 3 27B, Qwen3-VL-8B) to 41.7% (Qwen2-VL-2B).

	Params	Train		Validation		Full	
Model	(B)	EM %	95% CI	EM %	95% CI	EM %	95% CI
Qwen3-VL-235B-A22B-Thinking	235	17.04	16.69–17.39	14.52	14.06–15.02	16.28	16.01–16.55
Qwen3-VL-32B-Instruct	32	16.02	15.70–16.34	11.04	10.57–11.45	14.58	14.32–14.84
Qwen3-VL-8B-Instruct	8	12.08	11.80–12.38	10.73	10.32–11.13	11.68	11.46–11.92
Qwen3-VL-4B-Instruct	4	11.10	10.82–11.38	6.37	6.04–6.70	9.72	9.51–9.94
Qwen3-VL-2B-Instruct	2	0.29	0.24–0.34	0.11	0.07–0.16	0.24	0.20–0.27
Qwen2.5-VL-72B-Instruct	72	10.27	10.01–10.54	4.56	4.27–4.87	8.59	8.40–8.79
Qwen2.5-VL-32B-Instruct	32	10.08	9.83–10.36	5.43	5.13–5.74	8.72	8.52–8.93
Qwen2.5-VL-7B-Instruct	7	6.54	6.33–6.75	3.48	3.25–3.75	5.64	5.49–5.82
Qwen2.5-VL-3B-Instruct	3	5.43	5.23–5.64	2.90	2.68–3.14	4.69	4.53–4.86
Qwen2-VL-72B-Instruct	72	8.70	8.46–8.95	5.70	5.39–6.00	7.82	7.64–8.01
Qwen2-VL-7B-Instruct	7	5.13	4.93–5.32	3.35	3.09–3.60	4.61	4.46–4.76
Qwen2-VL-2B-Instruct	2	2.29	2.15–2.42	1.48	1.31–1.64	2.05	1.94–2.15
Gemma 3 27B-it	27	5.61	5.40–5.81	9.83	9.43–10.21	6.85	6.66–7.02
Gemma 3 12B-it	12	5.20	5.01–5.39	4.95	4.65–5.25	5.13	4.97–5.29
Gemma 3 4B-it	4	0.56	0.50–0.63	0.23	0.16–0.30	0.46	0.42–0.51
MedGemma 3 27B-it	27	5.41	5.20–5.60	6.36	6.02–6.70	5.68	5.51–5.86
Llama-3.2-90B-Vision	90	10.62	10.33–10.91	8.06	7.68–8.45	9.85	9.62–10.09
Llama-3.2-11B-Vision	11	0.33	0.28–0.38	0.19	0.13–0.25	0.29	0.25–0.33
LLaVA-1.5-13B	13	1.66	1.55–1.77	0.53	0.44–0.64	1.33	1.23–1.41
Majority class baseline	–	11.49	–	13.41	–	12.06	–

Table 8 in Appendix E shows per-tool precision and recall. The model learns to detect several tools with high F1 scores (Suction Coagulator: 0.989, Drill: 0.876, Suction: 0.809) but completely fails on others (Suction microdebrider: 0% recall despite 497 ground truth instances in validation). This discrepancy arises from the procedure-based train/validation split: tools that appear predominantly in validation procedures were rarely seen during training (Table 7). For example, Suction microdebrider has only 94 training instances versus 497 in validation, and Aspirating dissector has 88 training instances versus 2,319 in validation.

Qualitative analysis reveals that fine-tuned models produce syntactically correct JSON outputs with valid tool names (eliminating output validation failures common in zero-shot outputs), but generalization to unseen tool distributions remains poor.

3.3 LoRA with classification head learns in-sample but fails to generalize out-of-sample

Takeaways Dedicated classification objectives are more effective than autoregressive JSON generation for surgical tool detection, yielding the strongest VLM-based performance in our study. The train-validation gap remains.

Detailed Results Having established that LoRA fine-tuning with JSON generation improves over zero-shot baselines, we test whether a dedicated classification objective can do better. We replace JSON generation with a linear classification head trained with binary cross-entropy loss (Section 2.3). The classification head produces per-tool probability scores, enabling threshold-independent metrics such as ROC-AUC and AUPRC that are not available from discrete JSON outputs. Figure 5 shows training dynamics.

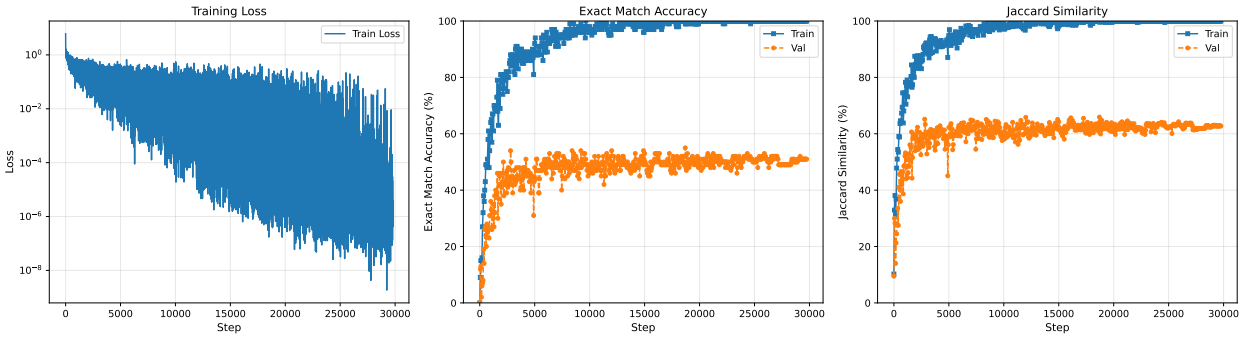

Figure 5: Training dynamics for LoRA fine-tuning with classification head on SDSC-EEA ($r = 1024$). Left: Training loss (log scale). Center: Exact match accuracy. Right: Jaccard similarity. The classification head achieves the highest validation accuracy among all VLM-based methods (51.08%), outperforming JSON generation at the same LoRA rank (47.63%, Figure 4). The persistent train–validation gap reflects limited generalization to held-out procedures. Metrics are computed on fixed random subsets of 100 frames from each set, approximately 100 times throughout training.

After 10 epochs, the fine-tuned model achieves 51.08% exact match accuracy (95% CI: 50.39%–51.81%) and 61.33% Jaccard similarity on the validation set ($n = 20{,}016$), substantially outperforming both the majority class baseline (13.41%) and the pre-training baseline (9.83%). The model also achieves 80.5% macro-averaged ROC-AUC and 37.6% macro-averaged AUPRC across the 23 tool classes present in the validation set. Table 10 in Appendix F shows per-tool ROC-AUC and AUPRC.

This approach achieves the highest validation accuracy among all VLM-based methods, outperforming JSON generation with the same LoRA rank (Section 3.2, 47.63% exact match), suggesting that explicit multi-label classification objectives are more effective than autoregressive generation for this task.

3.4 Scaling LoRA adapter rank does not meaningfully improve out-of-sample accuracy

Takeaways Increasing LoRA rank improves training accuracy but produces only limited gains on held-out procedures. This suggests that the main bottleneck is not insufficient adapter capacity or compute, but failure to generalize under distribution shift.

Detailed Results The experiments in Sections 3.2-3.3 use a single, large LoRA rank ($r = 1024$). A natural question is whether the validation accuracy gap reflects insufficient model capacity. We sweep LoRA rank from 2 to 1,024, increasing trainable parameters by nearly three orders of magnitude (Table 3; methodology in Section 2.3). Figure 6 shows accuracy as a function of rank.

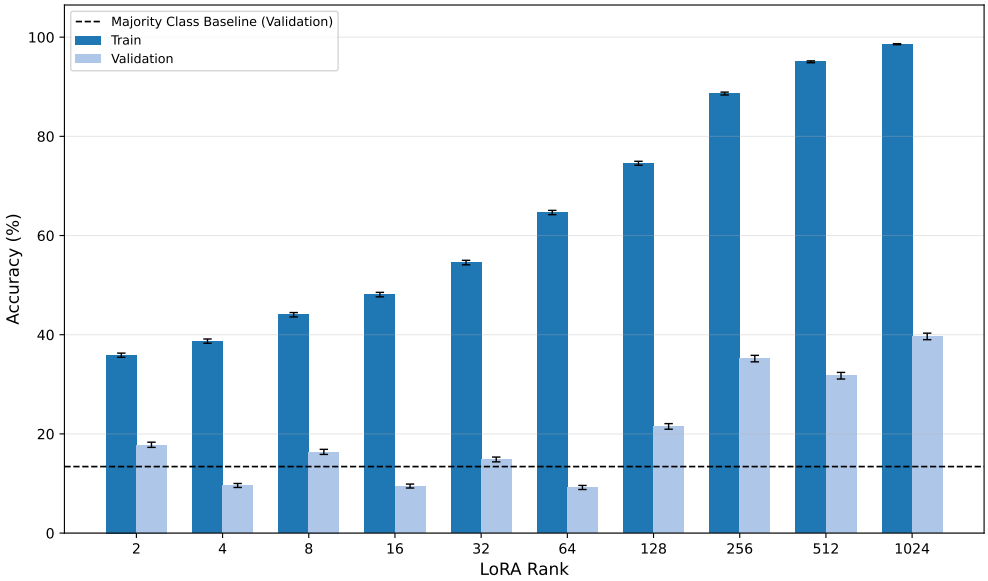

Figure 6: Exact match accuracy vs. LoRA rank on SDSC-EEA. Gemma 3 27B with LoRA adapters and a linear classification head, trained for 3 epochs at each rank ($r \in \{2, 4, \ldots, 1024\}$). Training accuracy (dark blue) increases monotonically from 35.9% to 98.6%, while validation accuracy (light blue) remains below 40% across all ranks. The widening gap demonstrates that scaling adapter capacity alone cannot overcome the procedure-level distribution shift. Error bars: 95% bootstrap CIs ($B = 1{,}000$).

Table 3: LoRA rank sweep on SDSC-EEA: configurations with trainable parameter counts for each rank.

Rank	Params (M)
2	4.7
4	9.4
8	18.8
16	37.5
32	75.0
64	150.0
128	300.0
256	600.0
512	1200.1
1024	2400.2

Training accuracy increases monotonically with rank, from 35.9% (95% CI: 35.5%–36.3%) at $r = 2$ to 98.6% (95% CI: 98.5%–98.7%) at $r = 1024$. Validation accuracy peaks at $r = 1024$ with 39.6% (95% CI: 39.0%–40.3%), though the relationship is non-monotonic.

3.5 Specialized supervised model

Takeaways For this narrow perceptual task, a small specialized vision model outperforms all VLM-based approaches while using orders of magnitude fewer parameters. This suggests the surgical AI performance is currently limited less by larger generalist models than by the availability of task-specific data.

Detailed Results The VLM-based approaches above require a 27B-parameter model, with corresponding training time and inference cost. A natural question is whether a small, specialized model can match that performance at a fraction of the cost. YOLOv12-m (Section 2.4), with only 26M parameters—over 1,000× fewer—achieves 54.73% exact match accuracy (95% CI: 54.03%–55.44%) and 64.00% Jaccard similarity (95% CI: 63.37%–64.58%) on the validation set (n = 20,016), with 70.06% top-1 accuracy (95% CI: 69.43%–70.70%), outperforming the best VLM (Gemma 3 27B with classification head, $r = 1024$; 51.08% exact match, 61.33% Jaccard). We select YOLO as a natural baseline for this task given its established success in object detection. However, YOLO is trained with bounding box supervision, while VLMs receive only set-level labels. To verify that YOLO's advantage is not solely due to this richer supervisory signal, we train a ResNet-50 (23.6M parameters) using the same set-level labels as VLMs—without any bounding box information. This CNN achieves 39.6% exact match accuracy (95% CI: 38.9%–40.3%), outperforming all zero-shot VLMs (Appendix J).

Table 4 compares per-tool metrics between the two models. ROC-AUC is computed from Gemma's sigmoid outputs and YOLO's maximum detection confidence per class.

Table 4: Per-tool comparison: YOLOv12-m vs. Gemma 3 27B (classification head) on the SDSC-EEA validation set ($n = 20{,}016$). YOLO achieves higher F1 and recall on all 15 tools, while Gemma achieves higher precision on 4 tools. For ROC-AUC, the two models are complementary: YOLO leads on 7 tools, Gemma on 8. Sorted by ground truth count (N). Best per row in bold.

		Precision		Recall		F1		ROC-AUC	
Tool	N	YOLO	Gemma	YOLO	Gemma	YOLO	Gemma	YOLO	Gemma
Suction	10685	**.732**	.673	**.963**	.885	**.832**	.764	**.875**	.819
Rongeur	2790	.948	**.960**	**.716**	.222	**.816**	.361	.866	**.920**
Cotton Patty	2143	**.877**	.576	**.819**	.706	**.847**	.635	**.981**	.929
Drill	2116	.943	**.945**	**.959**	.790	**.951**	.861	**.984**	.983
Rhoton Dissector	1462	**.554**	.540	**.882**	.590	**.680**	.564	**.945**	.920
Surgical Knife	1422	**.947**	.920	**.904**	.049	**.925**	.092	**.953**	.939
Suction Coagulator	1188	.982	**1.00**	**.995**	.622	**.988**	.767	.998	**1.00**
Backbiting rongeur	1041	.852	.741	**.243**	.019	**.378**	.038	.641	**.918**
Scissor	996	.458	**.622**	**.673**	.136	**.545**	.223	**.840**	.766
Surgicel	739	**1.00**	.971	**.635**	.628	**.776**	.763	.825	**.908**
Curette	708	.842	**.949**	**.468**	.239	**.601**	.382	.744	**.940**
Grasper	509	**.289**	.032	**.686**	.020	**.406**	.024	**.840**	.708
Bipolar Forceps	263	**.752**	.000	**.346**	.000	**.474**	.000	.675	**.723**
Straight Forceps	173	.454	**.642**	**.775**	.249	**.573**	.358	.908	**.943**
Irrigation	112	.809	**1.00**	**.339**	.018	**.478**	.035	.701	**.776**

3.6 Robustness: Performance on CholecT50

Takeaways The results on SDSC-EEA reproduce on CholecT50: the broad pattern that zero-shot performance is poor, that fine-tuning is necessary, and that smaller models outperform VLMs at a fraction of the size. Additionally, frontier proprietary models from the GPT, Gemini, and Claude families underperform a fine-tuned open-weight LLM and a specialised computer vision model.

Detailed Results To assess whether our findings generalize beyond neurosurgery, we evaluate on CholecT50, an independent laparoscopic cholecystectomy dataset with 6 instrument classes (Section 2.5). Zero-shot Gemma 3 27B achieves 6.87% exact match accuracy (95% CI: 6.55%–7.22%), which is below the majority class baseline (34.76%).

Fine-tuning Gemma 3 27B with LoRA ($r = 128$) and a classification head reaches 83.02% exact match accuracy (95% CI: 82.52%–83.56%) and 88.79% Jaccard similarity (95% CI: 88.43%–89.18%). YOLOv12-m achieves 81.37% exact match accuracy (95% CI: 80.87%–81.92%) and 88.00% Jaccard similarity (95% CI: 87.62%–88.34%), with 93.80% top-1 accuracy (95% CI: 93.45%–94.12%). Table 6 compares per-tool metrics between the two fine-tuned models.

Per-tool metrics for the zero-shot setting are in Appendix I: grasper achieves the highest F1 (0.627), while bipolar has 12,096 false positives vs. 838 true positives.

Additionally, since CholecT50 is a public dataset, we can also evaluate the performance of five closed-weight frontier VLMs from the GPT 5.4 [OpenAI, 2026], Gemini 3 [Gemini Team, 2026], and Claude 4.6 [Anthropic, 2026] families[1] using the same prompt template and validation frames. Table 5 compares all models evaluated on CholecT50.

[1] We cannot evaluate SDSC-EEA on proprietary frontier models because we cannot send private SDSC-EEA data to third-party APIs. All evaluations on SDSC-EEA were conducted locally on open-source VLMs.

Table 5: Tool detection exact match accuracy (%) on the CholecT50 validation set (n = 19,923 frames, 6 instrument classes) with 95% bootstrap confidence intervals (B = 1,000). Models are sorted by exact match accuracy. The majority class baseline predicts the most common tool set for every frame. Output validation failures are counted as incorrect predictions.

Model	Params (B)	EM %	95% CI
Gemma 3 27B + LoRA cls. head	27	**83.02**	**82.52–83.56**
YOLOv12-m	0.026	81.37	80.87–81.92
Gemini 3 Flash Preview	—	69.15	68.49–69.73
Gemini 3.1 Pro Preview	—	66.21	65.58–66.88
Claude Opus 4.6	—	52.37	51.67–53.03
GPT-5.4	—	32.09	31.40–32.72
Claude Sonnet 4.6	—	30.73	30.07–31.37
Gemma 3 27B-it	27	6.87	6.55–7.22
Majority class baseline	—	34.76	—

Two closed-weight frontier models—Gemini 3 Flash Preview and Gemini 3.1 Pro Preview—achieve 69.15% and 66.21% zero-shot exact match accuracy respectively, far surpassing the open-weight Gemma 3 27B (6.87%) and approaching the fine-tuned models. Second, performance varies dramatically across model families: both Google models exceed 66%, Claude Opus reaches 52%, while GPT-5.4 and Claude Sonnet fall below the majority class baseline. Third, the fine-tuned models (83% and 81%) still outperform even the best zero-shot API model by 14 percentage points, confirming that task-specific training remains valuable even as frontier models improve.

