

SUREON: A Benchmark and Vision-Language Model for Surgical Reasoning

Alejandra Perez^{1,*} Anita Rau^{1,*} Busisiwe Mlambo¹ Lee White¹ Aina Brunet²
Juan Verde^{3,4} Juan Caicedo⁵ Joël L. Lavanchy^{6,7} Adam Schmidt¹
Chinedu Nwoye¹ Omid Mohareri^{1,†} Muhammad Abdullah Jamal^{1,†}

¹Intuitive Surgical Inc., USA ²Hospital Universitari Bellvitge, Spain ³IHU Strasbourg, France

⁴University of Strasbourg, CNRS, France ⁵Fundación Santa Fe de Bogotá, Colombia

⁶University Digestive Health Care Center – Clarunis, Switzerland ⁷University of Basel, Switzerland

{alejandra.perezrondon, anita.rau}@intusurg.com

<https://aperezr20.github.io/sureon>

Abstract

A capable surgical AI model should reason from visual evidence; yet, many existing surgical benchmarks can be partially solved without visual input, as models may exploit procedural priors or common anatomy–instrument interactions to infer answers from language alone. To address this, we introduce SUREON, a video question-answering (QA) dataset and benchmark designed to evaluate visually grounded surgical reasoning. Our pipeline generates QA pairs from narrated surgical academic videos, refines incorrect options using multiple models to reduce linguistic biases, and uses trained adversarial guessers to filter out questions answerable without visual evidence. The resulting dataset contains 76K training pairs covering 168 procedures and 10 specialties, alongside a benchmark of 340 questions across 10 categories verified by expert surgeons. Benchmarking leading commercial and open-source models shows that surgical visual reasoning remains largely unsolved—no model exceeds 53% accuracy and even expert surgeons reach only 64.9%. We also introduce SUREON-VLM, a baseline trained on our curated data that avoids language shortcuts, outperforms existing open-source models on SUREON, and matches larger specialist models on multiple standard surgical benchmarks. Our training ablations show that distractor refinement and blind curation strengthen visual grounding, increasing the multimodal gain from 8.7 to 29.1 percentage points while reducing the training set from 176.6K to 76K filtered samples. SUREON thus offers both a verified testbed for surgical visual reasoning and training data that strengthens visual grounding.

1. Introduction

Surgeons perceive complex visual cues in the operative scene and constantly interpret their clinical meaning, such as whether a dissection plane is safe or what the current tissue state implies for the next step. As AI systems are increasingly explored towards intraoperative decision support and surgical autonomy [6, 47, 53], a growing body of work has turned from traditional perception models to Vision Language Models (VLMs) to capture expert-like surgical understanding and reasoning [46, 61, 76]. Yet existing benchmarks fall short of evaluating the visually grounded reasoning that defines surgical expertise, restricting progress on models with such capabilities.

In the general domain, and increasingly in diagnostic medical fields, evaluation of VLMs has progressed from classification, recognition, and localization tasks to compositional and causal reasoning problems on visual inputs [10, 17, 68, 69, 78]. But unlike radiology and pathology, where image–report pairs already exist [16, 28], surgical video is recorded ad hoc without paired reports. The field has therefore leaned on rigid expert taxonomies, and the benchmarks built on them inherit those closed label spaces [18, 34, 46, 50, 56, 76]. Moving beyond closed label spaces requires supervision that captures how surgeons describe visual evidence, uncertainty, and clinical implications in natural language, which is hard to obtain at scale.

A particularly rich and scalable source of supervision for surgical VLMs is expert-narrated video lectures that are available for training and communication of best practices. Previous work converts these lectures into open-ended instruction data to train generative surgical assistants [31, 32, 60]. Such pipelines scale up data generation by leveraging state-of-the-art (SOTA) Large Language

^{*,†}Equal contribution.

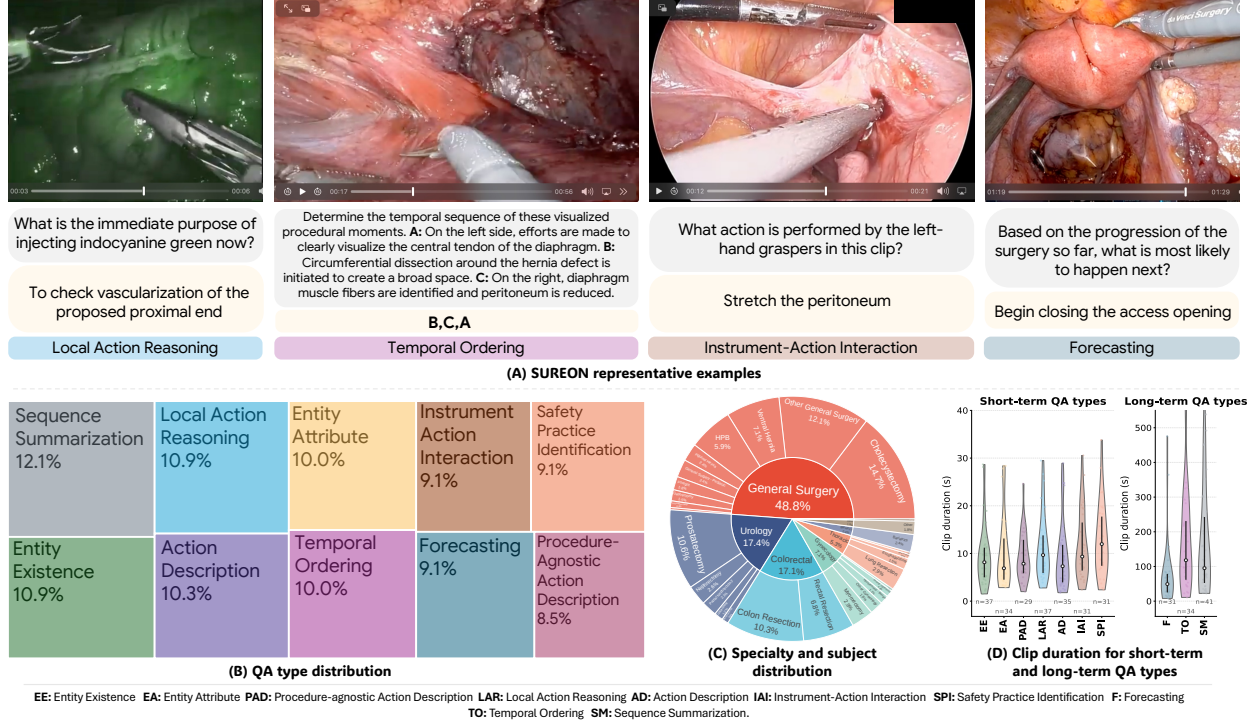


Figure 1. **Overview of SUREON Benchmark.** (A) Representative examples for some VQA types (B) VQA type distribution. (C) Hierarchical distribution of specialties and surgical subjects. (D) Clip-duration distributions per QA type, grouped by short-term and long-term categories.

Models (LLMs), but they provide limited verification that benchmark questions indeed require visual evidence. It remains unclear whether a high score reflects visual reasoning or just memorized surgical facts—a shortcut recent studies have documented in surgery [54] and beyond [4]. Video lectures are the rare setting in which visual cues and their clinical meaning are made explicit in language; a benchmark built from them should thus evaluate visual reasoning grounded in the surgical scene, not medical knowledge.

The challenge, therefore, is not to generate surgical QA pairs at scale, but to verify that answering them requires vision. Standard multimodal benchmarking construction delegates this task to state-of-the-art LLMs: they generate questions, answers, and distractors, judge sample quality, and grade the responses of competing models [33, 35, 63]. In surgery, every link of this chain is circular. Existing multimodal models perform poorly on visual surgical tasks, and a model cannot grade an exam it would fail itself. The fallback, routing every sample through surgeons, does not scale: expert annotation is the field’s long-standing bottleneck. But skipping verification has a cost: in the general domain, large fractions of popular multimodal benchmarks have been shown to be solvable by text-only models, defeating the purpose of multimodal evaluation [15].

Rather than entrusting one general-domain model to gen-

erate, judge, and grade its own exam in a domain where it performs poorly, we propose a multi-stage agentic pipeline that yields **SUREON**, a video QA dataset for surgical visual reasoning. Our data generation pipeline first identifies *Semantic Grounding Moments* (SGMs), time intervals where the surgeon explicitly ties a semantic cue to the visual scene, and converts them into QA pairs spanning ten question types including local action reasoning, forecasting, and safety. Second, retrieval-augmented, multi-model distractor refinement aligns generated distractors with the style of correct answers to minimize statistical biases and mixes LLM generators from different families to remove single-model artifacts. Third, a blind-testing protocol discards samples that an ensemble of finetuned and off-the-shelf LLM & VLM guessers answers consistently without the video, and an iterative hardening search applies further adversarial pressure to surviving test samples. Lastly, a team of six surgeons verify a subset of the final dataset.

We introduce **SUREON**, a resource that consists of over 70K rigorously curated video QA pairs for training (**SUREON-DB**) and 340 expert-reviewed samples for benchmarking. Extensive benchmarking shows that leading commercial and open-source models struggle to reason about visual cues in surgical videos. We further show that a model trained on **SUREON-DB** yields higher multimodal accuracy

gains with fewer training samples than naively generated surgical QA data as it is less reliant on language shortcuts. This model substantially outperforms existing open-source models, and is competitive with leading commercial systems.

2. Related Work

Surgical vision–language benchmarks. Surgical benchmarks for multi-modal understanding were first built by translating the annotations of existing datasets into language, reformulating phase, action, and tool labels into VQA pairs [22, 49, 61, 71]. Since these benchmarks were confined to a single procedure, other efforts have aggregated annotations from multiple annotated sources into cross-procedure benchmarks such as SurgVLM-Bench [76], EndoBench [34], and SurgMLLBench [18] to assess generalization across interventions and tasks. While such annotations give a broad, systematic representation of a scene’s observable semantics, these benchmarks stay bound to the rigid label sets of their sources, never reaching perception or reasoning beyond their fixed taxonomies. To mitigate these limitations, recent work has leveraged narrated surgical procedures to source supervision from expert lectures: SurgCoT [60] uses narration to probe reasoning such as causal action ordering and cue–action alignment, though its open-ended QA format restricts systematic scoring, while SurgPub-Video [32] structures narration-derived questions as multiple choice for unambiguous evaluation. However, none of these approaches filter out questions answerable from language priors alone—as is now standard in vision-language benchmark construction in other domains [10, 14, 66, 67]—nor do they evaluate models when no visual input is provided, despite evidence that Multi-modal Large Language Models (MLLMs) score deceptively high on visual perception benchmarks under blind testing [4, 44, 54]. Concurrent work SurgCheck [54] diagnoses this shortcut reliance via paired questions that replace entity names in the question with spatial grounding cues, but is a post-hoc diagnostic tool that requires scene-graph annotations and stays constrained within existing perception ontologies. SUREON instead addresses shortcut reliance at construction time by curating questions under blind-test conditions to encourage visual reliance during both training and evaluation and targeting visual reasoning capabilities beyond fixed label spaces.