Table 6: Per-tool comparison: YOLOv12-m vs. Gemma 3 27B (classification head) on CholecT50 validation set (n = 19,923). Unlike SDSC-EEA (Table 4), results are mixed: Gemma achieves higher F1 on 4 of 6 tools; YOLO leads on irrigator and bipolar. Gemma has higher precision on all 6 tools; YOLO has higher recall on 5 of 6. Sorted by Gemma F1. Best per row in bold.

	Precision		Recall		F1		ROC-AUC	
Tool	YOLO	Gemma	YOLO	Gemma	YOLO	Gemma	YOLO	Gemma
hook	.953	**.972**	**.986**	.977	.969	**.974**	**.992**	.989
clipper	.902	**.936**	**.910**	.895	.906	**.915**	.967	**.989**
grasper	.860	**.899**	**.953**	.927	.904	**.913**	**.958**	.941
irrigator	.910	**.929**	**.819**	.703	**.862**	.800	.955	**.969**
bipolar	**.920**	.944	**.776**	.743	**.842**	.831	.920	**.959**
scissors	.884	**.888**	.578	**.599**	.699	**.715**	.825	**.947**

We additionally sweep LoRA rank from 2 to 1,024 on CholecT50 using the same protocol as Section 3.4. Figure 7 shows accuracy as a function of rank. Training accuracy increases from 67.1% (95% CI: 66.8%–67.5%) at $r = 2$ to 95.9% (95% CI: 95.8%–96.1%) at $r = 1024$. Validation accuracy increases from 64.7% (95% CI: 64.0%–65.4%) at $r = 2$ to 85.1% (95% CI: 84.5%–85.6%) at $r = 1024$. Unlike SDSC-EEA, where validation accuracy exhibits a non-monotonic relationship with rank and remains below 40% even at $r = 1024$, CholecT50 validation accuracy increases monotonically across all tested ranks and reaches 85.1% at $r = 1024$.

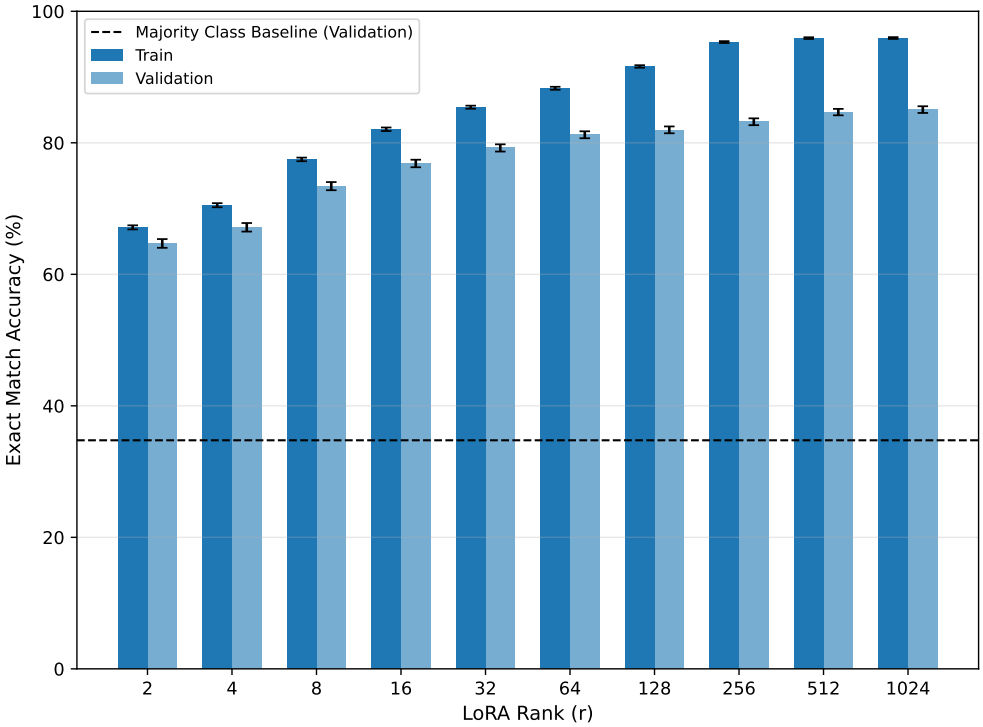

Figure 7: Exact match accuracy vs. LoRA rank on CholecT50. Unlike SDSC-EEA (Figure 6), validation accuracy increases monotonically across all ranks, reaching 85.1% at $r = 1024$, with a much smaller train–validation gap. The lower tool diversity (6 vs. 31 classes) and more uniform video-level distribution make CholecT50 more amenable to LoRA fine-tuning. Dashed line: majority class baseline (34.76%). Error bars: 95% bootstrap CIs ($B = 1{,}000$).

4 Discussion

Despite the growing enthusiasm for some form of Med-AGI, our results on surgical tool detection demonstrate that significant obstacles still exist when training generally-capable AI architectures to perform specialized medical tasks. As shown in Section 3.1 (Table 2, Figures 2–3), 19 open-weight VLMs spanning 2B–235B parameters fail to surpass a trivial majority class baseline on surgical tool detection, despite steady gains on general benchmarks. Fine-tuning closes part of the gap—Section 3.3 achieves 51.08% exact match accuracy with a classification head—but a persistent train–validation gap (Figure 5) and the rank sweep in Section 3.4 (Figure 6) confirm that scaling adapter capacity alone does not resolve the generalization bottleneck. Thus, our results suggest advancing surgical AI may require more task-specialized approaches and—contrary to conventional wisdom—may not be directly solvable by the "scaling law" approach of increasing computation or architecture size [Hestness et al., 2017, Kaplan et al., 2020, Ho et al., 2025].

Meanwhile, Section 3.5 shows that YOLOv12-m, a 26M-parameter model, outperforms all VLM-based approaches (Tables 4), and Section 3.6 reproduces the same pattern (including on proprietary frontier VLMs) on CholecT50 (Tables 5-6, Figure 7), a dataset that is available in the public domain and may have been included in the training of the underlying VLM. Taken together, these results show that specialized computer vision models match or outperform VLMs at one thousandth of the cost measured in the number of parameters, which is directly proportional to compute and latency. The efficiency and superiority of these specialized models indicate that the next critical

advances in Surgical AI will most likely arise from a focused large-scale community effort to defragment data, achieve consensus and labeling at scale, and make data and labels available in an administrative/operational context that would facilitate better training of such specialized models.

Our results also suggest that one way to reconcile generality with performance is to treat the VLM as an orchestrator that can fit or select specialized perception modules on demand. The complementary strengths visible in Table 4—where YOLO leads on recall and F1 across all tools while Gemma leads on ROC-AUC for 8 of 15 tools—suggest that hybrid systems combining both model types could outperform either alone. A promising direction for future research is exploring the best approaches for build and improving such hierarchical systems, with the generalist model delegating to specialized models for high-precision subtasks.

Need for Community-Driven Progress: Perspectives from the SDSC. Some authors of this paper are members of the Surgical Data Science Collective (SDSC), a nonprofit dedicated to advancing open, collaborative, and clinically-grounded approaches to surgical AI. Our experiences in developing collaborative AI tools for surgeons suggest that assembling large-scale data, ontologies, and labels is a critical prerequisite step for building useful clinical tools. This involves building curated datasets, domain specific innovations, and efficient annotation frameworks. Moreover, such an effort must be supported by a community- and consensus-focused effort, led by coalitions of aligned organizations. By fostering multi-institutional collaboration, standardizing data-sharing protocols, and developing open access tools, SDSC and similar organizations seeks to rapidly advance surgical AI to achieve clinical relevance, equitable access, and real world impact.

The SDSC believes that surgical AI is constrained less by model scale than by the availability of clinically relevant data. Despite advances in foundation models, both the SDSC's practical experience and the results in this paper indicate that performance on basic perceptual tasks remains limited under realistic distribution shift—as evidenced by the widening train–validation gap across LoRA ranks (Section 3.4, Figure 6) and the uneven per-tool recall driven by procedure-level tool imbalance (Section 3.2, Table 7). This indicates that significant gaps remain in domain-specific data coverage and suggests that improvements will depend on the development of large-scale, standardized surgical datasets that capture variability across procedures, institutions, and patient populations. In this setting, the SDSC and similar organizations can play an important role in enabling cross-institutional data aggregation and establishing shared standards.

In addition to the need for specialized data, this paper's findings indicate that the most promising path forward may not be pushing towards more powerful end-to-end AI models, but rather developing hybrid systems in which generalist models are complemented by specialized components, consistent with the strong performance of smaller task-specific models (Sections 3.5 and 3.6). Accordingly, the research and development of such hybrid models has been a key focus at the SDSC [Masson-Forsythe et al., 2024, Cook et al., 2025a,b].

5 Exploratory Next Steps: Natural Language vs. Operating Room

As an exploratory next step, in Appendix N, we conduct a preliminary experiment where LLMs appear to give almost entirely correct responses to questions relating to pituitary tumor surgery. However, as we have demonstrated earlier, the same models fail at a simple task of tool detection. This result is not surprising: consider that neurosurgeons train primarily through practice. The Accreditation Council for Graduate Medical Education (ACGME) requires 7 years of residency, typically completed after 2 years of rotations in medical school, compared to only 2 years of classroom and anatomy lab education ACGME [2025]. The idea that tacit knowledge in such jobs is more

important than what can be written down is not new. In the context of job market automation, this is often referred to as Polanyi's paradox Autor [2014]. The challenge with medical AI is that the data for pretraining foundation models lacks information from the operating room. So, while it is conceivable that scaling foundation models could solve Med-AGI without domain-specific data, the evidence so far does not support this idea.

6 Limitations

This study has several limitations. First, our evaluation is restricted to surgical tool detection: Tool detection is a prerequisite for any system aspiring to general surgical intelligence, but we do not evaluate higher-order capabilities such as phase recognition, decision support, or anomaly detection. It is possible that VLMs offer greater advantages on these more abstract tasks, where language-mediated reasoning plays a larger role. Second, our VLM evaluation focuses on open-weight models with a specific prompting and decoding setup. Stronger closed-source models, alternative prompting strategies, or more extensive instruction tuning could yield different results. Third, the degree to which our conclusions generalize to other surgical specialties, institutions, and recording conditions remains an open question, although the consistency of the takeaways on CholecT50 with those that we found on our own data suggests the broad pattern holds across at least two distinct surgical domains. Finally, while we did our best to conduct scaling experiments within our computational means, it remains possible that future models may show non-linear "emergent" jumps in performance if model size and training duration scale past a yet-to-be-discovered threshold [Wei et al., 2022].

7 Conclusion

In this paper, we evaluate how much recent progress in large vision-language models and scaling can bring us closer to the popular notion of Med-AGI by using surgical tool detection in endoscopic endonasal neurosurgery as a case study. Section 3.1 shows that across 19 open-weight VLMs spanning 2B to 235B parameters, zero-shot performance on held-out procedures remains at or near a trivial majority class baseline, despite large gains on general benchmarks like MMBench. Section 3.3 demonstrates that LoRA fine-tuning with a classification head improves performance substantially, reaching 51.08% exact match accuracy, but a persistent train–validation gap reflects limited generalization under procedure-level distribution shift. Section 3.4 rules out insufficient capacity as the cause: scaling LoRA rank by nearly three orders of magnitude saturates training accuracy near 99% while validation accuracy remains below 40%. Section 3.5 shows that YOLOv12-m, a 26M-parameter object detection model—over 1,000× smaller than the VLM—outperforms all VLM-based approaches at a fraction of the training time and inference cost. Section 3.6 replicates these findings (including comparisons with proprietary frontier VLMs) on CholecT50, a public laparoscopic cholecystectomy dataset, confirming that the same patterns hold across surgical domains.

Our findings suggest that progress toward reliable surgical AI is likely more constrained by limited amounts of specialized data than by the scale of potential AI architectures and training resources. Small specialized models can outperform large foundation models on narrow surgical tasks while being orders of magnitude more efficient. Thus, future efforts to pool and label surgical data across institutions will be crucial to improving Surgical AI capabilities.

Funding and Support. This project is jointly funded by the Booth School of Business at UChicago, the Center for Applied AI at Chicago Booth, and the Surgical Data Science Collective (SDSC). Collaborative data sharing between Chicago Booth and the SDSC was facilitated by the Tolan Center for Healthcare at Chicago Booth and the SDSC Engineering Team. Computational experiments were conducted on the Pythia Supercomputer Cluster at Chicago Booth.

References

ACGME. Acgme program requirements for graduate medical education in neurological surgery, 2025. URL https://www.acgme.org/globalassets/pfassets/programrequirements/2025-reformatted-requirements/160_neurologicalsurgery_2025_reformatted.pdf.

Anthropic. Introducing Claude Opus 4.6, February 2026. URL https://www.anthropic.com/news/claude-opus-4-6.

Leopold Aschenbrenner. Situational awareness: The decade ahead, 2024. URL https://situational-awareness.ai.

David Autor. Polanyi's paradox and the shape of employment growth. Working Paper 20485, National Bureau of Economic Research, September 2014. URL http://www.nber.org/papers/w20485.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL https://arxiv.org/abs/2303.12712.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant

Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Jack Cook, Jonathan Chainey, Ruth Lau, Margaux Masson-Forsythe, Ayesha Syeda, Kaan Duman, Daniel Donoho, Dhiraj A Pangal, and Juan Vivanco Suarez. Enhancing surgical video phase recognition with advanced ai models for endoscopic pituitary tumor surgery. *Journal of Neurological Surgery Part B: Skull Base*, 86(S 01):S335, 2025a.

Jack Cook, Ayesha Syeda, Margaux Masson-Forsythe, Dhiraj Pangal, and Daniel Donoho. 1255 enhancing surgical computer vision: A real-time monitoring system for model performance and data quality. *Neurosurgery*, 71(Supplement_1):202, 2025b.

Krit Duangprom, Tryphon Lambrou, and Binod Bhattarai. Estimating 2d keypoints of surgical tools using vision-language models with low-rank adaptation, 2025. URL https://arxiv.org/abs/2508.20830.

Gemini Team. Gemini 3.1 pro: A smarter model for your most complex tasks, February 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Gemma-Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti,

Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024. doi: 10.1038/s41591-024-03097-1. URL https://doi.org/10.1038/s41591-024-03097-1.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. A rosetta stone for ai benchmarks. *arXiv preprint arXiv:2512.00193*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. Pytorch distributed: Experiences on accelerating data parallel training, 2020. URL https://arxiv.org/abs/2006.15704.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024b. URL https://arxiv.org/abs/2307.06281.

Margaux Masson-Forsythe, Juan Vivanco Suarez, Muhammad Ammar Haider, James K Liu, and Daniel A Donoho. Ai-based surgical tools detection from endoscopic endonasal pituitary videos. *Journal of Neurological Surgery Part B: Skull Base*, 85(S 01):S224, 2024.

Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, September 2024. Accessed: 2025-12-30.

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.

OpenAI. Introducing GPT-5.4, March 2026. URL https://openai.com/index/introducing-gpt-5-4/.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine, 2024. URL https://arxiv.org/abs/2404.18416.

Samuel Schmidgall, Joseph Cho, Cyril Zakka, and William Hiesinger. Gp-vls: A general-purpose vision language model for surgery, 2024. URL https://arxiv.org/abs/2407.19305.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer, 2022. URL https://arxiv.org/abs/2206.11053.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao,

Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Surgical Data Science Collective (SDSC). Surgical data science collective: From surgical data to clinical discovery, 2026. URL https://www.surgicalvideo.io. Accessed: 2026-03-05.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors, 2025. URL https://arxiv.org/abs/2502.12524.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025. doi: 10.1038/s41467-025-62385-7. URL https://doi.org/10.1038/s41467-025-62385-7.

A Tool Distribution Across Train/Validation Splits

Table 7 shows the number of frames containing each tool in the training set (n = 47,618 frames, 53 procedures) and validation set (n = 20,016 frames, 13 procedures). Because the split is performed at the procedure level, the per-tool distribution across splits is highly uneven. Several tools appear almost exclusively in one split: for example, Aspirating dissector has 88 training instances versus 2,319 validation instances, and Sonopet pineapple tip has 1,991 training instances versus zero in validation.

Table 7: Number of frames containing each tool in the training (n = 47,618) and validation (n = 20,016) splits, with percentage of frames in each split. Sorted by total count in descending order.

	Train		Validation		
Tool	*N*	%	*N*	%	Total
Suction	29,692	62.4	10,685	53.4	40,377
Cotton Patty	8,822	18.5	2,143	10.7	10,965
Grasper	6,812	14.3	509	2.5	7,321
Rhoton Dissector	4,249	8.9	1,462	7.3	5,711
Curette	4,829	10.1	708	3.5	5,537
Rongeur	1,730	3.6	2,790	13.9	4,520
Scissor	2,928	6.1	996	5.0	3,924
Drill	1,639	3.4	2,116	10.6	3,755
Surgicel	2,847	6.0	739	3.7	3,586
Surgical Knife	1,364	2.9	1,422	7.1	2,786
Suction Coagulator	1,400	2.9	1,188	5.9	2,588
Aspirating dissector	88	0.2	2,319	11.6	2,407
Bipolar Forceps	1,934	4.1	263	1.3	2,197
Sonopet pineapple tip	1,991	4.2	0	0.0	1,991
Straight Forceps	1,602	3.4	173	0.9	1,775
Doppler	1,592	3.3	0	0.0	1,592
unknown	1,172	2.5	23	0.1	1,195
Backbiting rongeur	144	0.3	1,041	5.2	1,185
Tissue shaver	934	2.0	57	0.3	991
Cottle Elevator	855	1.8	0	0.0	855
Through cutting forceps	255	0.5	542	2.7	797
Suction microdebrider	94	0.2	497	2.5	591
Floseal Gel	502	1.1	40	0.2	542
Collagen Matrix	280	0.6	0	0.0	280
Floseal Applicator	199	0.4	40	0.2	239
Monopolar Electrocautery	190	0.4	0	0.0	190
Local Anesthesia Needle	139	0.3	0	0.0	139
Straight Curette	102	0.2	16	0.1	118
Needle	31	0.1	0	0.0	31
Tisseel Applicator	23	0.0	0	0.0	23

B Zero-Shot Evaluation Prompt Template

The following prompt template is used for zero-shot tool detection evaluation across all vision-language models.