Multimodal models in surgery. Data scarcity first made CLIP-style dual-encoder models predominant in surgery, pretrained on lecture narration through contrastive alignment and extended with hierarchical objectives and LLM-based text denoising [42, 72, 73]. Although these models can match a scene to a label, they cannot describe, explain, or reason. Generative surgical MLLMs followed to cover these needs, some trained on instruction data derived

Table 1. Comparison against existing surgical VL *datasets* (top) and *benchmarks* (bottom). SUREON contributes both a large-scale training database (SUREON-DB) and an expert-verified evaluation benchmark (SUREON), and supports perception and reasoning across diverse procedures and video modalities. **Cap.**: Captioning; **QA**: Question Answering; **P**: Perception **R**: Reasoning; **?**: blind-test curation.

Dataset / Benchmark	Format	#Proc. Types	#Clips	#VL Pairs	P	#VL Pair Types	R	?
<i>Training datasets</i>								
SVL [73]	Cap.	>1	25.5K	25.5K	✓	1	✗	✗
SurgLaVi [42]	Cap.	217	240K	240K	✓	3	✗	✗
SurgLaVi-β [42]	Cap.	170	112.7K	112.7K	✓	3	✗	✗
Surg-396K [62]	QA	3	—	396K	✓	7	✗	✗
Surg-QA [31]	QA	—	44K	102K	✓	5	✓	✗
SurgVLM-DB [76]	QA	16	—	7.79M	✓	10	✓	✗
SurgPub-Video [32]	QA	75	10.9K	48.5K	✓	5	✗	✗
SurgAtlas [9]	Cap.+QA	>5K	393K	793.2K	✓	10	✓	✗
SUREON-DB (Ours)	QA	168	63.1K	76.1K	✓	10	✓	✓
<i>Evaluation benchmarks</i>								
Cholec80-VQA [49]	QA	1	—	43K	✓	3	✗	✗
EndoVis-18-VQA [49]	QA	1	—	11.8K	✓	3	✗	✗
EndoVis-VQLA [7]	QA	1	—	9.5K	✓	4	✗	✗
PSI-AVA-VQA [50]	QA	1	—	10.3K	✓	3	✗	✗
CoPESD [61]	QA	1	—	17.7K	✓	2	✗	✗
SSG-VQA [71]	QA	1	—	960K	✓	5	✗	✗
SurgVLM-Bench [76]	QA	16	—	6K	✓	6	✓	✗
SurgPub-Video [32]	QA	75	—	3.3K	✓	5	✗	✗
SurgCoT [60]	QA	35	2.8K	78.5K	✓	5	✓	✗
SurgAtlas-Bench [9]	QA	—	—	2.96K	✓	10	✓	✗
SUREON (Ours)	QA	69	335	340	✓	10	✓	✓

from the classification and spatial annotations of existing datasets [50, 76] and others on narration-sourced supervision [32, 62], with recent work targeting reasoning rather than perception alone [56]. Concurrently, annotation- and narration-derived supervision have been integrated into a single training corpus with better generalization [9].

Automated VQA generation. Since manual annotation does not scale, recent benchmarks synthesize VQA pairs automatically from annotations or narrations with LLMs, typically in multiple-choice form for objective scoring [77]. However, naively generated options introduce *language shortcuts*: distractors leave statistical artifacts in length, specificity, or style that models exploit without visual grounding [65], and many questions stay answerable from textual priors alone, allowing that blind LLMs achieve high performance [15]. Two remedies have emerged to suppress this leakage: refining questions through chain-of-thought rewriting or length-matched distractors [10, 67], and removing text-answerable items via blind testing that discards any question a text-only LLM answers consistently across trials [52, 66, 74]. Such curation is often drastic: ExAct [67] retains only $\sim 0.9\%$ of its pool, for instance. These pipelines, however, rely almost exclusively on off-the-shelf LLMs prompted zero-shot, which underestimate exploitable leakage; a text-only model *fine-tuned* on the answer distribution can still solve items that such zero-

shot testers pass [4]. This fine-tuned-guesser signal has so far been used primarily to audit completed benchmarks rather than to curate samples during generation. In contrast, SUREON integrates distractor refinement, blind filtering, and fine-tuned guesser feedback into a single construction pipeline for surgical video QA.

3. SUREON Dataset, Benchmark, and Model

Moving beyond the closed-space recognition of predefined taxonomies, SUREON tests whether VLMs can reason over actions, states, safety-relevant events, and procedural progression in surgical video. Built from a heterogeneous pool of narrated minimally-invasive laparoscopic and robot-assisted procedures across multiple specialties, it comprises two complementary resources: **SUREON-DB**, a large and diverse training database of 76,061 automatically pseudo-labeled Video QA pairs for adapting VLMs to surgery, spanning 10 specialties and 168 procedure types, and **SUREON**, a 340-sample, surgeon-verified benchmark for rigorous assessment. This benchmark is highly diverse (8 specialties and 69 procedure types; fewer than DB as the benchmark is drawn from the smaller expert-verified validation split) and comparable in size with widely used benchmarks in other domains [20, 70, 75]. Unlike previous work (see Table 1), this data allows not only training surgical VLMs but also systematically assessing their visual reasoning capabilities. For more details, key attributes are summarized in Table 2.

SUREON unifies a broad set of capabilities within a single evaluation framework, organized into 10 categories across two temporal granularities (Figure 2). In *short temporal windows*, it evaluates recognition of anatomical structures, instruments, instrument-action interactions, entity states, and safety-driven actions, together with locally grounded action reasoning. In *longer contexts*, SUREON evaluates procedural reasoning through sequence summarization, temporal ordering, and prediction of the next procedural step. The first eight categories are posed as single-answer multiple-choice questions, while sequence summarization and temporal ordering are open-ended.

3.1. Dataset Construction

As illustrated in Figure 3, we introduce a four-stage automated pipeline to build SUREON. The pipeline addresses three challenges in surgical video-language dataset construction: (i) narrations are *selective* and do not densely describe all visible content, (ii) naively LLM generated multiple-choice distractors introduce *linguistic shortcuts* that can be exploited by other LLMs to successfully guess the correct answer, and (iii) some generated questions may be answerable from *prior procedural knowledge* without requiring visual evidence. To mitigate these, we first identify *Semantic Grounding Moments* (SGMs) from narration transcripts and use them as anchors for QA generation, then we

Table 2. Key attributes of SUREON database and benchmark.

Attribute	DB	Bench
Samples (QAs)	76,061	340
Labeling	auto, pseudo	expert-verified
Question categories	10	10
short-window / long-horizon	7 / 3	7 / 3
MCQ / open-ended	8 / 2	8 / 2
Surgical specialties	10	8
Subjects	37	27
Procedure types	168	69
Unique clips (temporal windows)	63,123	335
Source narrated videos	2,191	165
Expert surgeon reviewers	–	6
Median clip duration, short-window (s)	7.3	8.5
Median clip duration, long-horizon (s)	113.0	80.6

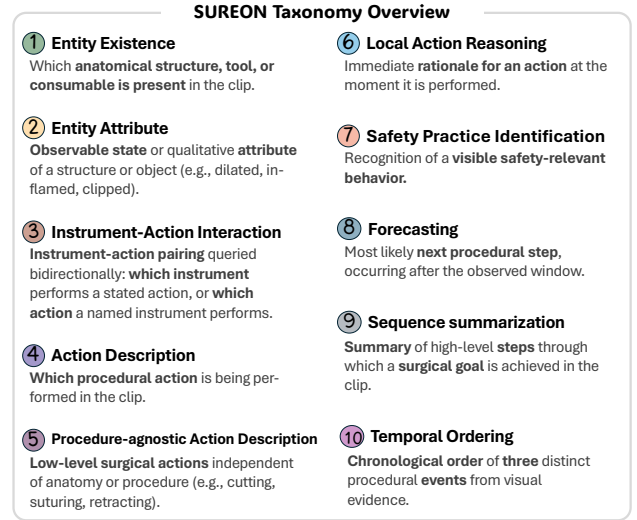


Figure 2. SUREON Question-Answer Types. Taxonomy definition for all ten QA types in the dataset

systematically refine multiple-choice distractors to mitigate linguistic biases, and finally curate the dataset by removing questions that are solvable without accessing the video, followed by expert verification to consolidate the benchmark.

Stage 1: Raw transcript-grounded QA synthesis

In lectures, surgeons describe predominantly clinically salient events, transitions, and safety-relevant observations, so transcripts alone do not offer dense visual supervision for VLM learning. We therefore anchor QA synthesis to *Semantic Grounding Moments* (SGMs) in the narration (examples in Figure 6 in the Appendix B.2.2).

Given a video v_i and its aligned transcript of time-stamped narrated sentences (formalized in Appendix B.2), each SGM, z_{ij}^k is a temporal window in which the narration provides task-specific supervision for a visually observable

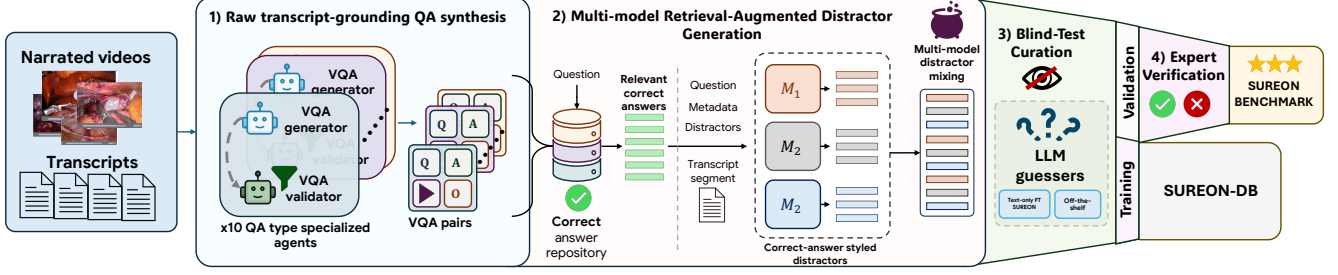


Figure 3. **The SUREON construction pipeline.** A four-stage automated pipeline converting narrated surgical videos into visually grounded QA. (1) Ten category-specialized generator-validator agent pairs locate *Semantic Grounding Moments* in the transcript and synthesize validated QA pairs. (2) Retrieval-augmented distractor generation (RADG) conditions three LLMs (M_1 – M_3) on retrieved exemplars of correct answers, and mixing across models yields distractors free of single-model shortcuts. (3) Blind-test curation discards samples that LLM *guessers* solve without the video, keeping only questions that require visual evidence (guessers include both an off-the-shelf and fine-tuned models on SUREON’s text distribution). (4) Surgeons verify the validation split for correctness and grounding. The curated training split forms the large-scale **SUREON-DB**; the expert-verified validation split forms the compact **SUREON BENCHMARK**.

surgical concept:

$$z_{ij}^k = (v_i[t_s, t_e], s_{ij}, c_k), \quad (1)$$

where $v_i[t_s, t_e]$ is the video clip, s_{ij} is the supporting transcript context, c_k is the target semantic category, and i, j, k index video, clip, and category.

To extract SGMs at scale and convert them into QA samples, we design a pipeline of K generator-validator agent pairs, one per semantic category in Figure 2. This decomposition follows the “divide and conquer” principle of agentic data synthesis [21, 38], in which role and category specialization yields higher data quality than prompting a single LLM. Formally, the generation-validation pipeline is described as follows:

- **Generation:** Given a transcript T_i , generator $\mathcal{A}_k$ identifies the intervals in which the surgeon explicitly describes content relevant to its assigned category and produces a set of instructional samples $\{x_{ij}^k\}_j$, each defined as:

$$x_{ij}^k = (v_i[t_s, t_e], Q_{ij}^k, A_{ij}^k, \mathcal{O}_{ij}^k), \quad (2)$$

where Q_{ij}^k is the question, A_{ij}^k the answer, and $\mathcal{O}_{ij}^k$ the set of multiple-choice options, when applicable.

- **Validation:** The paired validator filters the generator’s output by enforcing category-specific constraints alongside four universal criteria: hallucination filtering, visual grounding, temporal-window correctness, and semantic deduplication (detailed in Appendix B.2.3). The output of Stage 1 is a transcript-grounded set of surgical video QA pairs $\mathcal{D}_{\text{SGM}} = \{x_{ij}^k\}_{i,j,k}$.

Additional implementation details, agent system prompts, per-category agent specifications (Table 7), and the full per-category sample breakdown (Table 6) are provided in Appendix B.2.

Stage 2: Multi-Model Retrieval-Augmented Distractor Generation for Bias Mitigation

The naively generated distractors from Stage 1 introduce two compounding biases that let models guess correctly without video. A *distributional gap*, where correct answers and distractors differ in length, specificity, clinical detail, and phrasing, reveals which option *looks* correct without any visual grounding; and *model-specific artifacts*—distractors from a single LLM share syntactic and lexical patterns—provide a further shortcut that becomes especially exploitable during fine-tuning. We mitigate both with a two-component strategy: (i) Retrieval-Augmented Distractor Generation (RADG) to align the distractor distribution with that of correct answers, and (ii) multi-model mixing to suppress model-specific artifacts:

(i) **RADG.** To close the distributional gap between correct answers and distractors, we aggregate all correct answers from the training set into a typed retrieval pool for each question category, c :

$$\mathcal{P}_c = \{A_{ij}^k \mid c_k = c\}, \quad (3)$$

and encode each answer with a dense embedding model for retrieval. For each question Q_{ij}^k , we retrieve answer-style exemplars from $\mathcal{P}_{c_k}$, using video metadata to favor answers from procedures similar to the current one while preserving category-wide diversity, and merge them into a unified exemplar set E_{ij}^k (Appendix B.3.1). Each generator M_ℓ , instantiated as a state-of-the-art LLM, is then conditioned on the question, its correct answer, the local transcript context, and the exemplar set E_{ij}^k to produce distractors that are factually incorrect for the transcript context, clinically plausible, and stylistically aligned with the retrieved exemplars.

(ii) **Multi-model mixing.** Each of the three distractor-generator LLMs (M_1, M_2, M_3) produces N distractors, and the final set is drawn at random from the combined $3N$ -

candidate pool $\mathcal{D}_{ij}^{k,\text{pool}}$, which empirically reduces short-cut learning compared to employing distractors from a single model (ablation in Sec. B.3). The output of Stage 2 is the refined dataset $\mathcal{D}_{\text{ref}}$, in which each Stage-1 sample is paired with its refined distractor set $\mathcal{O}_{ij}^{k,\text{ref}}$.

Stage 3: Visual-Dependence Curation via Blind-Testing Filtering

Distractor refinement reduces linguistic artifacts in the answer options, yet some samples remain answerable without the video—either from residual language shortcuts, or because LLMs can resolve certain questions from procedural knowledge acquired during pretraining. The goal of Step 3 is to remove these samples. To this end, we employ a *blind-testing curation protocol*: an ensemble of LLM *guessers* $\mathcal{G}$ is queried without video access, receiving only the question Q_{ij}^k and the refined options $\mathcal{O}_{ij}^{k,\text{ref}}$. Any sample the *guessers* can solve is discarded. To separate genuine shortcut exploitation from lucky guesses, we query each guesser multiple times at non-zero temperature and score it by its *consistency*, the fraction of trials in which it recovers the correct answer, rather than by a single correct response, inspired by prior work [52, 66] (formalized in Appendix B.4.2).

To detect candidates answerable from prior knowledge alone, and to additionally account for language shortcuts that only become exploitable once a model is fine-tuned on the answer distribution, our guesser ensemble combines off-the-shelf and blind-fine-tuned models (full specification in Appendix B.4.2). We curate the training set for these fine-tuned guessers via stratified 4-fold cross-validation—guessers trained on the complementary folds and scored on the held-out fold—yielding the curated training set SUREON-DB. The validation split is evaluated by the guessers trained on the full training split and evaluated blindly, yielding $\mathcal{D}_{\text{val}}$.

As a final curation step for $\mathcal{D}_{\text{val}}$, we employ *distractor hardening via iterative blind testing*. We fine-tune a stronger blind guesser on SUREON-DB and, for each retained sample, iteratively resample distractor configurations from the pool $\mathcal{D}_{ij}^{k,\text{pool}}$ —subject to the cross-model diversity constraint—until the guesser fails to recover the correct answer under blind testing. The retained configuration yields the hardened validation set $\mathcal{D}_{\text{val}}^*$; samples for which no such configuration exists are discarded (Appendix B.4.3, Alg. 1).

Stage 4: Expert Verification

Samples from the hardened validation set $\mathcal{D}_{\text{val}}^*$ undergo *manual verification by 6 expert surgeons*. Each reviewer is presented with the video clip, the aligned transcript segment, the question, the correct answer, and all distractors, and accepts a sample only when its temporal alignment, answer correctness, visual grounding, and distractor validity all hold simultaneously (Appendix B.5.2). Any sample that is rejected or flagged is discarded under a conser-

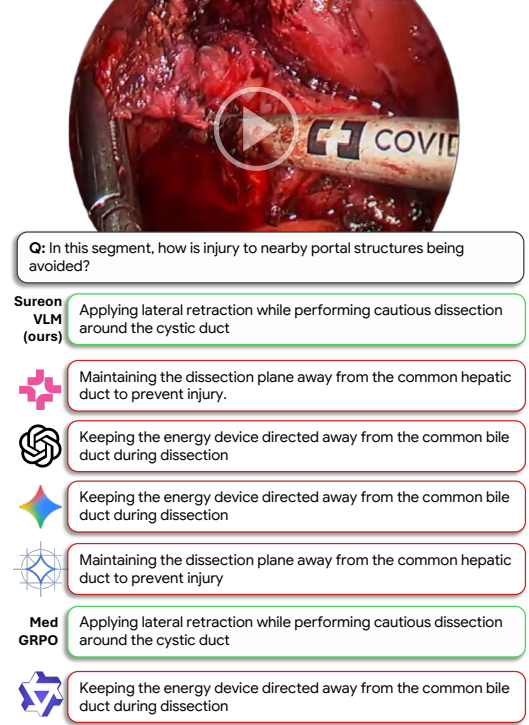


Figure 4. **Qualitative comparison** on a Safety Practice Identification question of SUREON-VLM, (✚) Molmo2, (🌀) GPT-5.1, (🌟) Gemini 3.1 Pro, (🌀) MedGemma 1.5, MedGRPO, and (🌀) Qwen3-VL-32B-Thinking. Green boxes denote correct responses and red boxes denote incorrect responses. *Options are presented along with the question, but are omitted for simplicity.*

vative retention policy, ensuring the final SUREON benchmark $\mathcal{B}_{\text{SUREON}} \subset \mathcal{D}_{\text{val}}^*$ satisfies both visual-dependence criteria and expert clinical judgment. A dual-reviewed subset is additionally used to assess inter-rater reliability (Appendix B.5.2).

3.2. Sureon-VLM

Beyond evaluation, SUREON-DB provides a large training dataset. To assess its value, we train a surgical VLM and use it to probe two questions. First, does the standard surgical-VLM recipe generalize to reasoning? Existing surgical assistants are trained by transforming classification-annotated surgery datasets into templated instructions augmented with LLMs [62, 76]; this supervision is mostly perceptual (presence, localization, phase), leaving open whether it extends to reasoning. And second, does consolidating public templated data with SUREON yield a single model strong at both perception and reasoning?