Prompt Template

```
Identify all surgical tools visible in this surgical video frame.

Valid tool names:
- Aspirating dissector
- Backbiting rongeur
- Bipolar Forceps
- Collagen Matrix
- Cotton Patty
- Cottle Elevator
- Curette
- Doppler
- Drill
- Floseal Applicator
- Floseal Gel
- Grasper
- Irrigation
- Local Anesthesia Needle
- Monopolar Electrocautery
- Needle
- Rhoton Dissector
- Rongeur
- Scissor
- Sonopet- pineapple tip
- Straight Curette
- Straight Forceps
- Suction
- Suction Coagulator
- Suction microdebrider
- Surgicel
- Surgical Knife
- Tisseel Applicator
- Tissue shaver
- Through cutting forceps
- unknown

Return your answer as a JSON object with this exact format:
{"detected_tools": ["Tool Name 1", "Tool Name 2"]}

Use ONLY the exact tool names from the list above.
If a tool appears multiple times, list it only once.
If no tools are visible, return: {"detected_tools": []}

Return ONLY valid JSON, nothing else.
```

C Output Validation Methodology

Model outputs are parsed by extracting the first valid JSON object from the response text using regex matching. An output is classified as an output validation failure if it meets any of the following conditions: (1) the response does not contain valid JSON or is missing the `detected_tools` key (JSON failure), (2) the `detected_tools` value is not an array of strings (schema failure), or (3)

the array contains tool names that do not exactly match any entry in the provided ontology of 31 valid tool names (ontology failure, e.g., misspellings, capitalization mismatches, or hallucinated tool names). All output validation failures are treated as empty predictions (`detected_tools: []`). This ensures that a model's inability to follow the output format or correctly name tools is penalized rather than silently excluded. Representative failure examples are shown in Appendix G.

D LoRA Fine-Tuning Configuration

All fine-tuning experiments use a fixed random seed of 42 for reproducibility. LoRA adapters [Hu et al., 2021] are applied to the query, key, value, and output projection matrices in both the language model and vision encoder attention layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `out_proj`).

JSON Generation (Section 3.2). LoRA rank $r = 1024$, scaling factor $\alpha = 2048$, dropout 0.05. Training: 10 epochs, learning rate 2×10^{-5}, effective batch size 64 (per-GPU batch size 1×8 gradient accumulation steps $\times$ 8 H200 GPUs), bfloat16 precision. Gradient checkpointing is used to reduce memory consumption. Training is distributed using PyTorch DDP [Li et al., 2020]. Training and evaluation required 80 wall-clock hours (640 GPU-hours on H200 GPUs). During training, exact match accuracy and Jaccard similarity are periodically evaluated on fixed random subsets of 100 training and 100 validation frames.

Classification Head (Section 3.3). The base model processes the image and prompt, and we apply mean pooling over the final hidden states (excluding padding tokens) to obtain a fixed-dimensional representation. A single linear layer (no hidden layers) maps this representation to 31 output logits (one per tool class), trained with binary cross-entropy loss averaged across all tool classes. At inference, we apply a sigmoid activation and threshold at 0.5 to obtain binary predictions. LoRA rank $r = 1024$, $\alpha = 2048$, dropout 0.05. Training: 10 epochs, learning rate 5×10^{-6}, effective batch size 32 (per-GPU batch size 1×4 gradient accumulation steps $\times$ 8 H200 GPUs).

Rank Sweep (Section 3.4). We sweep LoRA ranks $r \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$, setting $\alpha = 2r$ for each. Training: 3 epochs per configuration, effective batch size 32, with other settings matching the classification head configuration above. The full sweep required approximately 62 wall-clock hours (492 GPU-hours on H200 GPUs). Trainable parameters scale linearly with rank, from 4.7M at $r = 2$ to 2.4B at $r = 1024$. For each configuration, we report training and validation exact match accuracy with 95% confidence intervals.

E Per-Tool Metrics for LoRA Fine-Tuning with JSON Output

Table 8 shows per-tool classification metrics on the validation set ($n = 20{,}016$ frames) for Gemma 3 27B fine-tuned with LoRA to produce JSON outputs.

Table 8: Per-tool classification metrics for Gemma 3 27B with LoRA fine-tuning (JSON output). TP = true positives, FP = false positives, FN = false negatives, TN = true negatives.

Tool	TP	FP	FN	TN	Accuracy	Precision	Recall	F1
Aspirating dissector	0	0	2319	17697	0.884	0.000	0.000	0.000
Backbiting rongeur	141	44	900	18931	0.953	0.762	0.135	0.230
Bipolar Forceps	0	103	263	19650	0.982	0.000	0.000	0.000
Cottle Elevator	0	271	0	19745	0.987	0.000	0.000	0.000
Cotton Patty	1729	639	414	17234	0.947	0.730	0.807	0.767
Curette	103	3	605	19305	0.970	0.972	0.145	0.253
Doppler	0	31	0	19985	0.999	0.000	0.000	0.000
Drill	1637	109	479	17791	0.971	0.938	0.774	0.848
Floseal Applicator	0	17	40	19959	0.997	0.000	0.000	0.000
Floseal Gel	2	90	38	19886	0.994	0.022	0.050	0.030
Grasper	60	666	449	18841	0.944	0.083	0.118	0.097
Irrigation	9	104	103	19800	0.990	0.080	0.080	0.080
Monopolar Electrocautery	0	831	0	19185	0.959	0.000	0.000	0.000
Rhoton Dissector	799	3086	663	15468	0.813	0.206	0.546	0.299
Rongeur	522	1509	2268	15717	0.811	0.257	0.187	0.217
Scissor	530	703	466	18317	0.942	0.430	0.532	0.475
Sonopet- pineapple tip	0	1	0	20015	1.000	0.000	0.000	0.000
Straight Forceps	44	406	129	19437	0.973	0.098	0.254	0.141
Suction	9411	4289	1274	5042	0.722	0.687	0.881	0.772
Surgicel	347	36	392	19241	0.979	0.906	0.470	0.619
Through cutting forceps	1	11	541	19463	0.972	0.083	0.002	0.004
Tissue shaver	0	0	57	19959	0.997	0.000	0.000	0.000
unknown	2	1223	21	18770	0.938	0.002	0.087	0.003

F Per-Tool Metrics for LoRA Fine-Tuning with Classification Head

Table 9 shows per-tool classification metrics on the validation set (n = 20,016 frames) for Gemma 3 27B fine-tuned with LoRA and a linear classification head.

Table 9: Per-tool classification metrics for Gemma 3 27B with LoRA fine-tuning and classification head. TP = true positives, FP = false positives, FN = false negatives, TN = true negatives.

Tool	TP	FP	FN	TN	Accuracy	Precision	Recall	F1
Aspirating dissector	0	0	2319	17697	0.884	0.000	0.000	0.000
Backbiting rongeur	20	7	1021	18968	0.949	0.741	0.019	0.038
Bipolar Forceps	0	10	263	19743	0.986	0.000	0.000	0.000
Collagen Matrix	0	0	0	20016	1.000	0.000	0.000	0.000
Cottle Elevator	0	3	0	20013	1.000	0.000	0.000	0.000
Cotton Patty	1513	1112	630	16761	0.913	0.576	0.706	0.635
Curette	169	9	539	19299	0.973	0.949	0.239	0.382
Doppler	0	1	0	20015	1.000	0.000	0.000	0.000
Drill	1672	98	444	17802	0.973	0.945	0.790	0.861
Floseal Applicator	0	0	40	19976	0.998	0.000	0.000	0.000
Floseal Gel	2	1	38	19975	0.998	0.667	0.050	0.093
Grasper	10	307	499	19200	0.960	0.032	0.020	0.024
Irrigation	2	0	110	19904	0.995	1.000	0.018	0.035
Local Anesthesia Needle	0	1	0	20015	1.000	0.000	0.000	0.000
Monopolar Electrocautery	0	0	0	20016	1.000	0.000	0.000	0.000
Needle	0	0	0	20016	1.000	0.000	0.000	0.000
Rhoton Dissector	863	735	599	17819	0.933	0.540	0.590	0.564
Rongeur	620	26	2170	17200	0.890	0.960	0.222	0.361
Scissor	135	82	861	18938	0.953	0.622	0.136	0.223
Sonopet- pineapple tip	0	0	0	20016	1.000	0.000	0.000	0.000
Straight Curette	0	0	16	20000	0.999	0.000	0.000	0.000
Straight Forceps	43	24	130	19819	0.992	0.642	0.249	0.358
Suction	9452	4604	1233	4727	0.708	0.673	0.885	0.764
Suction Coagulator	739	0	449	18828	0.978	1.000	0.622	0.767
Suction microdebrider	0	0	497	19519	0.975	0.000	0.000	0.000
Surgical Knife	69	6	1353	18588	0.932	0.920	0.049	0.092
Surgicel	464	14	275	19263	0.986	0.971	0.628	0.763
Through cutting forceps	0	1	542	19473	0.973	0.000	0.000	0.000
Tisseel Applicator	0	0	0	20016	1.000	0.000	0.000	0.000
Tissue shaver	2	4	55	19955	0.997	0.333	0.035	0.064
unknown	0	26	23	19967	0.998	0.000	0.000	0.000

Table 10 shows per-tool ROC-AUC and AUPRC.

Table 10: Per-tool ROC-AUC and AUPRC for Gemma 3 27B with LoRA fine-tuning and classification head.

Tool	ROC-AUC	AUPRC
Suction Coagulator	1.000	0.995
Drill	0.983	0.935
Straight Forceps	0.943	0.351
Curette	0.940	0.600
Surgical Knife	0.939	0.578
Cotton Patty	0.929	0.699
Rongeur	0.920	0.729
Rhoton Dissector	0.920	0.540
Backbiting rongeur	0.918	0.482
Surgicel	0.908	0.701
Floseal Gel	0.878	0.351
Straight Curette	0.841	0.008
Suction	0.819	0.847
Through cutting forceps	0.797	0.092
Irrigation	0.776	0.117
Scissor	0.766	0.259
Tissue shaver	0.737	0.099
Bipolar Forceps	0.723	0.043
Grasper	0.708	0.047
Aspirating dissector	0.655	0.158
unknown	0.618	0.001
Floseal Applicator	0.494	0.002
Suction microdebrider	0.298	0.016
Macro Average	0.805	0.376

The model achieves high ROC-AUC (> 0.9) for tools well-represented in training (Suction Coagulator, Drill, Straight Forceps, Curette, Surgical Knife, Cotton Patty, Rongeur, Rhoton Dissector, Backbiting rongeur, Surgicel), but lower values for tools with limited training data or those appearing predominantly in validation procedures.

Our per-tool evaluation metrics (ROC-AUC and AUPRC) are reported for the 23 surgical instruments that appear in the validation set.

G Zero-Shot Output Validation Failure Examples

Output validation failures in zero-shot evaluation are not merely JSON formatting issues. Table 11 shows representative failed outputs from Qwen2-VL-2B-Instruct, the model with the highest output validation failure rate (41.7%). In most cases, the model produces syntactically valid JSON but hallucinates tool names that do not exist in the provided list, such as "Stirrup Curtain," "Parallel Shears," "Microlaryngeal electrodes," and "Semitendinosus skin dissection." These hallucinated names are not surgical instruments and indicate a fundamental failure in visual recognition, not a formatting limitation.

Table 11: Representative output validation failure examples from Qwen2-VL-2B-Instruct zero-shot evaluation. Each row shows the model's raw output and the failure category. Outputs are from randomly sampled frames (seed=42).

Model Output	Failure Category
{''detected_tools'': [''SCissor'']}	Ontology: misspelled tool name ("SCissor" vs. "Scissor")
[''Curette'', ''Floseal Gel'', ''Straight Curette'', ''Scissors'', ''Single- Precision Scissors'', ''Suction microdebrider'']	JSON + Ontology: missing JSON wrapper; "Scissors" and "Single- Precision Scissors" not in valid list
{''detected_tools'': [''Backbiting Rongeur'', ''Rhoton Dissector'']}	Ontology: capitalization mismatch ("Rongeur" vs. "rongeur")
{''detected_tools'': [''Collagen Matrix'', ''Curette'', ''Tourniquet'']}	Ontology: hallucinated tool ("Tourniquet")
{''detected_tools'': [''Local Anesthesia Needle'']]	JSON: malformed JSON (extra closing bracket)
{''detected_tools'': [''Aspirating dissector'', ''Froseal Applicator'', ''Froseal Gel'', ''Grasper'', ''Intraoperative Hemostat'', ''Rhoton Dissector'']}	Ontology: misspelled tools ("Froseal") and hallucinated tool ("Intraoperative Hemostat")
{''detected_tools'': [''Curette'', ''Rhoton Dissector'', ''Stirrup Curtain'', ''Parallel Shears'']}	Ontology: hallucinated tools ("Stirrup Curtain," "Parallel Shears")
{''detected_tools'': [''Microlaryngeal electrodes'', ''Laryngeal transducers'', ''Semitendinosus skin dissection'', ''Subglottic surgery'', ''Endoscopic cautery'', ''Suture'', ''Hydrofracture'']}	Ontology: all 7 predicted tools are hallucinated; none are surgical instruments

H YOLOv12-m Training Configuration

Table 12 shows the training configuration for YOLOv12-m used in Section 3.5. All hyperparameters use YOLO default values; no hyperparameter search was performed.

Table 12: YOLOv12-m training configuration (300-epoch run with best exact match set accuracy).

Parameter	Value
Model	YOLOv12-m (26M parameters)
Pretrained weights	`yolo12m.pt` (COCO)
Epochs	300
Batch size	32 total (4/GPU × 8 GPUs)
Image size	1280×1280
Hardware	8×L40S GPUs
Training time	11.3 wall-clock hours (90.6 GPU-hours)
Learning rate schedule	Cosine
Mixed precision	AMP (automatic)
Early stopping patience	20 epochs
Mosaic close epoch	10
Data caching	Disabled (disk-based)
Confidence threshold (eval)	0.25
Random seed	42

I Per-Tool Metrics for CholecT50 Evaluation

Tables 13–17 show per-tool classification metrics on the CholecT50 validation set (n = 19,923 frames) for zero-shot Gemma 3 27B, fine-tuned Gemma 3 27B, and YOLOv12-m.

Table 13: Per-tool classification metrics for Gemma 3 27B **zero-shot** on CholecT50. TP = true positives, FP = false positives, FN = false negatives, TN = true negatives.

Tool	TP	FP	FN	TN	Accuracy	Precision	Recall	F1
grasper	6978	2740	5552	4653	0.584	0.718	0.557	0.627
hook	2092	883	8651	8297	0.522	0.703	0.195	0.305
bipolar	838	12096	246	6743	0.381	0.065	0.773	0.120
irrigator	247	4355	696	14625	0.747	0.054	0.262	0.089
clipper	59	686	627	18551	0.934	0.079	0.086	0.083
scissors	163	3993	313	15454	0.784	0.039	0.342	0.070

In the zero-shot setting, bipolar has 12,096 false positives, irrigator has 4,355, and scissors has 3,993. Hook has a recall of 0.195. Grasper achieves the highest F1 (0.627).

Table 14: Per-tool classification metrics for Gemma 3 27B **fine-tuned** (LoRA + classification head) on CholecT50. TP = true positives, FP = false positives, FN = false negatives, TN = true negatives.

Tool	TP	FP	FN	TN	Accuracy	Precision	Recall	F1
grasper	11614	1307	916	6086	0.888	0.899	0.927	0.913
hook	10491	306	252	8874	0.972	0.972	0.977	0.974
bipolar	805	48	279	18791	0.984	0.944	0.743	0.831
irrigator	663	51	280	18929	0.983	0.929	0.703	0.800
clipper	614	42	72	19195	0.994	0.936	0.895	0.915
scissors	285	36	191	19411	0.989	0.888	0.599	0.715

After fine-tuning, the model achieves a macro ROC-AUC of 0.966 and macro AUPRC of 0.883.

Table 15 shows per-tool ROC-AUC and AUPRC.

Table 15: Per-tool ROC-AUC and AUPRC for Gemma 3 27B fine-tuned on CholecT50. Sorted by ROC-AUC in descending order.

Tool	ROC-AUC	AUPRC
hook	0.989	0.985
clipper	0.989	0.909
irrigator	0.969	0.836
grasper	0.941	0.960
bipolar	0.959	0.856
scissors	0.947	0.753
Macro Average	0.966	0.883

After fine-tuning, hook achieves the highest F1 (0.974), and all tools achieve F1 > 0.7. The largest change from zero-shot to fine-tuned is for bipolar (F1: $0.120 \rightarrow 0.831$) and irrigator (F1: $0.089 \rightarrow 0.800$).