To answer these questions, we assemble a unified instruction-tuning corpus from two complementary sources: (i) image-level annotations from public datasets (tool and anatomy localization, localized action recognition,

and presence/state labels derived from them); and (ii) video clips from workflow annotations (phase, step, and action recognition). Following common practice, each annotation is mapped to a schema, rendered into natural-language QA via task-specific templates, and augmented with an LLM for linguistic diversity (Appendix C). This yields $\mathcal{D}_{\text{pub}}$ (5.5M QA pairs over 22 public datasets), which we combine with the SUREON training split into $\mathcal{D}_{\text{all}} = \mathcal{D}_{\text{SUREON-DB}} \cup \mathcal{D}_{\text{pub}}$.

We then fine-tune Qwen3-VL-8B-Instruct [8] on $\mathcal{D}_{\text{all}}$ using 8 frames per clip to obtain SUREON-VLM. We study data composition and staging order (single-stage vs. two-stage schedules) and select the best configuration; the full ablation is reported in Appendix Table 19.

4. Experiments

Experiments on the SUREON benchmark show that (i) no existing model can reliably answer complex questions about surgery, (ii) SUREON measures visual reasoning, not shortcut learning, and (iii) training on SUREON-DB does not lead to overfitting to a specific task, but we show that SUREON-VLM outperforms existing open-source VLMs on SUREON and has strong performance on traditional classification tasks in surgery.

We benchmark commercial general-purpose VLMs (GPT-5.1 [55] (high reasoning), Gemini 3.1 Pro [23] (high reasoning)); open generalist models (Qwen3-VL-8B-Instruct and Qwen3-VL-32B-Thinking [8], Gemma 4 [24], Molmo2 [19]); and open medical VLMs (MedGemma 1.5 [51], MedGRPO), and include a human expert baseline. We also evaluate SUREON-VLM, our model fine-tuned on $\mathcal{D}_{\text{all}}$, along with the same model trained on SUREON-DB alone or the public surgical corpus alone.

Of the ten SUREON question types, eight are multiple-choice and scored by accuracy. Temporal ordering (TO) is scored by exact-match accuracy, while sequence summarization (SM) is open-ended and scored by a validated LLM judge (Appendix D). We report the mean over the accuracy-based types and report SM separately. Additionally, we report multimodal gain [15] as a complementary metric assessing how much a model’s answers reflect the surgical video rather than language priors. It is defined as the accuracy difference between a model given the visual input (S_v) and the same model answering blind (S_{wv}): $\text{MG} = S_v - S_{wv}$. We provide additional evaluation and inference details in Appendix (D.1).

Visual reasoning in surgery is an unsolved task.

Table 3 reports per-category accuracy with 95% confidence intervals on the average. Gemini 3.1 Pro obtains the highest numerical mean ($\mu = 52.2$), followed by SUREON-VLM ($\mu = 48.2$) and GPT-5.1 ($\mu = 46.8$). However, the differences among the strongest models are

Table 3. **SUREON benchmarking results.** Each model is evaluated with image inputs (👁) and *blind* (👁, no visual input). SUREON is split into perception (AD, EE, EA, PAD, IAI) and reasoning (F, TO, SM, LAR, SPI), all reported as accuracy. μ is the pooled accuracy over all accuracy-scored items (matching the bootstrap CIs in Appendix D.1, Table 14); SM is scored with an LLM judge (*LLMJ*). MG (multi-modal gain) = $S_v - S_{wv}$ is reported per model.

Model	Eval	SUREON										
		μ	Perception					Reasoning				
			AD	EE	EA	PAD	IAI	LAR	SPI	F	TO	SM
Random		14.6	14.3	14.3	14.3	14.3	14.3	14.3	14.3	14.3	16.7	—
Human	👁	64.9 $\pm$ 5.3	74.3	51.3	73.5	96.5	74.2	35.1	71.0	54.8	61.8	—
<i>Commercial</i>												
GPT-5.1 (high) [55]	👁	46.8 $\pm$ 5.7	57.1	54.0	38.2	37.9	61.3	46.0	45.2	25.8	52.9	29.3
	👁	25.8 $\pm$ 4.9	17.1	35.1	38.2	20.7	32.3	18.9	22.6	16.1	29.4	12.2
	MG	+21.1	+40.0	+18.9	+0.0	+17.2	+29.0	+27.0	+22.6	+9.7	+23.5	+17.1
Gemini 3.1 Pro [23]	👁	52.2 $\pm$ 5.5	54.3	48.6	41.2	69.0	54.8	56.8	61.3	19.4	64.7	0.0
	👁	31.8 $\pm$ 5.2	14.3	37.8	38.2	24.1	22.6	59.5	32.3	9.7	41.2	9.8
	MG	+20.4	+40.0	+10.8	+2.9	+44.8	+32.3	-2.7	+29.0	+9.7	+23.5	-9.8
<i>Open-source generalist</i>												
Qwen3VL-32B-T [8]	👁	34.1 $\pm$ 5.3	40.0	27.0	41.2	62.1	45.2	27.0	19.4	12.9	35.3	9.8
	👁	27.8 $\pm$ 4.9	28.6	35.1	38.2	13.8	45.2	27.0	25.8	16.1	17.6	7.3
	MG	+6.3	+11.4	-8.1	+2.9	+48.3	+0.0	+0.0	-6.5	-3.2	+17.6	+2.4
Qwen3VL-8B-I [8]	👁	23.1 $\pm$ 4.7	28.6	16.2	17.6	34.5	38.7	18.9	16.1	12.9	26.5	0.0
	👁	13.4 $\pm$ 3.9	8.6	5.4	11.8	24.1	29.0	13.5	19.4	6.5	5.9	0.0
	MG	+9.7	+20.0	+10.8	+5.9	+10.3	+9.7	+5.4	-3.2	+6.5	+20.6	+0.0
Gemma 4 [24]	👁	35.8 $\pm$ 5.3	48.6	27.0	32.4	51.7	41.9	29.7	35.5	16.1	41.2	4.9
	👁	21.4 $\pm$ 4.5	11.4	21.6	29.4	34.5	25.8	21.6	16.1	9.7	23.5	4.9
	MG	+14.4	+37.1	+5.4	+2.9	+17.2	+16.1	+8.1	+19.4	+6.5	+17.6	+0.0
Molmo2 [19]	👁	15.0 $\pm$ 4.0	31.4	10.8	20.6	24.1	16.1	8.1	12.9	3.2	8.8	0.0
	👁	17.1 $\pm$ 4.2	17.1	8.1	29.4	24.1	25.8	5.4	16.1	6.5	23.5	0.0
	MG	-2.0	+14.3	+2.7	-8.8	+0.0	-9.7	+2.7	-3.2	-3.2	-14.7	+0.0
<i>Open-source medical</i>												
MedGemma 1.5 [51]	👁	10.7 $\pm$ 3.5	11.4	10.8	8.8	24.1	6.5	10.8	6.5	6.5	11.8	0.0
	👁	11.0 $\pm$ 3.5	14.3	8.1	8.8	0.0	6.5	18.9	6.5	6.5	26.5	0.0
	MG	-0.3	-2.9	+2.7	+0.0	+24.1	+0.0	-8.1	+0.0	+0.0	-14.7	+0.0
MedGRPO [56]	👁	16.7 $\pm$ 4.2	25.7	16.2	17.6	10.3	25.8	13.5	12.9	6.5	20.6	0.0
	👁	9.7 $\pm$ 3.3	2.9	21.6	2.9	6.9	12.9	5.4	3.2	3.2	26.5	0.0
	MG	+7.0	+22.8	-5.4	+14.7	+3.4	+12.9	+8.1	+9.7	+3.2	-5.9	+0.0
<i>Ours</i>												
SUREON-VLM (SUREON-only)	👁	44.1 $\pm$ 5.7	51.4	48.6	44.1	41.4	25.8	40.5	61.3	38.7	44.1	7.3
	👁	22.1 $\pm$ 4.7	31.4	16.2	23.5	6.9	19.4	16.2	35.5	32.3	17.6	2.4
	MG	+22.1	+20.0	+32.4	+20.6	+34.5	+6.5	+24.3	+25.8	+6.5	+26.5	+4.9
SUREON-VLM (Public-only)	👁	20.7 $\pm$ 4.5	42.9	16.2	14.7	34.5	29.0	10.8	19.4	6.5	14.7	2.4
	👁	12.0 $\pm$ 3.7	8.6	8.1	8.8	20.7	22.6	8.1	19.4	6.5	8.8	2.4
	MG	+8.7	+34.3	+8.1	+5.9	+13.8	+6.5	+2.7	+0.0	+0.0	+5.9	+0.0
SUREON-VLM	👁	48.2 $\pm$ 5.7	65.7	48.6	47.1	41.4	38.7	43.2	58.1	35.5	52.9	14.6
	👁	19.1 $\pm$ 4.3	22.9	8.1	20.6	6.9	25.8	13.5	32.3	29.0	14.7	2.4
	MG	+29.1	+42.9	+40.5	+26.5	+34.5	+12.9	+29.7	+25.8	+6.5	+38.2	+12.2

not statistically significant (Appendix D.1). In contrast, the human expert baseline reaches ($\mu = 64.9$) and significantly outperforms every evaluated model, indicating remaining headroom for surgical visual understanding. Expert performance also varies by category: surgeons approach ceiling on Procedure-Agnostic Action Description (96.5%), but perform substantially lower on Entity Existence (51.3%) and Local Action Reasoning (35.1%), likely reflecting specialty-specific anatomical knowledge and the added difficulty of inferring the purpose of an observed action. Models show a different pattern, with Forecasting among the hardest categories, suggesting that anticipating procedural progression remains harder than describing the current scene. Given the wider confidence intervals at the category level, we interpret these differences as broad

trends rather than fine-grained rankings.

SUREON measures visual reasoning. Across the strongest models, multimodal gain remains substantial: +29.1 for SUREON-VLM, +20.4 for Gemini 3.1 Pro, and +21.1 for GPT-5.1. Notably, SUREON-VLM achieves the largest MG, indicating that its benchmark performance depends strongly on the surgical video rather than on language priors. This visual dependence is progressively strengthened by the SUREON construction pipeline. Table 4 shows that MG increases from 8.7 with the raw Stage-1 data to 16.8 after retrieval-augmented distractor generation and to 29.1 after blind-test curation. The final improvement occurs while reducing the training set from 176.6K to 76K samples. Thus, stronger visual grounding results from removing supervision that rewards language shortcuts rather than simply increasing training-set size.