Table 16: Per-tool classification metrics for **YOLOv12-m** on CholecT50.

Tool	TP	FP	FN	TN	Accuracy	Precision	Recall	F1
grasper	11938	1945	592	5448	0.873	0.860	0.953	0.904
hook	10590	517	153	8663	0.966	0.953	0.986	0.969
bipolar	841	73	243	18766	0.984	0.920	0.776	0.842
irrigator	772	76	171	18904	0.988	0.910	0.819	0.862
clipper	624	68	62	19169	0.993	0.902	0.910	0.906
scissors	275	36	201	19411	0.988	0.884	0.578	0.699

For YOLOv12-m, scissors has the lowest recall and F1 among all tools (476 validation instances). YOLOv12-m achieves higher F1 than Gemma on irrigator (0.862 vs. 0.800), while Gemma achieves higher F1 on grasper (0.913 vs. 0.904).

Table 17: Per-tool ROC-AUC for YOLOv12-m on CholecT50 (using maximum detection confidence per class as the continuous score).

Tool	ROC-AUC
hook	0.992
clipper	0.967
grasper	0.958
irrigator	0.955
bipolar	0.920
scissors	0.825
Macro Average	0.936

J Robustness Check: CNN without Bounding Box Supervision

Our comparison between YOLOv12-m and VLM-based classifiers evaluates tool presence only: both models are scored on whether the predicted tool set matches the ground truth set, ignoring spatial

localization. However, the two approaches differ in their training signal. YOLO is trained with bounding box supervision, while VLMs receive only set-level labels during fine-tuning. YOLO's localization objective may confer an indirect advantage for presence detection by forcing the model to ground each prediction spatially, reducing hallucinated detections. Conversely, VLMs must learn tool presence from a weaker supervisory signal. This asymmetry could mean that YOLO's advantage partly reflects the richer information content of bounding box annotations rather than a fundamental architectural superiority for the presence detection task.

To test this, we train a ResNet-50 (23.6M parameters; He et al. 2015) for multi-label tool classification using *only* set-level labels—the same supervisory signal available to VLMs—with no bounding box information. The model uses ImageNet-pretrained weights, a dropout layer ($p = 0.5$) followed by a 31-class linear head, and is trained with binary cross-entropy loss (label smoothing $\epsilon = 0.1$). We use differential learning rates (backbone: 10^{-4}, head: 10^{-3}), AdamW optimizer (weight decay 10^{-2}), cosine annealing schedule, and aggressive data augmentation (random resized crops, color jitter, random erasing, rotation). Training uses 8×L40S GPUs for 50 epochs with a total batch size of 512.

Figure 8 shows training dynamics. The model achieves 39.6% exact match accuracy (95% CI: 38.9%–40.3%) on the validation set (n = 20,016), with 52.6% Jaccard similarity, 70.3% top-1 accuracy, and 0.673 micro F1.

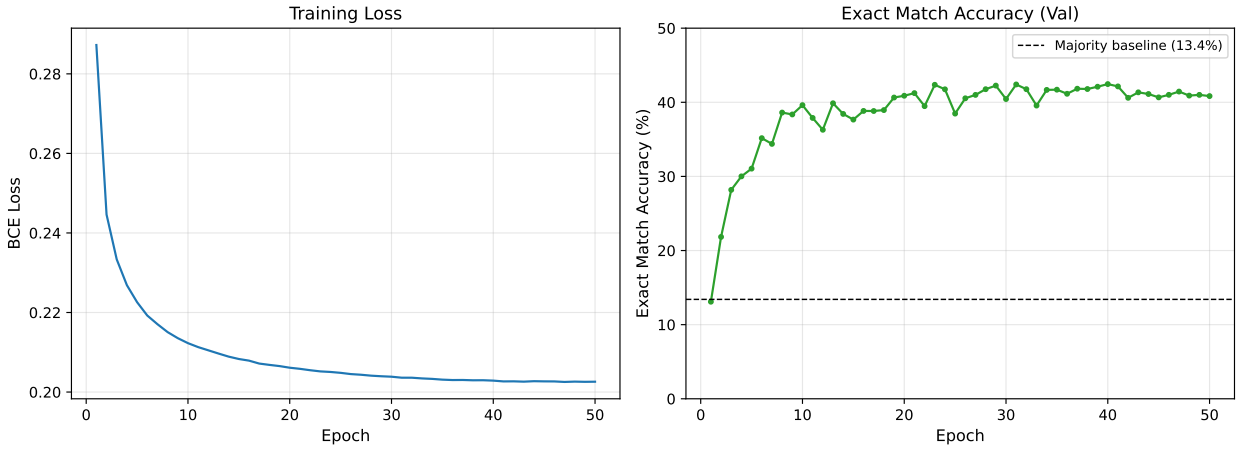

Figure 8: Training dynamics for ResNet-50 multi-label classification without bounding box supervision. Left: Training loss (binary cross-entropy with label smoothing). Right: Exact match accuracy on the validation set evaluated at each epoch. The dashed line indicates the majority class baseline (13.4%).

This result exceeds all zero-shot VLMs and matches the 3-epoch LoRA rank sweep at r = 1024 (39.6%, Section 3.4), but falls below the best fine-tuned VLM (Gemma 3 27B with LoRA classification head trained for 10 epochs: 51.08%, Section 3.3), despite using roughly 1,000× fewer parameters and receiving the same set-level supervision. The ResNet-50's performance also falls below YOLOv12-m (54.7%), suggesting that bounding box supervision does confer some advantage for presence detection. The fact that a 23.6M-parameter CNN trained with set-level labels alone outperforms all zero-shot VLMs—including models with up to 235B parameters—underscores the difficulty of surgical tool detection as a zero-shot task.

K Effect of Sampling Temperature on Zero-Shot Accuracy

To investigate whether sampling temperature affects zero-shot tool detection performance, we sweep the generation temperature of Gemma 3 27B-it from 0 (greedy decoding) to 2.0 in increments of 0.1, evaluating exact match accuracy on the full validation set (n = 20,016) at each setting. Figure 9 shows the results with 95% Wilson binomial confidence intervals.

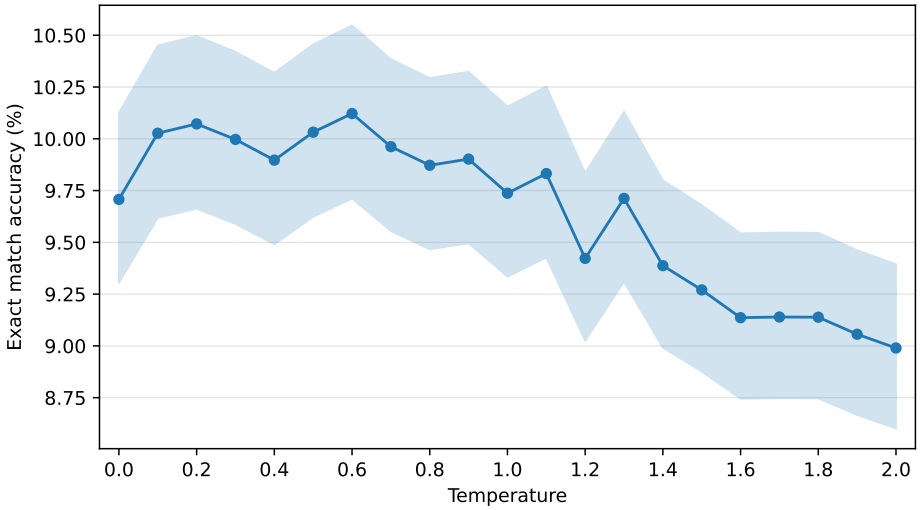

Figure 9: Zero-shot exact match accuracy of Gemma 3 27B-it on the SDSC-EEA validation set (n = 20,016) as a function of sampling temperature. Shaded region: 95% Wilson binomial CI. Accuracy is largely insensitive to temperature in the 0–0.7 range (∼10%), with a gradual decline at higher temperatures. At $T \geq 1.5$, output validation failures begin to appear (up to 38 at $T = 2.0$), indicating that high temperatures degrade the model's ability to produce valid JSON. All CIs overlap, confirming that temperature has no statistically significant effect on zero-shot performance.

L Per-Tool Metrics for Zero-Shot VLM Evaluation

The following tables present per-tool classification metrics for each zero-shot vision-language model evaluated on the full dataset ($n \approx$ 67,634 frames). Output validation failures are treated as empty predictions (`detected_tools: []`) for the purpose of computing per-tool precision, recall, and F1. Only tools with at least one true positive, false positive, or false negative are shown. Tools are sorted by F1 score in descending order.

Table 18: Per-tool metrics for Qwen3-VL-235B-A22B-Thinking (235B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	28745	12845	10868	0.691	0.726	0.708
Cotton Patty	2329	387	8498	0.858	0.215	0.344
Sonopet pineapple tip	367	356	1624	0.508	0.184	0.270
Grasper	2216	8314	5106	0.210	0.303	0.248
Drill	467	260	2930	0.642	0.137	0.226
Rongeur	593	4131	3577	0.126	0.142	0.133
Straight Forceps	292	3286	1403	0.082	0.172	0.111
Bipolar Forceps	110	848	2087	0.115	0.050	0.070
Rhoton Dissector	193	883	5520	0.179	0.034	0.057
Suction Coagulator	88	1003	2298	0.081	0.037	0.051
Floseal Gel	13	115	514	0.102	0.025	0.040
Surgicel	36	512	3551	0.066	0.010	0.017
Backbiting rongeur	11	256	1174	0.041	0.009	0.015
Scissor	30	225	3895	0.118	0.008	0.014
Floseal Applicator	2	115	214	0.017	0.009	0.012
Suction microdebrider	16	2728	575	0.006	0.027	0.010
Aspirating dissector	12	407	2369	0.029	0.005	0.009
Irrigation	2	175	384	0.011	0.005	0.007
Cottle Elevator	3	158	852	0.019	0.004	0.006
Surgical Knife	7	17	2780	0.292	0.003	0.005
Through cutting forceps	1	108	796	0.009	0.001	0.002
Curette	1	4	5538	0.200	0.000	0.000
Collagen Matrix	0	206	280	0.000	0.000	0.000
Doppler	0	32	1482	0.000	0.000	0.000
Local Anesthesia Needle	0	22	139	0.000	0.000	0.000
Monopolar Electrocautery	0	18	190	0.000	0.000	0.000
Needle	0	117	31	0.000	0.000	0.000
Straight Curette	0	2	102	0.000	0.000	0.000
Tisseel Applicator	0	8	23	0.000	0.000	0.000
Tissue shaver	0	9	991	0.000	0.000	0.000
unknown	0	0	1195	0.000	0.000	0.000

Table 19: Per-tool metrics for Qwen3-VL-32B-Instruct (32B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	33716	12951	7507	0.722	0.818	0.767
Grasper	1441	2642	6187	0.353	0.189	0.246
Cotton Patty	1089	146	10088	0.882	0.097	0.175
Rhoton Dissector	1324	9249	4489	0.125	0.228	0.162
Bipolar Forceps	395	2933	1861	0.119	0.175	0.141
Rongeur	519	2156	4577	0.194	0.102	0.134
Floseal Gel	85	1148	460	0.069	0.156	0.096
Curette	352	1985	5226	0.151	0.063	0.089
Surgical Knife	242	3498	2561	0.065	0.086	0.074
Backbiting rongeur	63	634	1125	0.090	0.053	0.067
Surgicel	143	1415	3516	0.092	0.039	0.055
Aspirating dissector	146	3783	2263	0.037	0.061	0.046
Sonopet pineapple tip	44	37	1948	0.543	0.022	0.042
Floseal Applicator	14	701	234	0.020	0.056	0.029
Scissor	61	268	3871	0.185	0.016	0.029
Suction Coagulator	94	4420	2511	0.021	0.036	0.026
Drill	38	190	3729	0.167	0.010	0.019
Straight Forceps	18	568	1761	0.031	0.010	0.015
Monopolar Electrocautery	5	1182	185	0.004	0.026	0.007
Through cutting forceps	2	32	796	0.059	0.003	0.005
Irrigation	1	126	394	0.008	0.003	0.004
Collagen Matrix	0	905	280	0.000	0.000	0.000
Cottle Elevator	0	0	866	0.000	0.000	0.000
Doppler	0	0	1601	0.000	0.000	0.000
Local Anesthesia Needle	0	5	139	0.000	0.000	0.000
Needle	0	108	31	0.000	0.000	0.000
Straight Curette	0	326	118	0.000	0.000	0.000
Suction microdebrider	0	1368	592	0.000	0.000	0.000
Tisseel Applicator	0	210	23	0.000	0.000	0.000
Tissue shaver	0	24	993	0.000	0.000	0.000
unknown	0	0	1249	0.000	0.000	0.000

Table 20: Per-tool metrics for Qwen3-VL-8B-Instruct (8B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	15594	4603	24890	0.772	0.385	0.514
Grasper	1587	3204	5735	0.331	0.217	0.262
Surgical Knife	228	886	2559	0.205	0.082	0.117
Cotton Patty	649	35	10320	0.949	0.059	0.111
Rhoton Dissector	522	4399	5191	0.106	0.091	0.098
Aspirating dissector	842	19376	1565	0.042	0.350	0.074
Drill	116	115	3639	0.502	0.031	0.058
Rongeur	135	1045	4851	0.114	0.027	0.044
Floseal Applicator	7	263	232	0.026	0.029	0.028
Irrigation	14	660	372	0.021	0.036	0.026
Scissor	46	241	3879	0.160	0.012	0.022
Surgicel	34	335	3553	0.092	0.009	0.017
Floseal Gel	5	68	537	0.068	0.009	0.016
Suction Coagulator	22	1047	2566	0.021	0.009	0.012
Bipolar Forceps	10	171	2187	0.055	0.005	0.008
Tissue shaver	5	340	986	0.014	0.005	0.007
Curette	21	132	5518	0.137	0.004	0.007
Monopolar Electrocautery	1	361	189	0.003	0.005	0.004
Straight Forceps	3	12	1772	0.200	0.002	0.003
Backbiting rongeur	1	1	1184	0.500	0.001	0.002
Suction microdebrider	5	7259	586	0.001	0.008	0.001
Collagen Matrix	0	54	280	0.000	0.000	0.000
Cottle Elevator	0	23	855	0.000	0.000	0.000
Doppler	0	0	1592	0.000	0.000	0.000
Local Anesthesia Needle	0	7	139	0.000	0.000	0.000
Needle	0	21	31	0.000	0.000	0.000
Sonopet pineapple tip	0	20	1991	0.000	0.000	0.000
Straight Curette	0	34	118	0.000	0.000	0.000
Through cutting forceps	0	3	797	0.000	0.000	0.000
Tisseel Applicator	0	6	23	0.000	0.000	0.000
unknown	0	2	1195	0.000	0.000	0.000

Table 21: Per-tool metrics for Qwen3-VL-4B-Instruct (4B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	26189	10548	14295	0.713	0.647	0.678
Drill	1126	402	2629	0.737	0.300	0.426
Grasper	4922	24241	2400	0.169	0.672	0.270
Cotton Patty	1297	120	9672	0.915	0.118	0.209
Scissor	2016	18097	1909	0.100	0.514	0.168
Surgical Knife	251	3478	2536	0.067	0.090	0.077
Curette	290	3419	5249	0.078	0.052	0.063
Aspirating dissector	90	566	2317	0.137	0.037	0.059
Suction Coagulator	271	9345	2317	0.028	0.105	0.044
Irrigation	7	368	379	0.019	0.018	0.018
Local Anesthesia Needle	1	52	138	0.019	0.007	0.010
Through cutting forceps	5	167	792	0.029	0.006	0.010
Straight Curette	1	127	117	0.008	0.008	0.008
Surgicel	14	172	3573	0.075	0.004	0.007
Floseal Gel	2	57	540	0.034	0.004	0.007
Monopolar Electrocautery	1	110	189	0.009	0.005	0.007
Floseal Applicator	1	68	238	0.014	0.004	0.006
unknown	5	355	1190	0.014	0.004	0.006
Backbiting rongeur	2	30	1183	0.062	0.002	0.003
Straight Forceps	3	132	1772	0.022	0.002	0.003
Rongeur	8	230	4978	0.034	0.002	0.003
Rhoton Dissector	3	24	5710	0.111	0.001	0.001
Sonopet pineapple tip	1	23	1990	0.042	0.001	0.001
Bipolar Forceps	0	25	2197	0.000	0.000	0.000
Collagen Matrix	0	172	280	0.000	0.000	0.000
Cottle Elevator	0	84	855	0.000	0.000	0.000
Doppler	0	32	1592	0.000	0.000	0.000
Needle	0	59	31	0.000	0.000	0.000
Suction microdebrider	0	214	591	0.000	0.000	0.000
Tisseel Applicator	0	71	23	0.000	0.000	0.000
Tissue shaver	0	224	991	0.000	0.000	0.000