SUREON-VLM has broad capabilities. The surgical field’s data scarcity has pushed existing surgical VLMs toward classification-derived datasets. The data-composition ablations in Table 3 show that training on public surgical datasets alone performs poorly on SUREON ($\mu = 20.7$, $MG = 8.7$), whereas training on SUREON-DB alone reaches $\mu = 44.1$ with $MG = 22.1$. Both SUREON-only and combined training significantly outperform public-only training after multiple-comparison correction. When combining the two sources, performance raises SUREON from 44.1 to 48.2.

Public data nevertheless remains complementary. The full SUREON-VLM is trained on both sources so that reasoning-oriented supervision from SUREON-DB is combined with conventional surgical perception supervision. Table 5 shows that this produces a model that remains competitive on established surgical perception benchmarks while also performing strongly on SUREON. Despite its 8B backbone, SUREON-VLM matches or exceeds several larger specialist models on Cholec80 and Endoscapes-CVS, while clip-level inference further improves temporal recognition. The lower performance on HeiChole action recognition suggests that breadth still comes with trade-offs relative to models optimized for narrower tasks. Consolidating reasoning and perception supervision thus yields a single model that keeps the perception abilities of specialist systems while gaining the visual reasoning they lack (Table 3).

5. Conclusions

In this work, we introduce SUREON, a high-quality surgical visual reasoning benchmark that leverages expert narration in educational videos. Rather than naively generating VQA samples from lecture videos, which can introduce language-only shortcuts, SUREON combines distractor refinement, blind-test curation, and expert review to preserve questions that require visual evidence from the surgical video. As

Table 4. Benchmark construction ablation. SUREON-VLM trained on the distractors of Stage 1 (Sec. 3.1), Stage 2 (Sec. 3.1), and Stage 3 (Sec. 3.1) and evaluated on SUREON, with multimodal gain $MG = S_v - S_{wv}$ between visual (S_v) and blind (S_{wv}) accuracy. MG increases across stages, notably in Stage 3 with less than half the data.

Distractors	# training samples	S_v	S_{wv}	MG
Stage 1 (Baseline)	176.6K	22.4	13.7	8.7
Stage 2 (RADG)	176.6K	39.5	22.7	16.8
Stage 3 (Blind-Test Curation)	76K	48.2	19.1	29.1

Table 5. Comparison on public surgical benchmarks. (a) Macro-F1 (%) on phase and action recognition. (b) Cholec80 detailed metrics (%). (c) Endoscapes-CVS criterion accuracy (%). Inference is done at frame-level unless (**clip**) is specified.

(a) Macro F1 (%) on phase and action benchmarks.								
Method	Cholec80 [58]	HeiChole Phase [59]	HeiChole Action [59]					
GPT-5.1	36.0	29.0	18.0					
Gemini 3.1 Pro	47.0	35.0	21.0					
Qwen3-VL-8B	17.0	12.0	17.0					
SurgAtlas-VLM	68.2	54.5	55.8					
SUREON-VLM	55.33	47.1	14.5					
SUREON-VLM (clip)	78.85	71.4	clip eval not applicable					
(b) Cholec80 detailed.			(c) Endoscapes-CVS [39]					
Method	Acc	Rec.	Prec.	Method	Avg	C1	C2	C3
InternVL3-8B	23.9	15.2	15.2	InternVL3-8B	48.2	40.0	53.3	51.4
Qwen2.5-VL-7B	30.5	16.7	26.2	Qwen2.5-VL-7B	65.9	56.1	82.4	59.2
GPT-4o	36.4	31.0	33.0	GPT-4o	74.5	72.6	79.5	71.2
Gemini 2.0 Fl.	38.9	36.8	40.0	Gemini 2.0 Fl.	59.6	47.8	63.9	67.1
SurgVLM-7B	70.3	61.9	59.8	SurgVLM-7B	76.9	75.3	82.4	72.9
SurgVLM-72B	76.4	70.8	66.0	SurgVLM-72B	76.7	76.5	82.8	71.0
SurgAtlas-VLM	66.9	66.8	73.5	SurgAtlas-VLM	77.7	77.1	82.9	73.0
SUREON-VLM	71.9	60.8	57.4	SUREON-VLM	78.5	75.6	84.3	75.6
SUREON-VLM (clip)	87.45	82.21	79.78					

a result, SUREON supports the training and evaluation of models across multiple semantic categories spanning perception and reasoning tasks in surgery. Our experiments show two things. First, surgical visual reasoning remains far from solved: no model exceeds 53% accuracy, and multimodal gain does not exceed 29.1%. Second, verification matters beyond data volume: removing text-solvable samples raises multimodal gain from 8.7 to 29.1 while reducing the training set size by 57.0%. A model trained on this data matches larger specialist models on perception benchmarks while also demonstrating strong visual reasoning on SUREON, outperforming existing open-source models. Together, these findings show that progress in surgical vision-language models requires evaluating and training for dependence on visual evidence. SUREON is a step towards more reliable and visually grounded surgical reasoning.

References

- [1] Oluwatosin Alabi et al. Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *Scientific Data*, 12(1):825, 2025. 11, 12, 18
- [2] Max Allan et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019. 11, 12
- [3] Max Allan et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020. 11, 12
- [4] Mohammad Asadi, Jack W O’Sullivan, Fang Cao, Tahoura Nedae, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026. 2, 3, 4, 1
- [5] Nicolás Ayobi et al. Pixel-wise recognition for holistic surgical scene understanding. *MedIA*, page 103726, 2025. 11, 12, 18
- [6] Franco Badaloni, Gino Kuiper, Ronald de Jong, Romy van Jaarsveld, Yiping Li, Mart van der Neut, Alae Asag-Gau, Adam Zeyara, Marcel Breeuwer, Richard van Hillegersberg, et al. Are we developing the right intraoperative ai assistance? surgeons’ perspectives and desired functions. *Surgical Endoscopy*, 40(6):5259–5266, 2026. 1
- [7] Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. *arXiv preprint arXiv:2305.11692*, 2023. 3
- [8] Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 7, 6, 8, 13
- [9] Filippas Bellos, Andre S. Gala-Garza, Miaowei Wang, Alyssa M. Hardin, Ahmad M. Hider, Yayuan Li, Jing Bi, Susan Liang, Chenliang Xu, Donald S. Likosky, and Jason J. Corso. SurgAtlas: A large-scale surgical video–language dataset with 2,391 hours of open and minimally invasive surgery. *arXiv preprint arXiv:2606.25905*, 2026. 3, 11
- [10] James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19552–19564, 2025. 1, 3
- [11] Matthias Carstens et al. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):3, 2023. 11, 12
- [12] Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. Lemon: A large endoscopic monocular dataset and foundation model for perception in surgical settings. *arXiv preprint arXiv:2503.19740*, 2025. 13
- [13] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pages 2318–2335, 2024. 4
- [14] Jun Chen, Dannong Xu, Junjie Fei, Chun-Mei Feng, and Mohamed Elhoseiny. Document haystacks: Vision-language reasoning over piles of 1000+ documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24817–24826, 2025. 3
- [15] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 2, 3, 7
- [16] Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision*, pages 401–417. Springer, 2024. 1
- [17] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9155–9164, 2026. 1
- [18] Tae-Min Choi, Tae Kyeong Jeong, Garam Kim, Jaemin Lee, Yeongyoon Koh, In Cheul Choi, Jae-Ho Chung, Jong Woong Park, and Juyoun Park. Surgmllmbench: A multimodal large language model benchmark dataset for surgical scene understanding. *arXiv preprint arXiv:2511.21339*, 2025. 1, 3
- [19] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinyuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28652–28668, 2026. 7
- [20] Jasper Dekoninck, Nikola Jovanović, Tim Gehringer, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. 2026. 4
- [21] Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*, 2025. 5
- [22] Mauro Orazio Drago, Luca Carlini, Pelinsu Celebi Balyemez, Dennis Pierantozzi, Chiara Lena, Cesare Hassan, Danail Stoyanov, Elena De Momi, Sophia Bano, and Mobarak I Hoque. Surgvivqa: temporally grounded video question answering for surgical scene understanding. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2026. 3
- [23] Google DeepMind. Gemini 3 pro frontier safety framework report. Technical report, Google, 2025. 7
- [24] Google DeepMind. Gemma 4: Open multimodal and reasoning models. Model documentation, 2026. 7
- [25] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5
- [26] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020. 11, 12

- [27] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 5
- [28] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018. 1
- [29] Joël L. Lavanchy et al. Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. *IJCARS*, 2024. 11, 12
- [30] Andreas Leibetseder et al. Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In *Proceedings of the 9th ACM multimedia systems conference*, pages 357–362, 2018. 11, 12
- [31] Jiajie Li, Garrett Skinner, Gene Yang, Brian R Quaranto, Steven D Schwaartzberg, Peter CW Kim, and Jinjun Xiong. Llava-surg: towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981*, 2024. 1, 3
- [32] Yaoqian Li, Xikai Yang, Dunyuan Xu, Yang Yu, Litao Zhao, Xiaowei Hu, Jinpeng Li, and Pheng-Ann Heng. Surgpub-video: A comprehensive surgical video framework for enhanced surgical intelligence in vision-language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6628–6635, 2026. 1, 3
- [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2
- [34] Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *Advances in Neural Information Processing Systems*, 38, 2026. 1, 3
- [35] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 2
- [36] Lena Maier-Hein et al. Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific data*, 8(1):101, 2021. 11, 12
- [37] Albert J Miao et al. Hemoset: The first blood segmentation dataset for automation of hemostasis management. In *2024 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2024. 11, 12
- [38] Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Cotas, Yadong Lu, Wei-ge Chen, Olga Vrousos, Corby Rosset, et al. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*, 2024. 5
- [39] Aditya Murali et al. Latent graph representations for critical view of safety assessment. *IEEE TMI*, pages 1–1, 2023. 8, 11, 12
- [40] Sahar Nasirihaghighi et al. Gynsurg: A comprehensive gynecology laparoscopic surgery dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13141–13147, 2025. 11, 12
- [41] Chinedu Innocent Nwoye et al. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *MedIA*, 78:102433, 2022. 11, 12
- [42] Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. Surglavi: Large-scale hierarchical dataset for surgical vision–language representation learning. *MedIA*, page 103982, 2026. 3, 2, 18
- [43] Dimitrios Psychogyios et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023. 11, 12, 18
- [44] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024. 3
- [45] Tobias Rueckert, David Rauber, Raphaela Maerkl, Leonard Klausmann, Suemeyye R Yildiran, Max Gutbrod, Danilo Weber Nunes, Alvaro Fernandez Moreno, Imanol Luengo, Danail Stoyanov, et al. Comparative validation of surgical phase recognition, instrument key-point estimation, and instrument instance segmentation in endoscopy: Results of the phakir 2024 challenge. *Medical Image Analysis*, page 103945, 2026. 11, 12
- [46] Samuel Schmidgall, Joseph Cho, Cyril Zakka, and William Hiesinger. Gp-vls: A general-purpose vision language model for surgery. *arXiv preprint arXiv:2407.19305*, 2024. 1
- [47] Samuel Schmidgall, Justin D Opfermann, Ji Woong Kim, and Axel Krieger. Will your next surgeon be a robot? autonomy and ai in robotic surgery. *Science robotics*, 10(104): eadt0187, 2025. 1
- [48] Klaus Schoeffmann, Heinrich Husslein, Sabrina Kletz, Stefan Petschnig, Bernd Münzer, and Christian Beecks. Video retrieval in laparoscopic video recordings with dynamic content descriptors. *Multim. Tools Appl.*, 77(13):16813–16832, 2018. 11, 12
- [49] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *MICCAI*, pages 33–43. Springer, 2022. 3
- [50] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. In *International conference on medical image computing and computer-assisted intervention*, pages 281–290. Springer, 2023. 1, 3
- [51] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 7, 5
- [52] Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, Muhammad Jehanzeb Mirza, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky,