Table 22: Per-tool metrics for Qwen3-VL-2B-Instruct (2B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	24482	14568	16002	0.627	0.605	0.616
Grasper	4424	30813	2898	0.126	0.604	0.208
Drill	581	2309	3174	0.201	0.155	0.175
Rongeur	1472	19535	3514	0.070	0.295	0.113
Scissor	1236	18066	2689	0.064	0.315	0.106
Curette	710	9347	4829	0.071	0.128	0.091
Surgical Knife	2110	43265	677	0.047	0.757	0.088
Cotton Patty	603	2503	10366	0.194	0.055	0.086
Rhoton Dissector	481	5552	5232	0.080	0.084	0.082
Suction Coagulator	732	18203	1856	0.039	0.283	0.068
Straight Forceps	476	16156	1299	0.029	0.268	0.052
Bipolar Forceps	100	2770	2097	0.035	0.046	0.039
Aspirating dissector	79	1986	2328	0.038	0.033	0.035
Tissue shaver	269	15834	722	0.017	0.271	0.031
Doppler	47	1623	1545	0.028	0.030	0.029
Cottle Elevator	240	19232	615	0.012	0.281	0.024
Through cutting forceps	156	13162	641	0.012	0.196	0.022
unknown	16	690	1179	0.023	0.013	0.017
Sonopet pineapple tip	19	351	1972	0.051	0.010	0.016
Backbiting rongeur	19	1345	1166	0.014	0.016	0.015
Floseal Gel	58	7239	484	0.008	0.107	0.015
Suction microdebrider	214	29996	377	0.007	0.362	0.014
Surgicel	28	464	3559	0.057	0.008	0.014
Irrigation	35	5748	351	0.006	0.091	0.011
Monopolar Electrocautery	29	8210	161	0.004	0.153	0.007
Collagen Matrix	5	1746	275	0.003	0.018	0.005
Floseal Applicator	18	7161	221	0.003	0.075	0.005
Local Anesthesia Needle	14	6382	125	0.002	0.101	0.004
Straight Curette	8	5957	110	0.001	0.068	0.003
Needle	2	1901	29	0.001	0.065	0.002
Tisseel Applicator	1	14970	22	0.000	0.043	0.000

Table 23: Per-tool metrics for Qwen2.5-VL-72B-Instruct (72B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	10304	3016	30180	0.774	0.255	0.383
Sonopet pineapple tip	174	272	1817	0.390	0.087	0.143
Grasper	472	1330	6850	0.262	0.064	0.103
Cotton Patty	391	285	10578	0.578	0.036	0.067
Rongeur	156	1070	4830	0.127	0.031	0.050
Scissor	107	585	3818	0.155	0.027	0.046
Curette	118	902	5421	0.116	0.021	0.036
unknown	92	3878	1103	0.023	0.077	0.036
Surgical Knife	61	622	2726	0.089	0.022	0.035
Surgicel	68	529	3519	0.114	0.019	0.033
Rhoton Dissector	106	1226	5607	0.080	0.019	0.030
Straight Forceps	36	1250	1739	0.028	0.020	0.024
Backbiting rongeur	14	275	1171	0.048	0.012	0.019
Drill	36	93	3719	0.279	0.010	0.019
Bipolar Forceps	20	329	2177	0.057	0.009	0.016
Irrigation	4	406	382	0.010	0.010	0.010
Aspirating dissector	8	419	2399	0.019	0.003	0.006
Cottle Elevator	2	254	853	0.008	0.002	0.004
Floseal Gel	1	63	541	0.016	0.002	0.003
Through cutting forceps	1	66	796	0.015	0.001	0.002
Collagen Matrix	1	658	279	0.002	0.004	0.002
Monopolar Electrocautery	1	1189	189	0.001	0.005	0.001
Suction Coagulator	1	94	2587	0.011	0.000	0.001
Doppler	0	17	1592	0.000	0.000	0.000
Floseal Applicator	0	142	239	0.000	0.000	0.000
Local Anesthesia Needle	0	7	139	0.000	0.000	0.000
Needle	0	6	31	0.000	0.000	0.000
Straight Curette	0	229	118	0.000	0.000	0.000
Suction microdebrider	0	64	591	0.000	0.000	0.000
Tisseel Applicator	0	38	23	0.000	0.000	0.000
Tissue shaver	0	139	991	0.000	0.000	0.000

Table 24: Per-tool metrics for Qwen2.5-VL-32B-Instruct (32B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	10990	4861	29494	0.693	0.271	0.390
Cotton Patty	929	518	10040	0.642	0.085	0.150
Scissor	269	2041	3656	0.116	0.069	0.086
Bipolar Forceps	236	3079	1961	0.071	0.107	0.086
Sonopet pineapple tip	77	50	1914	0.606	0.039	0.073
Curette	352	3950	5187	0.082	0.064	0.072
Surgical Knife	157	1770	2630	0.081	0.056	0.067
Rhoton Dissector	260	2364	5453	0.099	0.046	0.062
Rongeur	208	1972	4778	0.095	0.042	0.058
Grasper	230	948	7092	0.195	0.031	0.054
Drill	56	102	3699	0.354	0.015	0.029
Straight Forceps	29	864	1746	0.032	0.016	0.022
Through cutting forceps	10	307	787	0.032	0.013	0.018
Aspirating dissector	24	1249	2383	0.019	0.010	0.013
Floseal Gel	4	76	538	0.050	0.007	0.013
Irrigation	24	3707	362	0.006	0.062	0.012
Backbiting rongeur	6	115	1179	0.050	0.005	0.009
Floseal Applicator	2	230	237	0.009	0.008	0.008
Monopolar Electrocautery	9	3185	181	0.003	0.047	0.005
Local Anesthesia Needle	1	294	138	0.003	0.007	0.005
Suction Coagulator	5	223	2583	0.022	0.002	0.004
Tissue shaver	2	257	989	0.008	0.002	0.003
unknown	2	92	1193	0.021	0.002	0.003
Doppler	1	73	1591	0.014	0.001	0.001
Surgicel	2	22	3585	0.083	0.001	0.001
Collagen Matrix	0	166	280	0.000	0.000	0.000
Cottle Elevator	0	89	855	0.000	0.000	0.000
Needle	0	307	31	0.000	0.000	0.000
Straight Curette	0	607	118	0.000	0.000	0.000
Suction microdebrider	0	55	591	0.000	0.000	0.000
Tisseel Applicator	0	55	23	0.000	0.000	0.000

Table 25: Per-tool metrics for Qwen2.5-VL-7B-Instruct (7B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Surgical Knife	799	11805	1988	0.063	0.287	0.104
Grasper	480	1506	6842	0.242	0.066	0.103
Suction	2207	1031	38277	0.682	0.055	0.101
Curette	399	2817	5140	0.124	0.072	0.091
Scissor	185	1476	3740	0.111	0.047	0.066
Rhoton Dissector	192	1678	5521	0.103	0.034	0.051
Suction Coagulator	113	1792	2475	0.059	0.044	0.050
Backbiting rongeur	43	632	1142	0.064	0.036	0.046
Rongeur	112	890	4874	0.112	0.022	0.037
Bipolar Forceps	59	998	2138	0.056	0.027	0.036
Drill	56	334	3699	0.144	0.015	0.027
Aspirating dissector	55	2346	2352	0.023	0.023	0.023
unknown	19	462	1176	0.040	0.016	0.023
Straight Forceps	26	789	1749	0.032	0.015	0.020
Doppler	15	399	1577	0.036	0.009	0.015
Tissue shaver	11	1071	980	0.010	0.011	0.011
Cottle Elevator	10	1200	845	0.008	0.012	0.010
Cotton Patty	40	60	10929	0.400	0.004	0.007
Sonopet pineapple tip	8	251	1983	0.031	0.004	0.007
Local Anesthesia Needle	1	156	138	0.006	0.007	0.007
Floseal Applicator	3	652	236	0.005	0.013	0.007
Floseal Gel	2	110	540	0.018	0.004	0.006
Through cutting forceps	3	220	794	0.013	0.004	0.006
Monopolar Electrocautery	4	1302	186	0.003	0.021	0.005
Surgicel	9	211	3578	0.041	0.003	0.005
Irrigation	1	271	385	0.004	0.003	0.003
Collagen Matrix	0	110	280	0.000	0.000	0.000
Needle	0	743	31	0.000	0.000	0.000
Straight Curette	0	762	118	0.000	0.000	0.000
Suction microdebrider	0	109	591	0.000	0.000	0.000
Tisseel Applicator	0	99	23	0.000	0.000	0.000

Table 26: Per-tool metrics for Qwen2.5-VL-3B-Instruct (3B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Drill	573	1106	3182	0.341	0.153	0.211
Surgical Knife	434	5197	2353	0.077	0.156	0.103
Curette	498	4650	5041	0.097	0.090	0.093
Suction	1650	660	38834	0.714	0.041	0.077
Rongeur	276	2926	4710	0.086	0.055	0.067
Straight Forceps	175	4054	1600	0.041	0.099	0.058
Grasper	157	783	7165	0.167	0.021	0.038
Bipolar Forceps	55	686	2142	0.074	0.025	0.037
Local Anesthesia Needle	3	86	136	0.034	0.022	0.026
Through cutting forceps	29	2458	768	0.012	0.036	0.018
Rhoton Dissector	51	421	5662	0.108	0.009	0.016
Scissor	30	239	3895	0.112	0.008	0.014
Aspirating dissector	23	877	2384	0.026	0.010	0.014
Cotton Patty	62	139	10907	0.308	0.006	0.011
Collagen Matrix	3	308	277	0.010	0.011	0.010
Surgicel	19	214	3568	0.082	0.005	0.010
Irrigation	4	466	382	0.009	0.010	0.009
Floseal Gel	3	179	539	0.016	0.006	0.008
Sonopet pineapple tip	6	167	1985	0.035	0.003	0.006
Monopolar Electrocautery	3	1070	187	0.003	0.016	0.005
Backbiting rongeur	4	544	1181	0.007	0.003	0.005
Suction Coagulator	6	162	2582	0.036	0.002	0.004
Tissue shaver	3	429	988	0.007	0.003	0.004
Straight Curette	7	3403	111	0.002	0.059	0.004
Cottle Elevator	2	216	853	0.009	0.002	0.004
Floseal Applicator	1	541	238	0.002	0.004	0.003
Doppler	2	236	1590	0.008	0.001	0.002
Suction microdebrider	1	498	590	0.002	0.002	0.002
Needle	0	42	31	0.000	0.000	0.000
Tisseel Applicator	0	22	23	0.000	0.000	0.000
unknown	0	67	1195	0.000	0.000	0.000

Table 27: Per-tool metrics for Qwen2-VL-72B-Instruct (72B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	9085	4437	31399	0.672	0.224	0.336
Drill	458	907	3297	0.336	0.122	0.179
Grasper	1069	4589	6253	0.189	0.146	0.165
Curette	401	3817	5138	0.095	0.072	0.082
Bipolar Forceps	191	2473	2006	0.072	0.087	0.079
Scissor	219	1650	3706	0.117	0.056	0.076
Rhoton Dissector	327	3356	5386	0.089	0.057	0.070
Sonopet pineapple tip	107	1039	1884	0.093	0.054	0.068
Rongeur	208	2034	4778	0.093	0.042	0.058
Surgical Knife	91	1493	2696	0.057	0.033	0.042
Straight Forceps	64	1818	1711	0.034	0.036	0.035
Backbiting rongeur	52	2098	1133	0.024	0.044	0.031
Surgicel	59	334	3528	0.150	0.016	0.030
unknown	35	1182	1160	0.029	0.029	0.029
Aspirating dissector	75	2785	2332	0.026	0.031	0.028
Cotton Patty	146	70	10823	0.676	0.013	0.026
Suction Coagulator	52	1361	2536	0.037	0.020	0.026
Floseal Gel	10	509	532	0.019	0.018	0.019
Irrigation	19	2964	367	0.006	0.049	0.011
Local Anesthesia Needle	1	74	138	0.013	0.007	0.009
Doppler	6	135	1586	0.043	0.004	0.007
Tissue shaver	5	719	986	0.007	0.005	0.006
Cottle Elevator	3	281	852	0.011	0.004	0.005
Suction microdebrider	7	2096	584	0.003	0.012	0.005
Through cutting forceps	3	369	794	0.008	0.004	0.005
Collagen Matrix	3	979	277	0.003	0.011	0.005
Monopolar Electrocautery	7	2923	183	0.002	0.037	0.004
Straight Curette	3	1897	115	0.002	0.025	0.003
Floseal Applicator	1	436	238	0.002	0.004	0.003
Needle	0	9	31	0.000	0.000	0.000
Tisseel Applicator	0	83	23	0.000	0.000	0.000

Table 28: Per-tool metrics for Qwen2-VL-7B-Instruct (7B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	14191	7274	26293	0.661	0.351	0.458
Drill	1691	5743	2064	0.227	0.450	0.302
Curette	948	8601	4591	0.099	0.171	0.126
Grasper	680	4174	6642	0.140	0.093	0.112
Scissor	537	5600	3388	0.088	0.137	0.107
Surgical Knife	449	5921	2338	0.070	0.161	0.098
Suction Coagulator	338	7800	2250	0.042	0.131	0.063
Rongeur	219	2482	4767	0.081	0.044	0.057
Straight Forceps	226	5955	1549	0.037	0.127	0.057
Cotton Patty	276	513	10693	0.350	0.025	0.047
Bipolar Forceps	48	1001	2149	0.046	0.022	0.030
Aspirating dissector	61	1912	2346	0.031	0.025	0.028
Doppler	39	1172	1553	0.032	0.024	0.028
Tissue shaver	38	2435	953	0.015	0.038	0.022
Backbiting rongeur	21	876	1164	0.023	0.018	0.020
Rhoton Dissector	61	544	5652	0.101	0.011	0.019
Through cutting forceps	29	2265	768	0.013	0.036	0.019
Cottle Elevator	55	5065	800	0.011	0.064	0.018
Irrigation	93	12128	293	0.008	0.241	0.015
Floseal Gel	10	870	532	0.011	0.018	0.014
Surgicel	22	290	3565	0.071	0.006	0.011
Sonopet pineapple tip	12	288	1979	0.040	0.006	0.010
Suction microdebrider	19	3950	572	0.005	0.032	0.008
Monopolar Electrocautery	18	4194	172	0.004	0.095	0.008
Floseal Applicator	7	1680	232	0.004	0.029	0.007
Local Anesthesia Needle	4	1353	135	0.003	0.029	0.005
Straight Curette	6	3032	112	0.002	0.051	0.004
Collagen Matrix	0	91	280	0.000	0.000	0.000
Needle	0	837	31	0.000	0.000	0.000
Tisseel Applicator	0	700	23	0.000	0.000	0.000
unknown	0	134	1195	0.000	0.000	0.000

Table 29: Per-tool metrics for Qwen2-VL-2B-Instruct (2B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Curette	1110	11909	4429	0.085	0.200	0.120
Grasper	748	5279	6574	0.124	0.102	0.112
Rhoton Dissector	593	6046	5120	0.089	0.104	0.096
Drill	238	1733	3517	0.121	0.063	0.083
Aspirating dissector	283	5676	2124	0.047	0.118	0.068
Surgical Knife	343	7496	2444	0.044	0.123	0.065
Surgicel	139	2026	3448	0.064	0.039	0.048
Bipolar Forceps	114	2978	2083	0.037	0.052	0.043
Doppler	133	4515	1459	0.029	0.084	0.043
Rongeur	126	1728	4860	0.068	0.025	0.037
Suction	633	414	39851	0.605	0.016	0.030
Backbiting rongeur	28	1005	1157	0.027	0.024	0.025
Scissor	49	715	3876	0.064	0.012	0.021
Tissue shaver	25	1485	966	0.017	0.025	0.020
Cottle Elevator	28	1921	827	0.014	0.033	0.020
Straight Forceps	21	519	1754	0.039	0.012	0.018
Sonopet pineapple tip	22	468	1969	0.045	0.011	0.018
Suction Coagulator	27	507	2561	0.051	0.010	0.017
Floseal Gel	25	2934	517	0.008	0.046	0.014
Collagen Matrix	31	4494	249	0.007	0.111	0.013
Irrigation	14	1991	372	0.007	0.036	0.012
Through cutting forceps	12	1274	785	0.009	0.015	0.012
Suction microdebrider	8	878	583	0.009	0.014	0.011
Floseal Applicator	20	4768	219	0.004	0.084	0.008
Straight Curette	5	1170	113	0.004	0.042	0.008
Monopolar Electrocautery	5	1324	185	0.004	0.026	0.007
Cotton Patty	16	71	10953	0.184	0.001	0.003
Tisseel Applicator	1	732	22	0.001	0.043	0.003
Local Anesthesia Needle	1	790	138	0.001	0.007	0.002
Needle	0	371	31	0.000	0.000	0.000
unknown	0	33	1195	0.000	0.000	0.000

Table 30: Per-tool metrics for Gemma 3 27B-it (27B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	39555	26381	929	0.600	0.977	0.743
Rongeur	1424	14204	3562	0.091	0.286	0.138
Drill	154	32	3601	0.828	0.041	0.078
Bipolar Forceps	277	5299	1920	0.050	0.126	0.071
Grasper	329	1596	6993	0.171	0.045	0.071
Cotton Patty	232	69	10737	0.771	0.021	0.041
Through cutting forceps	216	14551	581	0.015	0.271	0.028
Straight Forceps	25	666	1750	0.036	0.014	0.020
Scissor	42	177	3883	0.192	0.011	0.020
Surgical Knife	24	182	2763	0.117	0.009	0.016
Sonopet pineapple tip	21	816	1970	0.025	0.011	0.015
Surgicel	25	169	3562	0.129	0.007	0.013
Curette	35	129	5504	0.213	0.006	0.012
Irrigation	49	8101	337	0.006	0.127	0.011
Monopolar Electrocautery	23	5683	167	0.004	0.121	0.008
Suction microdebrider	3	461	588	0.006	0.005	0.006
Cottle Elevator	1	60	854	0.016	0.001	0.002
Straight Curette	1	962	117	0.001	0.008	0.002
Rhoton Dissector	5	27	5708	0.156	0.001	0.002
Aspirating dissector	0	10	2407	0.000	0.000	0.000
Backbiting rongeur	0	0	1185	0.000	0.000	0.000
Collagen Matrix	0	538	280	0.000	0.000	0.000
Doppler	0	0	1592	0.000	0.000	0.000
Floseal Applicator	0	4	239	0.000	0.000	0.000
Floseal Gel	0	12	542	0.000	0.000	0.000
Local Anesthesia Needle	0	5	139	0.000	0.000	0.000
Needle	0	257	31	0.000	0.000	0.000
Suction Coagulator	0	54	2588	0.000	0.000	0.000
Tisseel Applicator	0	0	23	0.000	0.000	0.000
Tissue shaver	0	47	991	0.000	0.000	0.000
unknown	0	2	1195	0.000	0.000	0.000

Table 31: Per-tool metrics for Gemma 3 12B-it (12B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	26679	15734	13805	0.629	0.659	0.644
Grasper	4431	23958	2891	0.156	0.605	0.248
Suction Coagulator	1180	15044	1408	0.073	0.456	0.125
Rhoton Dissector	825	6977	4888	0.106	0.144	0.122
Scissor	327	2297	3598	0.125	0.083	0.100
Sonopet pineapple tip	101	611	1890	0.142	0.051	0.075
Local Anesthesia Needle	9	236	130	0.037	0.065	0.047
Drill	85	27	3670	0.759	0.023	0.044
Rongeur	158	2530	4828	0.059	0.032	0.041
Bipolar Forceps	65	1562	2132	0.040	0.030	0.034
Cotton Patty	181	250	10788	0.420	0.017	0.032
Surgical Knife	71	2207	2716	0.031	0.025	0.028
Surgicel	38	119	3549	0.242	0.011	0.020
Monopolar Electrocautery	5	311	185	0.016	0.026	0.020
Irrigation	149	24718	237	0.006	0.386	0.012
Straight Forceps	10	552	1765	0.018	0.006	0.009
Suction microdebrider	3	366	588	0.008	0.005	0.006
Through cutting forceps	2	47	795	0.041	0.003	0.005
Tissue shaver	3	282	988	0.011	0.003	0.005
Aspirating dissector	7	667	2400	0.010	0.003	0.005
Curette	9	32	5530	0.220	0.002	0.003
Backbiting rongeur	0	0	1185	0.000	0.000	0.000
Collagen Matrix	0	954	280	0.000	0.000	0.000
Cottle Elevator	0	76	855	0.000	0.000	0.000
Doppler	0	26	1592	0.000	0.000	0.000
Floseal Applicator	0	133	239	0.000	0.000	0.000
Floseal Gel	0	43	542	0.000	0.000	0.000
Needle	0	582	31	0.000	0.000	0.000
Straight Curette	0	1	118	0.000	0.000	0.000
Tisseel Applicator	0	0	23	0.000	0.000	0.000
unknown	0	3	1195	0.000	0.000	0.000

Table 32: Per-tool metrics for Gemma 3 4B-it (4B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	39017	26280	1467	0.598	0.964	0.738
Grasper	3699	24871	3623	0.129	0.505	0.206
Rongeur	1714	23095	3272	0.069	0.344	0.115
Surgical Knife	2314	43203	473	0.051	0.830	0.096
Bipolar Forceps	597	13902	1600	0.041	0.272	0.072
Suction Coagulator	136	1328	2452	0.093	0.053	0.067
Straight Forceps	1299	46446	476	0.027	0.732	0.052
Surgicel	96	489	3491	0.164	0.027	0.046
Monopolar Electrocautery	5	131	185	0.037	0.026	0.031
Curette	102	1512	5437	0.063	0.018	0.029
Cottle Elevator	200	15114	655	0.013	0.234	0.025
Aspirating dissector	90	5718	2317	0.015	0.037	0.022
Local Anesthesia Needle	1	5	138	0.167	0.007	0.014
Irrigation	100	15849	286	0.006	0.259	0.012
Drill	21	7	3734	0.750	0.006	0.011
Rhoton Dissector	32	415	5681	0.072	0.006	0.010
Scissor	14	123	3911	0.102	0.004	0.007
Doppler	6	448	1586	0.013	0.004	0.006
Cotton Patty	27	8	10942	0.771	0.002	0.005
Straight Curette	4	2463	114	0.002	0.034	0.003
Backbiting rongeur	1	28	1184	0.034	0.001	0.002
Needle	7	9710	24	0.001	0.226	0.001
Collagen Matrix	0	3	280	0.000	0.000	0.000
Floseal Applicator	0	0	239	0.000	0.000	0.000
Floseal Gel	0	0	542	0.000	0.000	0.000
Sonopet pineapple tip	0	0	1991	0.000	0.000	0.000
Suction microdebrider	0	38	591	0.000	0.000	0.000
Through cutting forceps	0	32	797	0.000	0.000	0.000
Tisseel Applicator	0	0	23	0.000	0.000	0.000
Tissue shaver	0	108	991	0.000	0.000	0.000
unknown	0	0	1195	0.000	0.000	0.000

Table 33: Per-tool metrics for MedGemma 3 27B-it (27B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	38413	24419	1970	0.611	0.951	0.744
Rongeur	2042	13802	2478	0.129	0.452	0.201
Surgical Knife	851	9387	1936	0.083	0.305	0.131
Curette	286	2206	5253	0.115	0.052	0.071
Sonopet pineapple tip	72	110	1919	0.396	0.036	0.066
Cotton Patty	344	29	10625	0.922	0.031	0.061
Grasper	190	295	7132	0.392	0.026	0.049
Suction Coagulator	218	12464	2370	0.017	0.084	0.029
Straight Forceps	36	870	1739	0.040	0.020	0.027
unknown	147	10757	1048	0.013	0.123	0.024
Monopolar Electrocautery	13	1375	177	0.009	0.068	0.016
Irrigation	52	9955	334	0.005	0.135	0.010
Suction microdebrider	33	6177	558	0.005	0.056	0.010
Scissor	15	44	3910	0.254	0.004	0.008
Straight Curette	4	3246	114	0.001	0.034	0.002
Tissue shaver	1	30	990	0.032	0.001	0.002
Doppler	1	14	1591	0.067	0.001	0.001
Bipolar Forceps	1	23	2196	0.042	0.000	0.001
Aspirating dissector	1	26	2406	0.037	0.000	0.001
Rhoton Dissector	2	19	5711	0.095	0.000	0.001
Surgicel	1	28	3586	0.034	0.000	0.001
Backbiting rongeur	0	10	1185	0.000	0.000	0.000
Collagen Matrix	0	636	280	0.000	0.000	0.000
Cottle Elevator	0	12	855	0.000	0.000	0.000
Drill	0	16	3755	0.000	0.000	0.000
Floseal Applicator	0	15	239	0.000	0.000	0.000
Floseal Gel	0	15	542	0.000	0.000	0.000
Local Anesthesia Needle	0	16	139	0.000	0.000	0.000
Needle	0	17	31	0.000	0.000	0.000
Through cutting forceps	0	44	797	0.000	0.000	0.000
Tisseel Applicator	0	15	23	0.000	0.000	0.000

Table 34: Per-tool metrics for Llama-3.2-90B-Vision (90B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	24755	6689	15626	0.787	0.613	0.689
Drill	1264	1005	2489	0.557	0.337	0.420
Cotton Patty	2569	2671	8400	0.490	0.234	0.317
Rongeur	1062	5120	3458	0.172	0.235	0.198
Curette	657	2324	4882	0.220	0.119	0.154
Grasper	669	1046	6653	0.390	0.091	0.148
Rhoton Dissector	705	7066	5008	0.091	0.123	0.105
Surgical Knife	167	1094	2620	0.132	0.060	0.083
Bipolar Forceps	153	2900	2044	0.050	0.070	0.058
Surgicel	97	1064	3490	0.084	0.027	0.041
Backbiting rongeur	53	1699	1132	0.030	0.045	0.036
Floseal Gel	24	927	517	0.025	0.044	0.032
Doppler	27	225	1565	0.107	0.017	0.029
Scissor	64	396	3861	0.139	0.016	0.029
Irrigation	12	588	374	0.020	0.031	0.024
Monopolar Electrocautery	67	5254	123	0.013	0.353	0.024
Through cutting forceps	18	846	779	0.021	0.023	0.022
Straight Forceps	27	1143	1748	0.023	0.015	0.018
Floseal Applicator	6	932	233	0.006	0.025	0.010
Sonopet pineapple tip	10	207	1981	0.046	0.005	0.009
Suction Coagulator	39	6827	2550	0.006	0.015	0.008
Collagen Matrix	3	553	277	0.005	0.011	0.007
unknown	45	11803	1150	0.004	0.038	0.007
Cottle Elevator	2	225	853	0.009	0.002	0.004
Suction microdebrider	7	5387	584	0.001	0.012	0.002
Tissue shaver	1	594	990	0.002	0.001	0.001
Aspirating dissector	0	478	2407	0.000	0.000	0.000
Local Anesthesia Needle	0	295	138	0.000	0.000	0.000
Needle	0	291	31	0.000	0.000	0.000
Straight Curette	0	467	118	0.000	0.000	0.000
Tisseel Applicator	0	789	23	0.000	0.000	0.000

Table 35: Per-tool metrics for Llama-3.2-11B-Vision (11B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	505	392	39979	0.563	0.012	0.024
Tissue shaver	10	467	981	0.021	0.010	0.014
Aspirating dissector	13	364	2394	0.034	0.005	0.009
Curette	23	190	5516	0.108	0.004	0.008
Bipolar Forceps	8	132	2189	0.057	0.004	0.007
Rongeur	15	163	4971	0.084	0.003	0.006
Suction microdebrider	2	168	589	0.012	0.003	0.005
Surgical Knife	8	309	2779	0.025	0.003	0.005
Scissor	8	87	3917	0.084	0.002	0.004
Grasper	13	115	7309	0.102	0.002	0.003
Drill	5	9	3750	0.357	0.001	0.003
Rhoton Dissector	5	19	5708	0.208	0.001	0.002
Suction Coagulator	1	53	2587	0.019	0.000	0.001
Cotton Patty	1	4	10968	0.200	0.000	0.000
Backbiting rongeur	0	20	1185	0.000	0.000	0.000
Collagen Matrix	0	3	280	0.000	0.000	0.000
Cottle Elevator	0	16	855	0.000	0.000	0.000
Doppler	0	19	1592	0.000	0.000	0.000
Floseal Applicator	0	2	239	0.000	0.000	0.000
Floseal Gel	0	0	542	0.000	0.000	0.000
Irrigation	0	99	386	0.000	0.000	0.000
Local Anesthesia Needle	0	26	139	0.000	0.000	0.000
Monopolar Electrocautery	0	38	190	0.000	0.000	0.000
Needle	0	11	31	0.000	0.000	0.000
Sonopet pineapple tip	0	0	1991	0.000	0.000	0.000
Straight Curette	0	13	118	0.000	0.000	0.000
Straight Forceps	0	17	1775	0.000	0.000	0.000
Surgicel	0	85	3587	0.000	0.000	0.000
Through cutting forceps	0	24	797	0.000	0.000	0.000
Tisseel Applicator	0	78	23	0.000	0.000	0.000
unknown	0	2	1195	0.000	0.000	0.000

Table 36: Per-tool metrics for LLaVA-1.5-13B (13B) zero-shot evaluation. P = precision, R = recall.

Tool	TP	FP	FN	P	R	F1
Suction	22423	14880	17960	0.601	0.555	0.577
Cotton Patty	6339	27699	4630	0.186	0.578	0.282
Grasper	6850	54394	472	0.112	0.936	0.200
Rhoton Dissector	5405	58418	308	0.085	0.946	0.155
Curette	5429	59769	110	0.083	0.980	0.153
Rongeur	3798	53913	722	0.066	0.840	0.122
Scissor	3876	61604	49	0.059	0.988	0.112
Surgicel	2338	39493	1249	0.056	0.652	0.103
Drill	1661	27772	2094	0.056	0.442	0.100
Suction Coagulator	1309	29208	1279	0.043	0.506	0.079
Surgical Knife	2406	56776	381	0.041	0.863	0.078
Aspirating dissector	2353	62594	54	0.036	0.978	0.070
Bipolar Forceps	2150	62184	47	0.033	0.979	0.065
Sonopet pineapple tip	905	28587	1086	0.031	0.455	0.057
Backbiting rongeur	110	2896	1075	0.037	0.093	0.052
Straight Forceps	821	28844	954	0.028	0.463	0.052
Doppler	1010	49108	582	0.020	0.634	0.039
unknown	163	7614	1032	0.021	0.136	0.036
Tissue shaver	924	61184	67	0.015	0.932	0.029
Cottle Elevator	355	29201	500	0.012	0.415	0.023
Floseal Gel	261	28761	281	0.009	0.482	0.018
Suction microdebrider	252	29226	339	0.009	0.426	0.017
Irrigation	374	63968	12	0.006	0.969	0.012
Floseal Applicator	91	28909	148	0.003	0.381	0.006
Monopolar Electrocautery	185	64062	5	0.003	0.974	0.006
Collagen Matrix	73	25090	207	0.003	0.261	0.006
Local Anesthesia Needle	123	49177	16	0.002	0.885	0.005
Straight Curette	53	29563	65	0.002	0.449	0.004
Needle	18	32276	13	0.001	0.581	0.001
Tisseel Applicator	17	46181	6	0.000	0.739	0.001
Through cutting forceps	0	137	797	0.000	0.000	0.000

M Qwen3-VL-235B-A22B-Thinking's thinking

Qwen3-VL-235B-A22B-Thinking Response

```
</think>
So, let's look at the surgical video frame. There are two tools visible. Let's
    check the list. The tools here seem to be a suction and maybe a grasper or
    something else. Wait, the left tool might be suction, and the right one...
    Let's see. The valid names: Suction is one. Then, the other tool--- maybe
    Grasper? Wait, in the image, the two instruments. Let's check each.

First, Suction: yes, one of the tools is likely suction. Then, the other tool---
    Grasper? Let's confirm. The list has Grasper. So detected tools would be
    Suction and Grasper. Wait, let's make sure. The image shows two surgical
    instruments. Let's check the options. Suction is a valid name. Grasper is also
    there. So the detected tools are Suction and Grasper.
</think>

{"detected_tools": ["Suction", "Grasper"]}
```

Qwen3-VL-235B-A22B-Thinking Response

```
</think>
So, let's look at the image. First, identify the tools. There's a white gauze-like
    material, which might be a Cotton Patty. Then there are two instruments: one
    looks like a suction tool (Suction), and another might be a Grasper. Wait,
    let's check the list.

Cotton Patty is in the list. Then, the metallic instrument with a tube--- Suction.
    Also, the other instrument that's holding or manipulating--- Grasper. Let's
    confirm.

In the image, the white pad is Cotton Patty. The tube-like instrument is Suction.
    The other tool with the jaws is Grasper. Are there any others? Let's check the
    valid names.

So detected tools: Cotton Patty, Suction, Grasper.
</think>

{"detected_tools": ["Cotton Patty", "Suction", "Grasper"]}
```

Qwen3-VL-235B-A22B-Thinking Response

```
</think>
So, let's look at the surgical video frame. There are two tools visible. First,
    the one on the left---what is that? Wait, the options include Suction, maybe?
    Wait, no. Wait, the tool on the left might be a Suction Coagulator? Or maybe
    Suction. Wait, let's check the list. The tools present: the left tool looks
    like a suction device, maybe Suction. The right tool---wait, the right one has
    a metallic tip, maybe a Suction Coagulator? Wait, no. Wait, let's think. The
    options include Suction. Let's see. In the image, there are two instruments.
    Let's check each.

First, the left instrument: it's a metallic tube, possibly a Suction tool. The
    right instrument: maybe a Suction Coagulator? Wait, no. Wait, the valid tools:
    Suction is a valid name. Let's check. The list has "Suction" as a tool. So if
    there's a suction device, that's one. Are there others?

Wait, the image shows two surgical instruments. Let's identify each. The left one:
    maybe Suction. The right one: maybe Suction Coagulator? Wait, no. Wait, the
    Suction Coagulator is a specific tool. Alternatively, maybe the right one is a
    Suction microdebrider? No, the microdebrider is different. Wait, let's check
    the options again.

Wait, the tools in the list: Suction is a valid entry. So if there's a suction
    device, that's one. Are there any other tools? Let's see. The image has two
    instruments. Let's check each.

First instrument (left): looks like a suction tube. So "Suction".

Second instrument (right): the tip is rounded, maybe a Suction Coagulator? Wait,
    Suction Coagulator is a tool that combines suction and coagulation. But maybe
```

in this case, the right tool is a Suction Coagulator. Wait, but let's check the options. The valid names include "Suction Coagulator" and "Suction".

Wait, but maybe both are suction-related. Wait, no. Let's think again. The left tool: maybe it's a Suction. The right tool: perhaps a Suction Coagulator. But I need to be precise.

Alternatively, maybe the right tool is a Suction microdebrider? No, the microdebrider has a different tip. Wait, the image shows two instruments. Let's check the list again.

Wait, the correct approach is to match the visible tools to the exact names. Let's see:

- Suction: yes, if there's a suction device.

Looking at the image, there are two instruments. Let's check each.

First, the left instrument: it's a metallic tube, likely a suction device. So "Suction".