- et al. Livexiv-a multi-modal live benchmark based on arxiv papers content. In *International Conference on Learning Representations*, pages 11470–11502, 2025. 3, 6
- [53] Gabriel Szyldo Shein, Ronit Brodie, and Yoav Mintz. Human-machine collaboration in ai-assisted surgery: Balancing autonomy and expertise. *Artificial Intelligence in Medicine and Surgery-An Exploration of Current Trends, Potential Opportunities, and Evolving Threats-Volume 1*, 2023. 1
- [54] Jongmin Shin, Ka Young Kim, Eunki Cho, Seong Tae Kim, and Namkee Oh. Surgcheck: Do vision–language models really look at images in surgical vqa? *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2026. 2, 3
- [55] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 7
- [56] Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, et al. Medgrp: Multi-task reinforcement learning for heterogeneous medical video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2788–2798, 2026. 1, 3, 7
- [57] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. 5, 13
- [58] Andru P Twinanda et al. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE TMI*, 36(1): 86–97, 2016. 8, 11, 12, 18
- [59] Martin Wagner et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *MedIA*, 86:102770, 2023. 8, 11, 12
- [60] Gui Wang, YongSong Zhou, Kaijun Deng, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. Surgcot: Advancing spatiotemporal reasoning in surgical videos through a chain-of-thought benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17011–17021, 2026. 1, 3
- [61] Guankun Wang et al. Copesd: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12636–12643, 2025. 1, 3
- [62] Guankun Wang et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *MedIA*, page 103789, 2025. 3, 6
- [63] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508, 2023. 2
- [64] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer, 2022. 11, 12
- [65] Zhecan Wang, Long Chen, Haoxuan You, Keyang Xu, Yicheng He, Wenhao Li, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang. Dataset bias mitigation in multiple-choice visual question answering and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8598–8617, 2023. 3
- [66] Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. Captionqa: Is your caption as useful as the image itself? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23741–23750, 2026. 3, 6
- [67] Han Yi, Yulu Pan, Feihong He, Xinyu Liu, Benjamin Zhang, Oluwatuminu Oguntola, and Gedas Bertasius. Exact: A video-language benchmark for expert action analysis. *Advances in Neural Information Processing Systems*, 38, 2026. 3
- [68] Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21655–21666, 2025. 1
- [69] Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, et al. Benchmarking visual state tracking in multimodal video understanding. *arXiv preprint arXiv:2606.03920*, 2026. 1
- [70] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 4
- [71] Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *IJCARS*, pages 1–9, 2024. 3
- [72] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37:122952–122983, 2024. 3
- [73] Kun Yuan et al. Learning multi-modal representations by watching hundreds of surgical video lectures. *MedIA*, 105: 103644, 2025. 3
- [74] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, 2025. 3
- [75] Daoguang Zan, Zhirong Huang, Wei Liu, Hanwu Chen, Shulin Xin, Linhao Zhang, Qi Liu, Li Aoyan, Lu Chen, Xiaojian Zhong, et al. Multi-swe-bench: A multilingual bench-

- mark for issue resolving. *Advances in Neural Information Processing Systems*, 38, 2026. 4
- [76] Zhitao Zeng et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*, 2025. 1, 3, 6, 11
- [77] Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29580–29590, 2025. 3
- [78] Yang Zhou, Tan L Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick S Goh. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. *Advances in Neural Information Processing Systems*, 37:6625–6647, 2024. 1

SUREON: A Benchmark and Vision-Language Model for Surgical Reasoning

Supplementary Material

This supplementary material provides additional details, results, and discussion supporting the main paper. Sec. **A** covers limitations and future work; the remaining sections expand on the main paper: Sec. **B** details the four-stage construction pipeline, including the final dataset configuration (main Sec. 3.1); Sec. **C** details the public-data training corpus $\mathcal{D}_{\text{pub}}$ (main Sec. 3.2); Sec. **D** specifies inference and scoring protocols (main Sec. 4); Sec. **E** reports the training data-composition and staging ablations (main Sec. 4); and Sec. **F** discusses ethical considerations.

A Limitations and future work	1
B SUREON Construction Details	2
B.1. Data Source	2
B.2. Stage 1: Raw transcript-grounded QA synthesis	2
B.3. Stage 2: Multi-Model Retrieval-Augmented Distractor Generation	4
B.4. Stage 3: Visual-Dependence Curation Details	7
B.5. Stage 4: Expert Verification Details	9
B.6. Final Dataset Configuration	10
C Training Data Construction from public dataset sources	11
C.1. Stage 1: Common Annotation Schema	11
C.2. Stage 2a: Image-Level QA Template Gener- ation	12
C.3. Stage 2b: Video Clip QA Template Generation	12
C.4. Stage 3: LLM-Based Linguistic Augmentation	13
D Evaluation	13
D.1. SUREON Benchmark Evaluation	13
D.2. Qualitative Error Analysis	15
D.3. Public Perception Benchmark Evaluation . .	16
E SUREON-VLM training details	17
F. Ethics	18

A. Limitations and future work

While SUREON advances visually grounded surgical reasoning evaluation, several limitations and avenues for future work remain:

- Our refinement and blind-test curation (Sec. **B.3–B.4**) reduce text-solvability by operating on the answer options. However, text-solvability is partly a property of how a question is framed: a question that names the operative context or telegraphs the procedural phase can leak its

answer regardless of the distractors. Jointly reframing questions alongside distractors is a natural extension that could reduce the number of removed samples in Stage 3.

- Our guesser ensemble is built from an open-source model (Qwen3-VL-8B), used off-the-shelf and in fine-tuned variants, rather than from commercial models. However, unlike previous work that only use models off-the-shelf, the fine-tuned guessers reveal language shortcuts that only become learnable once a model is exposed to the answer distribution – leakage that zero-shot off-the-shelf testers, including commercial ones, systematically underestimate [4]. Additionally, the precise set of samples removed during curation may vary under a different guesser family or as foundation models become more capable. We therefore complement overall benchmark accuracy with evaluation-time multimodal gain, $MG = S_v - S_{wv}$, which evaluates each model both with and without visual input and measures how much of its performance is attributable to the surgical video rather than language priors or shortcuts. This provides a model-specific check for visual dependence beyond the fixed curation ensemble and allows the benchmark to remain auditable as model capabilities improve.
- Every QA pair is anchored to a Semantic Grounding Moment, excluding visually salient content that lecturers do not verbalize. Thus, our benchmark under-samples reasoning that is observable but un-narrated, and its coverage is shaped by what surgeons choose to teach rather than by everything a model could be asked to see.
- Because SUREON is sourced from educational lecture videos, the case mix likely over-represents textbook presentations, well-rehearsed techniques, and controlled operative conditions relative to the full diversity of real-world practice. Atypical anatomies, intraoperative complications, and less common surgical approaches may therefore be underrepresented, and strong benchmark performance should not be assumed to transfer to the long tail of clinical variability.
- SUREON is built from edited, narrated, minimally-invasive lecture recordings. Intraoperative video differs (uncut, without narration, with degraded visibility from smoke, blood, and motion) so strong SUREON performance does not by itself establish readiness for live surgical settings. SUREON-DB is additionally pseudo-labeled and not expert-verified; both resources are intended for research and evaluation, and not for clinical use.

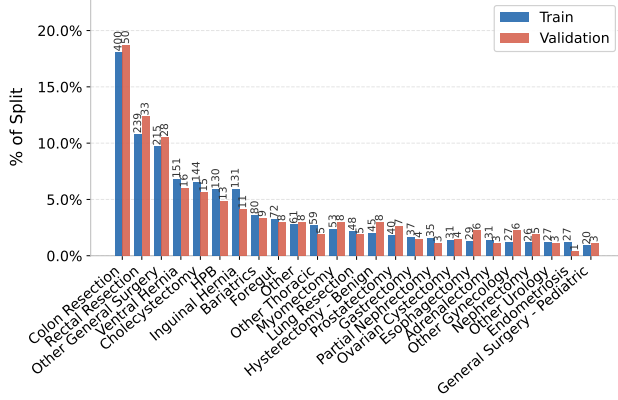


Figure 5. Video subject distribution across training and validation splits (Included subjects with at least 1% of representation).

B. SUREON Construction Details

B.1. Data Source

SUREON is built on the SurgLaVi- β dataset [42], which provides timestamped transcriptions of surgical academic lectures for 3,148 YouTube videos, along with metadata on the surgical specialty, subject, and procedure type of each video.