Second, the right instrument: the tip is a rounded metallic part, maybe a Suction Coagulator. Wait, but Suction Coagulator is a specific tool. Alternatively, maybe it's a different type. Wait, the list includes "Suction" as a separate entry. So if both are suction-related, but the exact names.

Wait, the problem says to list each tool only once even if multiple. But in this case, are there two different tools or same?

Wait, maybe the left is Suction and the right is Suction Coagulator. Let's check.

Suction Coagulator: a tool that suctions and coagulates. The right tool might be that. The left one is just Suction.

So detected tools would be ["Suction", "Suction Coagulator"].

Wait, but let's confirm. Let's look at the image again. The left instrument is a larger tube, maybe a suction, and the right one is a different type. Alternatively, maybe both are suction, but the list has "Suction" as a single entry. Wait, no---if they are different tools, but the names must be exact.

Wait, the valid tool names include "Suction" and "Suction Coagulator". So if one is Suction and the other is Suction Coagulator, then both are listed.

Alternatively, maybe the right tool is a Suction microdebrider? No, the microdebrider has a rotating tip, which isn't visible here.

Wait, another approach: check common endoscopic tools. In sinus surgery, for example, suction and coagulator are common. The image shows two instruments. Let's assume left is Suction, right is Suction Coagulator.

So the detected tools would be Suction and Suction Coagulator.

```
</think>

{"detected_tools": ["Suction", "Suction Coagulator"]}
```

M.1 Model Responses to Questions about Neurosurgery

Question 1

What roles do suction, cotton patties, and hemostatic agents play during pituitary tumor resection?

Gemma 3 27B Response

```
Okay, let's break down the roles of suction, cotton patties, and hemostatic agents
    during pituitary tumor resection. These are *critical* components of the
    surgical toolkit, and understanding their interplay is key to a successful and
    safe operation.  I'll cover each, then discuss how they work together.  This
    is a complex surgery, so I'll aim for a detailed but understandable
    explanation.  **Please remember I am an AI and this is for informational
    purposes only. It is not medical advice. Consult with a qualified healthcare
    professional for any health concerns or before making any decisions related to
    your health or treatment.**

**1. Suction**

* **Primary Role: Maintaining a Clear Operative Field & Removing Fluid/Debris.**
    This is arguably the *most* important element throughout the entire procedure.
     The surgeon needs to *see* what they're doing, and the surgical space within
    the sella (the bony cavity where the pituitary gland sits) is incredibly small
    and prone to bleeding.
* **Types of Suction Used:**
* **Irrigation Suction:**  This is the workhorse.  It simultaneously delivers
    irrigating fluid (usually normal saline) to wash away blood, debris, and tumor
    fragments *while* simultaneously suctioning it all away.  The fluid keeps the
    field clean and allows the surgeon to visualize the delicate structures.
    Different tip sizes and shapes are used depending on the specific area being
    worked on.  Gentle irrigation is crucial to avoid damaging brain tissue.
* **Ultrasonic Aspirator (CUSA):**  Often used for softer tumors (like pituitary
    adenomas). CUSA uses high-frequency vibrations to break up the tumor tissue
    into small fragments, which are then suctioned away.  It's more selective than
    traditional suction, helping to preserve normal pituitary tissue and
    surrounding structures.  It's particularly useful for tumors that are adherent
    to critical structures.
* **Bipolar Suction:**  Some suction devices combine suction with bipolar
    electrocautery. This allows for simultaneous coagulation of small bleeding
    vessels while suctioning away the blood.
* **Why it's crucial in pituitary surgery:**
* **Limited Space:** The sella is tiny. Blood quickly obscures the view.
* **Proximity to Critical Structures:** The pituitary gland is nestled near the
    optic nerves, the cavernous sinus (a major venous structure), and the
    brainstem.  Clear visualization is essential to avoid damage.
```

```
* **Fragile Tissue:** The pituitary gland and surrounding tissues are delicate.



**2. Cottonoid Patties (Cotton Patties/Neuro-Patties)**

* **Primary Role: Hemostasis (Stopping Bleeding) & Tissue Protection.**  These
    aren't your typical cotton balls. Neuro-Patties are specifically designed for
    neurosurgery. They are made of highly absorbent, tightly woven cotton.
* **How they work:**
* **Direct Pressure:**  The primary mechanism.  Gently packing the patties into
    the surgical bed applies direct pressure to bleeding points, helping to stop
    the hemorrhage.
* **Absorption:** They soak up blood, further contributing to a dry field.
* **Temporary Space Filling:**  They can temporarily fill small defects in the
    bone or dura (the membrane covering the brain) created during tumor removal.
* **Tissue Protection:** They can act as a barrier to protect delicate structures
    from inadvertent injury during manipulation.
* **Specific Use in Pituitary Surgery:**
* **Sella Floor Repair:** After the tumor is removed, there's often a defect in
    the sella floor.  A patty soaked in hemostatic agent (see below) is often
    placed here to promote healing and prevent cerebrospinal fluid (CSF) leak.
* **Packing Around Vessels:**  Used to gently pack around the cavernous sinus or
    other vessels to control bleeding.
* **Dural Closure Assistance:**  Can help to support the dural repair.
* **Important Considerations:**
* **Careful Removal:** Patties *must* be carefully removed after the surgery to
    avoid leaving any remnants behind, which could cause infection or
    inflammation.  They are often marked with radiopaque threads to aid in
    identification on post-operative imaging.
* **Not a Long-Term Solution:**  Patties are for temporary hemostasis and space
    filling. They don't promote permanent tissue repair on their own.



**3. Hemostatic Agents**

* **Primary Role:  Enhancing Hemostasis (Stopping Bleeding).** These are
    substances applied to bleeding surfaces to accelerate clot formation. They are
    often used *in conjunction* with suction and cotton patties.
* **Types of Hemostatic Agents Commonly Used in Pituitary Surgery:**
* **Gelatin-Based (e.g., Gelfoam):**  A porous sponge that provides a matrix for
    clot formation. It's relatively inexpensive and effective for controlling
    minor bleeding.
* **Oxidized Regenerated Cellulose (e.g., Surgicel):**  Another absorbent material
    that promotes clotting. It has some antibacterial properties.
* **Thrombin:** A potent clotting factor.  Often used in combination with collagen
    or other carriers.  Can be very effective for controlling more significant
    bleeding.  (There are bovine and human forms; human thrombin is preferred to
    minimize immune reactions).
```

* **Fibrin Sealants:** Contain fibrinogen and thrombin, mimicking the natural clotting cascade. They create a strong, adhesive clot. Often used for dural closure.
* **Polysaccharide-Based (e.g., Arista AH):** Derived from plant sources, these agents promote rapid clot formation and are biocompatible.
* **How they're used in pituitary surgery:**
* **Soaking Cotton Patties:** Patties are frequently soaked in a hemostatic agent before being placed in the surgical bed. This provides both direct pressure and a boost to clotting.
* **Direct Application:** Hemostatic agents can be applied directly to bleeding vessels or surfaces.
* **Dural Sealants:** Fibrin sealants are often used to reinforce the dural closure, reducing the risk of CSF leak.
* **Important Considerations:**
* **Appropriate Agent Selection:** The choice of hemostatic agent depends on the type and severity of bleeding, the location, and the surgeon's preference.
* **Potential Complications:** While generally safe, some hemostatic agents can cause inflammation or allergic reactions.

How They Work Together: A Surgical Scenario

Imagine the surgeon is removing a pituitary adenoma:

1. **Initial Exposure:** The surgeon gains access to the sella through the nose (transsphenoidal approach).
2. **Tumor Dissection:** Using microsurgical instruments, the surgeon carefully dissects the tumor from surrounding structures. **Continuous irrigation suction** is used to keep the field clear of blood and debris.
3. **Bleeding Control:** As vessels are encountered, the surgeon may use **bipolar cautery** for small vessels. For larger bleeding points, they might apply a **hemostatic agent** (like thrombin) directly or pack the area with a **cotton patty soaked in hemostatic agent**.
4. **Tumor Removal (CUSA):** If the tumor is soft, a **CUSA** might be used to debulk it, with simultaneous suction to remove the fragmented tissue.
5. **Sella Floor Repair:** Once the tumor is removed, any defects in the sella floor are addressed. A **cotton patty saturated with a hemostatic agent and potentially a dural sealant** is carefully placed to promote healing and prevent CSF leak.
6. **Dural Closure:** The dura is closed, often with sutures and reinforced with a **fibrin sealant**.
7. **Final Inspection:** The surgeon meticulously inspects the field with **suction** to ensure complete hemostasis and remove any remaining debris or packing material.

In conclusion: Suction, cotton patties, and hemostatic agents are not independent tools but rather a coordinated system used to achieve a safe and effective pituitary tumor resection. They work synergistically to provide

```
    visualization, control bleeding, and promote healing in this delicate and
    challenging surgery.



**Disclaimer:**  This information is for general knowledge and informational
    purposes only, and does not constitute medical advice. It is essential to
    consult with a qualified healthcare professional for any health concerns or
    before making any decisions related to your health or treatment.
```

N Gemma 3 27B Responses to Neurosurgery Questions

This appendix presents the complete responses from Gemma 3 27B to five neurosurgery questions. The model was prompted with each question and generated the responses shown below.

N.1 Model Configuration

Parameter	Value
Model	Gemma 3 27B Instruct
Model ID	`google/gemma-3-27b-it`
Precision	bfloat16
Max New Tokens	8,192
Temperature	0.0

Table 37: Gemma 3 27B generation configuration.

Question 2

Describe the standard steps of an endoscopic endonasal transsphenoidal approach to a pituitary adenoma.

Gemma 3 27B Response

```
## Endoscopic Endonasal Transsphenoidal Approach to Pituitary Adenoma: A
    Step-by-Step Description

The endoscopic endonasal transsphenoidal approach (EETS) is now the gold standard
    for removing most pituitary adenomas. Here's a detailed breakdown of the
    standard steps involved.  It's a complex procedure, and this is a general
    overview; specifics can vary based on tumor size, location, and surgeon
    preference.

**I. Preoperative Preparation:**

* **Imaging:** High-resolution MRI with and without contrast is *essential* for
    planning. CT scans may be used to assess sphenoid sinus anatomy.  Angiography
```

```
    may be considered for very large or invasive tumors to assess relationship to
    major vessels.
* **Hormonal Evaluation:**  Complete endocrine workup to assess pituitary function
    and identify hormone excess.
* **Visual Field Testing:**  To document pre-operative visual function, especially
    if the tumor is compressing the optic chiasm.
* **Patient Counseling:**  Detailed discussion of risks, benefits, alternatives,
    and expected recovery.
* **Medical Optimization:**  Management of any co-morbidities and potential
    cessation of blood thinners.



**II. Intraoperative Steps:**

**1. Anesthesia & Positioning:**

* **General Anesthesia:**  The patient is placed under general anesthesia.
* **Positioning:** Typically a supine position with the head slightly elevated and
    fixed in a three-point head holder.  This allows for optimal visualization and
    access.  Sometimes a slight reverse Trendelenburg position is used.

**2. Nasal Preparation & Endoscopic Access:**

* **Nasal Decongestion:**  Topical decongestants (like oxymetazoline) are used to
    shrink nasal mucosa.
* **Endoscopic Setup:** A rigid endoscope (typically 0- or 30-degree) is inserted
    through one nostril.  The surgeon operates by looking at a monitor displaying
    the endoscopic view.
* **Septal Mucosa Incision:** A small incision is made in the nasal septum (the
    cartilage dividing the nostrils). This is often unilateral (one side) but can
    be bilateral if needed for wider access.
* **Sphenoid Sinus Entry:** The posterior nasal septum is removed to expose the
    sphenoid sinus.  The sphenoid sinus is then widely opened.  This often
    involves removing any septations or disease within the sinus.  Image guidance
    (see below) is crucial here.

**3. Sellar Floor Exposure & Opening:**

* **Sellar Floor Identification:** The surgeon identifies the bony sphenoid planum
    (sellar floor), which is the bone separating the sphenoid sinus from the sella
    turcica (the bony cavity housing the pituitary gland).
* **Sellar Floor Removal:** Using a combination of micro-instruments (drills,
    curettes, and endoscopes), the surgeon carefully removes the bony sphenoid
    planum to create an opening into the sella turcica.  This is done under
    constant endoscopic visualization and often with the assistance of image
    guidance.  The dura (the membrane covering the brain) is usually visualized at
    this point.
* **Dural Opening:** A small incision is made in the dura mater to access the
    sella.  The size of the dural opening is tailored to the size of the tumor.

**4. Tumor Resection:**
```

```
* **Tumor Debulking:**  The surgeon begins to remove the tumor.  For softer
    tumors, this can be done with suction and micro-instruments.  For harder or
    fibrous tumors, a combination of techniques is used.
* **Internal Decompression:**  The tumor is often debulked from within, starting
    with the central portion.
* **Capsule Dissection:**  The tumor capsule is carefully dissected from the
    surrounding structures the pituitary gland, optic chiasm, cavernous sinuses,
    and sphenoid sinus walls.  This is the most critical and delicate part of the
    surgery.
* **Complete Resection (Goal):** The goal is to achieve gross total resection
    (GTR) of the tumor, meaning all visible tumor is removed.  However, this isn't
    always possible or safe, especially for invasive tumors.
* **Suction Irrigation:**  Continuous irrigation and suction are used throughout
    the resection to maintain a clear surgical field.

**5. Reconstruction & Closure:**

* **Dural Repair:**  The dural opening is meticulously repaired.  This is often
    done with a combination of:
* **Fat Graft:**  Fat harvested from the abdomen is used to provide a watertight
    seal.
* **Dural Substitutes:**  Collagen matrices or other synthetic materials may be
    used in conjunction with fat.
* **Fibrin Glue:**  Fibrin glue is often applied to reinforce the dural repair.
* **Sphenoid Sinus Reconstruction (Optional):**  In some cases, the sphenoid sinus
    is partially reconstructed with a flap or graft to prevent cerebrospinal fluid
    (CSF) leaks.
* **Nasal Packing:**  Nasal packing (e.g., with absorbable material or silicone
    splints) is placed to support the nasal mucosa and promote healing.



**III.  Important Adjuncts:**

* **Image Guidance (Neuronavigation):**  This is *crucial*.  It uses pre-operative
    MRI or CT scans to create a 3D map of the patient's anatomy, allowing the
    surgeon to precisely locate the tumor and surrounding structures during
    surgery.
* **Intraoperative Neuromonitoring (IONM):**  This may be used for larger or more
    complex tumors, especially those near the optic nerves.  It involves
    monitoring the function of the optic nerves and brainstem during surgery to
    detect any potential injury.
* **Fluorescence-Guided Surgery:**  Some surgeons use fluorescent dyes (like
    5-ALA) to help differentiate tumor tissue from normal pituitary tissue,
    improving the extent of resection.



**IV. Postoperative Care:**
```

```
* **ICU Monitoring:**  Patients are typically monitored in the intensive care unit
    (ICU) for 1-2 days.
* **Hormone Replacement:**  Postoperative hormone deficiencies are common and
    require hormone replacement therapy (e.g., cortisol, thyroid hormone, growth
    hormone).
* **CSF Leak Monitoring:**  Close monitoring for CSF leaks is essential.
* **Follow-up MRI:**  A follow-up MRI is performed within a few weeks to assess
    the extent of resection and rule out any complications.
* **Endocrine Follow-up:**  Regular endocrine follow-up is necessary to monitor
    hormone levels and adjust hormone replacement therapy as needed.




**Disclaimer:** This is a general overview and should not be considered medical
    advice.  The specific steps and techniques used in an EETS procedure can vary
    depending on the individual patient and the surgeon's experience.  Always
    consult with a qualified neurosurgeon for any medical concerns.
```

Question 3

Compare microscopic versus endoscopic approaches for pituitary surgery.

Gemma 3 27B Response

```
## Microscopic vs. Endoscopic Pituitary Surgery: A Detailed Comparison

Pituitary surgery aims to remove pituitary tumors (adenomas) that cause hormonal
    imbalances or compress surrounding structures like the optic nerves.  Both
    microscopic and endoscopic approaches are common, but they differ
    significantly in technique, advantages, and disadvantages. Here's a detailed
    comparison:

**1. Surgical Approach & Technique:**

* **Microscopic Surgery (Transsphenoidal):**
* **Access:** Traditionally, this involves a surgical incision inside the nose
    (transsphenoidal approach) to reach the sphenoid sinus, a cavity behind the
    nose.  A speculum is used to retract tissues, and a high-powered surgical
    microscope is essential for visualization.
* **Visualization:** Relies heavily on the surgeon's skill and the magnification
    provided by the microscope.  The surgeon directly views the pituitary gland
    and tumor through the microscope.
* **Instrumentation:** Uses specialized micro-instruments manipulated directly by
    the surgeon under microscopic guidance.
* **Sinus Involvement:**  Often requires significant opening of the sphenoid sinus
    to gain adequate access.
* **Endoscopic Surgery (Transsphenoidal):**
* **Access:** Also uses a transsphenoidal approach, but instead of a microscope
    and speculum, a thin, flexible endoscope (a tube with a camera and light
```

```
    source) is inserted through the nostrils.
* **Visualization:** Provides a wide-angle, high-definition view on a monitor.
    The surgeon views the surgical field on the monitor, manipulating instruments
    through channels in the endoscope.
* **Instrumentation:** Uses specialized endoscopic instruments that are smaller
    and more maneuverable than microscopic instruments.
* **Sinus Involvement:**  Generally requires less sphenoid sinus opening than
    microscopic surgery, potentially minimizing nasal complications.



**2. Advantages & Disadvantages:**

| Feature | Microscopic Surgery | Endoscopic Surgery |
|---|---|---|
| **Visualization** | Excellent depth perception, tactile feedback.  Surgeon
    directly views the field. | Wide-angle view, high definition.  Can visualize
    around corners.  Less depth perception. |
| **Surgical Precision** | Highly precise, especially for complex tumors.  Surgeon
    has direct control and feel. |  Precision improving rapidly with advancements
    in endoscope technology and surgeon experience. |
| **Tumor Size** | Well-suited for a wide range of tumor sizes, including larger
    and more complex tumors. |  Historically better for smaller, less complex
    tumors.  Increasingly used for larger tumors with experienced surgeons. |
| **Sinus Complications** | Higher risk of nasal congestion, sinusitis,
    cerebrospinal fluid (CSF) leak, and septal perforation due to wider sphenoid
    sinus opening. | Lower risk of these complications due to less sinus
    disruption. |
| **Nasal Morbidity** |  More significant nasal packing often required
    post-operatively, leading to discomfort. |  Less or no nasal packing often
    needed, leading to faster recovery and less discomfort. |
| **Recovery Time** |  Generally longer recovery due to more extensive sinus
    manipulation and nasal packing. |  Generally faster recovery with less pain
    and quicker return to normal activities. |
| **Learning Curve** |  Steeper learning curve for surgeons initially, but
    well-established technique. |  Steeper learning curve for surgeons
    transitioning from microscopic to endoscopic techniques. Requires specialized
    training. |
| **Cost** |  Can be slightly less expensive initially (equipment costs are
    established). |  Can be more expensive due to the cost of endoscopes and
    specialized instruments. |
| **Surgeon Experience** |  Widely available with many experienced surgeons. |
    Availability of highly experienced endoscopic pituitary surgeons is growing
    but may be limited in some areas. |
| **Anterior Pituitary Preservation** |  Generally good, but can be challenging
    with larger tumors. |  Potentially better preservation of normal anterior
    pituitary function due to less manipulation. |



**3. Specific Tumor Types & Considerations:**
```

```
* **Microadenomas (<10mm):**  Both approaches are effective. Endoscopic is often
    preferred due to faster recovery and lower morbidity.
* **Macroadenomas (>10mm):**  Microscopic surgery has traditionally been the gold
    standard, especially for complex macroadenomas with significant extension.
    However, experienced endoscopic surgeons are increasingly successful with
    these tumors.
* **Aggressive Tumors (e.g., invasive adenomas):** Microscopic surgery may be
    favored for more extensive resection and better control.
* **Recurrent Tumors:**  Endoscopic approach can be useful for accessing residual
    tumor tissue after previous microscopic surgery.
* **Craniopharyngiomas:**  Both approaches can be used, but endoscopic may be
    preferred for smaller lesions and to minimize brain retraction.
* **Cholesterol Granulomas:** Endoscopic approach is often preferred for these
    cysts due to its ability to effectively decompress the cyst and preserve
    pituitary function.




**4.  Current Trends & Future Directions:**

* **Increasing adoption of endoscopic surgery:**  Endoscopic techniques are
    becoming increasingly popular and are now considered the preferred approach by
    many surgeons, particularly for smaller and less complex tumors.
* **Hybrid approaches:**  Some surgeons utilize a combination of microscopic and
    endoscopic techniques to leverage the advantages of both.
* **Robotic assistance:**  Robotic-assisted endoscopic pituitary surgery is
    emerging as a potential future direction, offering even greater precision and
    dexterity.
* **Improved endoscope technology:**  Advancements in endoscope resolution,
    illumination, and instrument design are continually improving the capabilities
    of endoscopic surgery.




**In conclusion:**

Both microscopic and endoscopic approaches are viable options for pituitary
    surgery. The best approach depends on several factors, including the **size
    and type of tumor, the surgeon's experience, and the patient's overall
    health.**

**It's crucial to discuss the risks and benefits of each approach with a
    neurosurgeon specializing in pituitary surgery to determine the most
    appropriate treatment plan.**



**Disclaimer:** I am an AI chatbot and cannot provide medical advice. This
    information is for general knowledge and informational purposes only, and does
    not constitute medical advice. It is essential to consult with a qualified
```

```
healthcare professional for any health concerns or before making any decisions
related to your health or treatment.
```

Question 4

Compare the use of a Rhoton dissector with suction or curettes during pituitary adenoma resection. When is each preferred?

Gemma 3 27B Response

```
## Rhoton Dissector vs. Suction/Curettes in Pituitary Adenoma Resection: A
    Comparison

Both the Rhoton dissector and suction/curettes are crucial tools in
    transsphenoidal pituitary adenoma resection, but they serve different purposes
    and are used in different phases of the surgery. Here's a detailed comparison,
    including when each is preferred:

**1. Rhoton Dissector**

* **What it is:** A thin, blunt instrument with a rounded tip, typically made of
    titanium or a similar material. It comes in various sizes and angles.  It's
    designed for *dissection*---gently separating the tumor from surrounding
    structures.
* **Mechanism of Action:**  Works by creating a plane between the tumor and the
    pituitary gland, cavernous sinus, optic chiasm, and dura. It relies on blunt
    dissection, minimizing trauma.
* **Advantages:**
* **Precise Dissection:** Excellent for delicate dissection around critical
    neurovascular structures (optic chiasm, internal carotid artery).
* **Minimizes Trauma:**  Blunt dissection reduces the risk of bleeding and damage
    to surrounding tissues compared to sharp instruments or aggressive curettage.
* **Visualization:**  Helps to define the tumor capsule and identify the optimal
    plane for removal.
* **Soft/Cystic Tumors:** Particularly useful for softer, cystic adenomas where
    suction alone may not be effective.
* **Capsule Dissection:**  Essential for debulking and defining the tumor capsule,
    especially for larger or invasive adenomas.
* **Disadvantages:**
* **Time-Consuming:**  Blunt dissection is slower than curettage.
* **Requires Skill:**  Effective use requires significant surgical experience and
    a thorough understanding of anatomy.
* **Limited for Hard Tumors:** Less effective for very firm, fibrous adenomas.
* **Can't Remove Fluid/Debris:** Doesn't actively remove tumor material; it
    *exposes* it for removal by other means.

**2. Suction & Curettes**

* **What they are:**
```

```
* **Suction:**  A device that uses negative pressure to remove fluid, blood, and
    soft tissue fragments.  Often uses a Frazier tip or similar.
* **Curettes:**  Loop-shaped instruments (various sizes and shapes - e.g., Ball,
    Luer, Janzen) used to scrape and remove soft tissue.  Can be rigid or flexible.
* **Mechanism of Action:**
* **Suction:** Removes fluid and soft debris, improving visualization.
* **Curettes:**  Mechanically remove tumor tissue by scraping.
* **Advantages:**
* **Efficient Removal:**  Rapidly removes large volumes of soft tumor tissue.
* **Good for Soft/Friable Tumors:**  Effective for adenomas that are easily
    fragmented.
* **Hemostasis:** Suction can help control bleeding by removing blood from the
    surgical field.
* **Debulking:**  Quickly reduces tumor size, facilitating further dissection.
* **Disadvantages:**
* **Risk of Trauma:**  Aggressive curettage can damage the pituitary gland, optic
    chiasm, or cavernous sinus.
* **Poor Visualization:**  Can create a bloody field, obscuring visualization.
* **Not Ideal for Encapsulated Tumors:**  May not effectively remove tumors with a
    well-defined capsule without first being dissected.
* **Risk of CSF Leak:**  Aggressive curettage of the sella can increase the risk
    of cerebrospinal fluid leak.



**When is each preferred? A Phased Approach & Specific Scenarios**

The use of these tools is often *sequential* and *complementary* during a
    pituitary adenoma resection.  Here's a breakdown:

**Phase 1: Initial Entry & Tumor Exposure**

* **Rhoton Dissector:**  Primarily used for the initial entry into the sella,
    opening the sphenoid sinus, and identifying the dura.  It's also used to
    initially define the tumor margins and begin dissecting the tumor from the
    pituitary gland.
* **Suction:** Used to remove any blood or debris encountered during the initial
    entry and to maintain a clear surgical field.

**Phase 2: Tumor Debulking & Dissection**

* **Rhoton Dissector:**  Crucial for dissecting the tumor from the optic chiasm,
    cavernous sinus, and pituitary gland.  This is especially important for:
* **Optically Important Tumors:**  Tumors that are close to or compressing the
    optic chiasm.  Gentle dissection with the Rhoton is paramount.
* **Invasive Adenomas:**  Tumors that extend beyond the sella.
* **Larger Adenomas:**  To define the tumor capsule and create a plane for removal.
* **Suction & Curettes:** Used *after* initial dissection with the Rhoton.
* **Soft, Friable Tumors:** Curettes are used to remove the bulk of the tumor
    tissue that has been exposed by the Rhoton dissector. Suction simultaneously
    removes debris and maintains visualization.
* **Cystic Components:** Suction is excellent for removing cystic fluid.
```

Phase 3: Final Removal & Pituitary Reconstruction

* **Rhoton Dissector:** Used for meticulous removal of any remaining tumor fragments, especially in areas near critical structures.
* **Suction & Curettes:** Used to remove any residual tumor tissue and to carefully curette the sella floor to ensure complete removal. However, curettage is used more cautiously in this phase to minimize the risk of pituitary damage or CSF leak.

Specific Scenarios & Preferences:

Scenario	Preferred Tool(s)	Rationale
Small, Soft Prolactinoma	Suction & Curettes (primarily)	Often easily fragmented and removed with minimal dissection.
Large, Invasive Prolactinoma	Rhoton Dissector + Suction/Curettes	Requires careful dissection from surrounding structures before debulking.
Non-Functioning Adenoma with Chiasm Compression	Rhoton Dissector (primary) + Suction/Curettes	Optic chiasm preservation is paramount; gentle dissection is essential.
Firm, Fibrous Adenoma	Rhoton Dissector + Micro-instruments	Requires more meticulous dissection; curettage may be less effective.
Recurrent Adenoma	Rhoton Dissector + Suction/Curettes	Often involves scar tissue, requiring careful dissection.
Cystic Adenoma	Suction (primary) + Rhoton Dissector	Suction effectively removes the cyst fluid, and the Rhoton helps define the capsule.

Important Considerations:

* **Endoscopic vs. Microscopic Approach:** The specific techniques and tool preferences can vary slightly depending on whether the surgery is performed using an endoscope or a microscope. Endoscopes often allow for more angled visualization and may favor certain Rhoton dissector shapes.
* **Surgeon's Experience:** The surgeon's skill and preference play a significant role in the choice of instruments and techniques.
* **Intraoperative Monitoring:** Neurophysiological monitoring (e.g., visual evoked potentials) can guide the surgeon's dissection and minimize the risk of damage to critical structures.

Disclaimer: I am an AI chatbot and cannot provide medical advice. This information is for general knowledge and informational purposes only, and does not constitute medical advice. It is essential to consult with a qualified healthcare professional for any health concerns or before making any decisions related to your health or treatment.

Question 5

What anatomical landmarks are used to identify the sella turcica during transsphenoidal pituitary surgery, and why are they critical for avoiding complications?

Gemma 3 27B Response

```
## Anatomical Landmarks for Sella Turcica Identification in Transsphenoidal
    Pituitary Surgery & Why They Matter

Transsphenoidal pituitary surgery is a delicate procedure requiring precise
    anatomical knowledge. Identifying the sella turcica (Turkish saddle) and its
    surrounding structures is *critical* for successful surgery and minimizing
    complications. Here's a breakdown of the key landmarks and their importance:

**I. External Landmarks (Pre-operative Imaging & Intraoperative Guidance)**

These are identified primarily through pre-operative imaging (CT and MRI) and then
    confirmed intraoperatively using fluoroscopy, neuronavigation, and direct
    visualization.

* **Dorsum Sellae:** This is the posterior wall of the sella, formed by the
    posterior clinoid processes. It's the "back" of the saddle.  *Importance:*
    Defining the dorsum sellae helps establish the posterior limit of the surgical
    field.  Overly aggressive removal of the dorsum can lead to cerebrospinal
    fluid (CSF) leak and damage to the pituitary stalk.
* **Tuberculum Sellae:** The anterior projection of the sella, formed by the
    anterior clinoid processes. It's the "front" of the saddle. *Importance:*
    Helps define the anterior boundary.  Fracturing the tuberculum sellae can
    cause bleeding and potentially damage the optic nerves/chiasm.
* **Clivus:** The bony base of the skull anterior to the pons. The sella sits *on*
    the clivus. *Importance:*  The clivus provides a key reference point for the
    overall orientation and depth of the surgical approach.  Understanding its
    contours is vital to avoid injury to the brainstem.
* **Sphenoid Sinus:** The air-filled cavity *below* the sella. The surgeon
    accesses the sella by opening the sphenoid sinus. *Importance:*  The size and
    pneumatization (air-filling) of the sphenoid sinus vary significantly between
    individuals.  Knowing its anatomy is crucial for safe and effective sinus
    entry.  Variations like concha (bony shelves) within the sinus need to be
    identified to avoid damage.
* **Carotid Groove:**  Located laterally and inferior to the sella, housing the
    internal carotid arteries. *Importance:*  This is arguably the *most critical*
    landmark. The internal carotid arteries are immediately adjacent to the sella
    and are at high risk of injury during surgery.  Identifying the carotid groove
    and its relationship to the sella is paramount.
* **Optic Canal & Optic Strut:**  The bony canal through which the optic nerves
    pass, located superolaterally to the sella. The optic strut is the bony
    support between the optic canal and the sella. *Importance:*  The optic nerves
    and chiasm lie *above* the sella and are vulnerable to compression or injury.
    The optic canals and strut define the superior limits of the surgical field
    and help protect the visual pathways.
```

```
* **Superior Sella Rim:** The bony edge of the sella superiorly. *Importance:*
    Helps define the superior limit of the sella and the proximity of the optic
    structures.
* **Foramen Ovale:** Located posterolaterally, containing the accessory meningeal
    artery. *Importance:*  While not directly within the sella, it's a nearby
    structure that can be encountered during surgery and a potential source of
    bleeding.



**II. Intraoperative Landmarks (Once the Sphenoid Sinus is Entered)**

Once the surgeon has entered the sphenoid sinus, these landmarks become visible:

* **Sellar Floor:** The bony floor of the sella, which the surgeon must remove to
    access the pituitary gland. *Importance:*  Careful removal of the floor is
    essential to avoid dural tears and CSF leaks.
* **Dural Ring:** The circular opening in the dura mater covering the sella.
    *Importance:*  The dural ring is incised to gain access to the subarachnoid
    space and the pituitary gland.  The incision must be carefully planned to
    minimize tension and prevent CSF leaks.
* **Pituitary Gland:** The target of the surgery, usually appearing as a soft,
    pinkish-gray mass within the sella. *Importance:*  Visualizing the gland and
    any associated tumor is crucial for accurate resection.
* **Pituitary Stalk:** The connection between the pituitary gland and the
    hypothalamus. *Importance:*  Preserving the pituitary stalk is often desirable
    to maintain residual pituitary function.
* **Cavernous Sinus:** Located laterally to the sella, containing the internal
    carotid artery, abducens nerve (CN VI), and oculomotor nerve (CN III).
    *Importance:*  The cavernous sinus is a critical landmark for avoiding injury
    to the internal carotid artery and cranial nerves.



**Why are these landmarks critical for avoiding complications?**

Complications in transsphenoidal pituitary surgery can be devastating.  Precise
    anatomical knowledge and identification of these landmarks are essential to
    prevent:

* **Cerebrospinal Fluid (CSF) Leak:**  Damage to the dura mater or inadequate
    dural repair can lead to CSF leakage, potentially causing meningitis.
* **Internal Carotid Artery Injury:**  The internal carotid artery is in close
    proximity to the sella. Injury can cause stroke or massive hemorrhage.
* **Optic Nerve/Chiasm Damage:**  Compression or injury to the optic nerves/chiasm
    can result in vision loss (blindness).
* **Hypopituitarism:**  Damage to the pituitary gland or pituitary stalk can lead
    to hormone deficiencies.
* **Cranial Nerve Palsies:**  Injury to cranial nerves within the cavernous sinus
    (CN III, CN VI) can cause double vision or other neurological deficits.
* **Bleeding:**  Damage to blood vessels during surgery can lead to significant
    bleeding.
```

```
**Tools to Aid Landmark Identification:**

* **Preoperative Imaging (CT & MRI):**  Provides a detailed roadmap of the sella
    and surrounding structures.
* **Fluoroscopy:**  Real-time X-ray imaging used intraoperatively to confirm the
    position of instruments and identify bony landmarks.
* **Neuronavigation:**  A computer-assisted surgical planning and guidance system
    that uses pre-operative imaging to create a 3D map of the surgical field.
    This allows the surgeon to precisely locate anatomical structures during
    surgery.
* **Endoscopic Visualization:**  Provides a magnified view of the surgical field,
    allowing for better identification of anatomical landmarks.



In conclusion, successful and safe transsphenoidal pituitary surgery relies
    heavily on a thorough understanding of the sella turcica's anatomy and
    meticulous identification of its surrounding landmarks.  The use of advanced
    imaging and surgical guidance technologies further enhances the surgeon's
    ability to navigate this complex anatomical region and minimize the risk of
    complications.
```