Dataset splits. SurgLaVi- β was originally released as a pre-training corpus and therefore does not include predefined train/validation/test splits. We performed stratified video-level splitting based on procedure subject, ensuring that the training and validation distributions are balanced across subjects. Approximately 10% of videos were held out for validation, yielding 2,809 videos for training and 339 for validation. The resulting subject distributions are shown in Figure 5.

B.2. Stage 1: Raw transcript-grounded QA synthesis

B.2.1. Notation

Each video v_i is paired with an aligned transcript $T_i = \{(s_{ij}, [t_{ij}^{\text{start}}, t_{ij}^{\text{end}}])\}_{j=1}^{m_i}$ of m_i timestamped narrated sentences, where s_{ij} is the j -th narrated sentence and $[t_{ij}^{\text{start}}, t_{ij}^{\text{end}}]$ its spoken interval. From this transcript, a Semantic Grounding Moment is the tuple $z_{ij}^k = (v_i[t_s, t_e], s_{ij:j+r}, c_k)$, pairing a video clip $v_i[t_s, t_e]$ with its supporting transcript context $s_{ij:j+r}$ and a target semantic category c_k . Each SGM is converted into one or more QA samples $x_{ij}^k = (v_i[t_s, t_e], Q_{ij}^k, A_{ij}^k, \mathcal{O}_{ij}^k)$, where Q_{ij}^k is the question, A_{ij}^k the ground-truth answer, and $\mathcal{O}_{ij}^k$ the multiple-choice option set when applicable. This notation is used throughout the remaining construction stages.

B.2.2. Semantic Grounding Moments

Surgical narrations are sparse: even when instruments, anatomical structures, or actions are visible in the operative field, surgeons do not continuously annotate them in speech. If raw narration segments are used directly as scene descriptions, the resulting labels are incomplete and potentially misleading— for instance, a narrator may mention a vessel sealer while a bipolar forceps is simultaneously visible but left unaddressed. Treating such a segment as a full scene description would introduce label noise rather than grounded supervision. To address this, we identify *Semantic Grounding Moments* (SGMs): focused intervals in the transcript where the narrator explicitly anchors a semantic cue to the current surgical scene. SGMs serve as reliable anchors for structured QA generation, since they represent moments of intentional visual-linguistic correspondence. Figure 6 illustrates representative SGMs across multiple semantic categories in the transcript of a surgical video.

B.2.3. Multi-Agent Data synthesis details

Model and inference configuration. All generator and validator agents are built on gpt-5-2025-08-28 (OpenAI), with the `reasoning` parameter set to `high` to maximize generation quality. Generator agents are queried at temperature $T=0.1$ to allow mild variation across repeated runs while preserving factual consistency. Validator agents are queried at temperature $T=0.0$ to enforce deterministic, rule-based judgments. Because GPT-5 is a proprietary API model, exact regeneration remains dependent on continued access to this model version and may be affected by provider-side changes.

Specialized agent pairs. We instantiate $K=10$ generator-validator agent pairs (20 agents in total), one per semantic category, each equipped with detailed system prompts specifying its role, generation instructions, and validation constraints. Table 7 summarizes the key instructions and constraints for each agent pair (Full prompts available in the [anonymized repository](#)).

Multi-call generation. Each generator agent is invoked $R=3$ times per transcript. Given that a single call can miss opportunities for QA generation within a given transcript, we run three independent calls to increase overall recall on finding SGMs and generating QAs from them. The combined outputs of all R calls are then passed jointly to the paired validator, which validates each candidate sample and removes semantic near-duplicates.

Validation criteria. Additional to applying the category-specific constraints listed in Table 7, the validator enforces three universal filtering rules on every generated sample:

- *Hallucination filtering.* A sample is discarded if the question or correct answer introduces an entity, action, or clinical claim absent from the paired transcript segment.

[...]
00:00:17,444 --> 00:00:22,225
Preoperative imaging did not demonstrate evidence of nodal involvement or metastatic disease.
00:00:22,785 --> 00:00:31,127
After pneumoperitoneum is established with an open technique, two further five-millimeter ports are inserted under vision suprapubically and in the epigastric region.
00:00:31,687 --> 00:00:35,688
A fourth twelve-millimeter operating trocar is inserted in the left iliac fossa.
00:00:36,388 --> 00:00:39,989
A medial to lateral no-touch technique is our standard approach.
00:00:40,330 --> 00:00:44,611
The ileocecal junction is retracted laterally exposing the ileocolic vessels.
00:00:45,452 --> 00:00:58,357
The mesenteric window is opened immediately below the ileocolic vessels and after identification of the duodenum the right retrocolic space is developed cranially and laterally along the avascular plane of Toldt.
00:00:59,357 --> 00:01:02,038
A retrocecal appendix is here demonstrated.
00:01:03,510 --> 00:01:05,412
The ileocolic pedicle is then approached.
00:01:05,992 --> 00:01:10,395
The dissection is here performed with the Thunderbeat vessel sealing device.
00:01:11,016 --> 00:01:18,522
The ileocolic vein and artery are dissected separately and ligated at the origin of the fully exposed superior mesenteric vessels.
00:02:39,205 --> 00:02:43,068
The duodenum is progressively dissected off the transverse colon.
00:02:43,628 --> 00:02:52,815
A combination of blunt and sharp dissection is used, constantly keeping the duodenum in the operative field in order to prevent thermal injury.
00:03:07,150 --> 00:03:10,640
A right colic artery was not identified in this patient.
00:03:52,021 --> 00:03:58,532
The right branches of the middle colic vessels are dissected and separately ligated at their origin on the top of the pancreas.
00:05:15,046 --> 00:05:22,474
The procedure is then completed in a standardized fashion, with full mobilization of the hepatic flexure performed from below.
00:05:22,494 --> 00:05:31,384
In agreement with the principles of the no-touch technique, only after the vascular supply has been interrupted, the right colon is finally mobilized laterally.
00:05:43,028 --> 00:05:54,209
The specimen is then delivered through a five-centimeter right-sided transverse incision after wound protection and an unsewn end-to-side ileotransverse anastomosis is fashioned.
00:06:04,915 --> 00:06:11,281
The post-operative recovery was uneventful, and the patient was discharged home on post-operative day four.
00:06:11,882 --> 00:06:17,567
Histopathology demonstrated a T-free cancer with none out of twenty-two lymph nodes involved.

Figure 6. Semantic Grounding Moments (SGMs) identified in a transcript from a laparoscopic colon resection procedure, color-coded by semantic category: Entity Existence, Entity Attribute, Action Description, and Safety Practice Identification. Four categories are shown for visual clarity, as displaying all categories simultaneously would result in overlapping highlights.

- *Visual grounding.* A sample is discarded if the question and its ground-truth answer correspond to a narration comment that is not referring to the present visual scene (e.g. comments referring to events that took place in the past, generic procedural information, etc).
- *Temporal window correctness.* A sample is discarded if its assigned temporal window does not cover the transcript segment from which the correct answer is derived.
- *Semantic deduplication.* Near-duplicate samples arising from multiple generator calls are identified and collapsed. Only the most informative candidate is retained.

Structured output schemas. To enable downstream parsing, both agents are prompted to return structured JSON with reasoning traces justifying its decisions. Listing 1 and Listing 2 show the generator and validator output schemas, respectively.

Listing 1. Generator agent output schema. Each entry encodes the question, options, answer, temporal window, video scope, and a brief reasoning trace.

```
{
  "question": "...",
  "options": ["...", "...", "..."],
```

```
"answer": "...",
"time_window": {
  "start": "HH:MM:SS.xxx",
  "end": "HH:MM:SS.xxx"
},
"scope": "clip" | "full_video",
"reasoning": "1-3 sentences explaining
grounding, uniqueness, and category
fit."
}
```

Listing 2. Validator agent output schema. The validator returns a summary of kept and removed samples, with per-removed-sample issue annotations.

```
{
  "summary": {
    "total_candidates": N,
    "kept": K,
    "removed": R
  },
  "kept": [
    {
      "question": "...",
      "options": ["...", "...", "..."],
```

Table 6. Per-category sample counts after Stage 1 validation ($\mathcal{D}_{SGM}$), reported for the training split, validation split, and full corpus.

Category	Train		Validation		Total	
	Count	%	Count	%	Count	%
AD	27,776	15.7	3,430	15.8	31,206	15.7
EA	26,685	15.1	3,216	14.9	29,901	15.1
EE	23,131	13.1	2,919	13.5	26,050	13.1
LAR	18,951	10.7	2,238	10.3	21,189	10.7
TO	17,600	10.0	2,055	9.5	19,655	9.9
PAD	16,191	9.2	2,029	9.4	18,220	9.2
SM	14,926	8.5	1,784	8.4	16,735	8.4
IAI	13,133	7.4	1,675	7.7	14,808	7.5
SPI	9,609	5.4	1,192	5.5	10,801	5.4
F	8,591	4.9	1,092	5.0	9,683	4.9
Total	176,593	100.0	21,655	100.0	198,248	100.0

```

    "answer": "...",
    "time_window": {"start": "...", "end": "..."},
    "scope": "clip",
    "reasoning": "...",
  },
  ],
  "removed": [
    {
      "index": 0,
      "issues": ["...", "..."]
    }
  ]
}

```

Quantitative summary. Stage 1 processes the transcripts of all 3,148 videos in the corpus, of which 2,478 (78.7%) produced at least one validated QA sample (train: 2,211 | val: 267), yielding 198,248 agent-validated samples in total – 176,593 in the training split and 21,655 in the validation split, leading to a total cost of approximately 20,000 USD. Table 6 reports the sample count and relative frequency per semantic category for each split and the full corpus. The per-category proportions are broadly consistent across splits.

B.3. Stage 2: Multi-Model Retrieval-Augmented Distractor Generation

Stage 1 produces transcript-grounded QA pairs with naively generated multiple-choice distractors. Inspection of these distractors reveals two compounding failure modes. First, correct answers and distractors differ in word length, vocabulary, and level of clinical specificity because distractors are generated from the LLM’s base knowledge distribution while the correct answers are derived from the transcript. Overall, this creates surface shortcuts that allow LLMs to identify the correct answer without any visual grounding

Q: What is the surgeon aiming to prevent by proceeding slowly and hemostatically through the dense scar? (LAR)

Stage 1

- Accidental injury to a vas deferens that may be hidden in the scar ✓ (correct answer)
- Accidental entry into the hernia sac
- Thermal injury to the peritoneum
- Bleeding from the inferior epigastric vessels

Stage 2

- To maintain visualization by preventing blood obscuration
- To avoid injuring the spermatic cord structures
- To achieve hemostasis of the soft tissue surrounding the hernia sac
- To prevent excessive dissection that would compromise the hernia repair
- To minimize lymphatic disruption and bleeding during the tissue separation
- To avoid obscuring the hernia sac with excessive bleeding
- Accidental injury to a vas deferens that may be hidden in the scar ✓ (correct answer)

Q: Given the intraoperative context shown, what would be the expected next procedural action? (F)

Stage 1

- Accidental injury to a vas deferens that may be hidden in the scar ✓ (correct answer)
- Accidental entry into the hernia sac
- Thermal injury to the peritoneum
- Bleeding from the inferior epigastric vessels

Stage 2

- Advance the dissection along the lateral pelvic fascia
- Perform intraoperative endoscopic evaluation to confirm intraoperative anatomy
- Begin the first layer of posterior reconstruction by suturing the posterior fascia to the sphincter
- Release the ProGrasp providing cephalad retraction on the bladder
- Place a suture to control the dorsal venous complex ✓ (correct answer)
- Begin the vesicourethral anastomosis by starting the posterior wall with a double-arm running suture
- Conduct an intraoperative assessment to confirm the integrity of the repair

Q: What action is performed by the left-hand graspers in this clip? (IAI)

Stage 1

- Lift the uterus
- Opening the anterior leaf of the broad ligament ✓ (correct answer)
- Grasp the round ligament
- Hold the ovarian ligament

Stage 2

- Opening the anterior leaf of the broad ligament ✓ (correct answer)
- Coagulating the left ovarian ligament
- Desiccating and cutting the left uterine vessels
- Opening the posterior peritoneum of the pouch of Douglas
- Dividing the ovarian ligament
- Inspecting the pelvic sidewall for ureteral injury
- Mobilizing the sigmoid colon laterally

Figure 7. **Qualitative comparison of automatically generated distractors at Stage 1 and Stage 2 on representative examples.** Baseline (Stage 1) distractors are **vaguer** and shorter than the correct answers, exposing a length and specificity shortcut. In contrast, distractors produced by Retrieval-Augmented Distractor Generation (RADG) and multi-model mixing (Stage 2) are harder, **matching** the correct answer in length, phrasing, and clinical specificity to reduce blind guessing.

(see Stage 1 distractors in Figure 7 for representative examples). Second, distractors generated by a LLMs share model-specific syntactic patterns and vocabulary, introducing an additional model-artifact shortcut that is even more harmful during fine-tuning, when the model can learn from these patterns.

Stage 2 mitigates both failure modes through a two-component strategy: Retrieval-Augmented Distractor Generation (RADG) to align the distractor distribution and style with that of correct answers, and *multi-model mixing* to suppress model-specific artifacts. The following sections detail the implementation of each component.

B.3.1. Retrieval-Augmented Distractor Generation (RADG)

Embedding model and pool construction. All correct answers in the training set are encoded with BAAI/bge-m3 [13], a dense text embedding model

Table 7. Per-category specialized agent properties. For each of the 10 semantic categories, we report the answer format and the primary instruction and constraints enforced at generation and validation.

Category	Format	Generator — Key Instructions	Validator — Key Rules
<i>Short-Term</i>			
Entity Existence (EE)	MCQ	Anchored existence-only questions; answer entity must not appear in question phrasing	Reject if action, location, or attribute implied; unique answer enforced per window
Entity Attribute (EA)	MCQ	Observable qualitative states only; numeric or measured attributes prohibited	Reject if action, causality, or non-visual condition; no quantitative measures
Instrument–Action Interaction (IAI)	MCQ	Instrument must be explicitly named in transcript; bidirectional query formulation	Bidirectional consistency enforced; exactly one instrument–action pair allowed
Action Description (AD)	MCQ	Standard procedural verbs only; no intent, outcome, or instrument information	Action-only constraint; answer must be visually observable, not inferred from narration
Procedure-agnostic Action Description (PAD)	MCQ	Labels ≤ 3 words; anatomy-, instrument-, and procedure-free phrasing required	Reject if label > 3 words or references anatomy or instruments; one action per window
Local Action Reasoning (LAR)	MCQ	Reason must be explicitly narrated; immediate and local purpose only	Narration-backed correct answer required; global strategy reasoning rejected
Safety Practice Identification (SPI)	MCQ	Explicit safety cue required in transcript; behavior must be directly video-observable	Transcript justifies window selection only; answer identifiable from video alone
<i>Mid-, Long-Term</i>			
Forecasting (F)	MCQ	Mid-procedure context only; multiple completed steps required; generic question phrasing	Multiple completed steps must support the predicted action; window ends before predicted action
Sequence Summarization (SM)	Open-ended	Goal-directed multi-step workflow; ordered procedural steps across clips	Temporal sequence of steps required; single-action or unordered answers rejected
Temporal Ordering (TO)	Open-ended	Three distinct events; randomized label presentation	Order-neutral question phrasing; unambiguous temporal separability required

optimized for retrieval-augmented generation and semantic similarity search. Each embedding is stored in a vector database alongside four metadata fields: procedure type, specialty, subject, and question type.

Per-question retrieval. For each question Q_{ij}^k we query four metadata-filtered pools, retrieving the top 100 candidates per pool ranked by cosine similarity to the query embedding: answers matching the same procedure type, specialty, and subject as the source video, plus a broad pool spanning all answers of the same question type with no further metadata restriction. This yields a total of 400 exemplar candidates of correct answers in the dataset across pools.

Merging and deduplication. Candidates from the four pools are merged into the exemplar set E_{ij}^k by iterating through them in priority order (procedure $\rightarrow$ specialty $\rightarrow$ subject $\rightarrow$ broad). Candidates already present in the merged list are discarded, yielding a deduplicated exemplar set of at most $K=100$ answers that is then passed to distractor generators as answer-style references for the present question.

Distractor Generators. We instantiate three distractor-generator LLMs available on Hugging Face under the following identifiers:

- M_1 : google/medgemma-27b-text-it [51]
- M_2 : meta-llama/Llama-3.1-70B [25]

- M_3 : Qwen/Qwen3.5-35B-A3B [57]

Each generator produces $N=6$ distractors per question, yielding a combined pool of $3N=18$ candidates.

Generation conditioning. Each generator M_ℓ is conditioned on the question Q_{ij}^k , its correct answer A_{ij}^k , the original option set $\mathcal{O}_{ij}^k$, the local transcript context window H_{ij}^k , the retrieved exemplar set E_{ij}^k , and video-level metadata m_i :

$$\mathcal{O}_{ij}^{k,\ell} = M_\ell(Q_{ij}^k, A_{ij}^k, \mathcal{O}_{ij}^k, H_{ij}^k, E_{ij}^k, m_i). \quad (4)$$

Implementation details. All distractor generators are served via vLLM [27] on $4 \times$ NVIDIA H200 GPUs with $T=0.1$. Requests are dispatched asynchronously with a concurrency limit of 128 (35B and 27B models) or 64 (122B models) in-flight calls. End-to-end distractor generation for the entire dataset takes approximately 6.4 days: Qwen3.5-35B (4.5 d), LLaMA-70B (1.2 d), and MedGemma-27B (0.7 d).

The same system prompt is applied identically to all three generators and is shown in Figure 8. We instruct generators to produce distractors whose length, vocabulary, and specificity are indistinguishable from the correct-answer style of the provided exemplar, while still being grounded on the transcript context.

Multi-model mixing. Each of the three generators contributes $N=6$ distractors to the combined pool $\mathcal{D}_{ij}^{k,\text{pool}} =$

```

## Role
You are a **surgical VLM dataset quality agent** specialized in rewriting
**distractor options** for multiple-choice questions so that
distractors are **distributionally indistinguishable** from real correct answers
while remaining factually wrong for the specific question.
...

## Why This Task Exists
This dataset is used to train vision-language models (VLMs). A text-only model
should NOT be able to identify the correct answer from linguistic style alone.
If distractors look stylistically different from correct answers, the model learns
a shortcut. Your job is to eliminate that shortcut.
...

## Input
You will receive:
- **video_title**: title of the surgical video
- **specialty**: surgical specialty (e.g., General Surgery, Urology)
- **subject**: subject/anatomy focus (e.g., Colorectal, Hepatobiliary)
- **procedure_type**: specific procedure (e.g., Laparoscopic Cholecystectomy)
- **question**: the multiple-choice question
- **answer**: the correct answer (DO NOT change this)
- **existing_options**: the current distractors (for reference)
- **transcript_segment**: the relevant portion of the surgical transcript for this question's
time window
- **answer_examples**: a curated list of real correct answers from OTHER videos for this same
question type – these represent the target distribution your distractors must match
...

## Task
Generate **6 new medium-level distractor options** (options_unbiased) that:
1. **Match the style and distribution of the provided answer_examples** – same
vocabulary register, specificity level, phrasing patterns, and length
2. Are **clearly incorrect** for this specific question (cannot be correct given
the transcript segment)
3. Are **clinically plausible** – a surgical professional would recognize them as
real entities/actions/concepts in the appropriate domain
4. Are **not trivially different** from the correct answer (no minor synonyms)
5. Do **not** include the correct answer or any close synonym
6. Are **distinct from each other**
...

## Critical Constraint: Distribution Matching
Study the **answer_examples** carefully. Note:
- The level of specificity (generic vs. precise anatomical/instrument names)
- The typical length (short phrases vs. full sentences)
- The vocabulary register (lay vs. clinical)
- The entity types (anatomy, instruments, actions, procedures, etc.)
Your 6 new distractors MUST feel like they could have been a correct answer to a
DIFFERENT question of the same type, drawn from the same dataset.
...

## What NOT to Do
- Do NOT generate generic filler options like "none of the above" or "all of the above"
- Do NOT reuse the correct answer or paraphrase it
- Do NOT generate options that are obviously incorrect to any clinician
- Do NOT generate options that are too easy to eliminate from text alone
- Do NOT change the question or correct answer
...

## Output Format
Return a **single JSON object** only.
{
  json
  {
    "options_unbiased": ["...", "...", "...", "...", "...", "..."]
  }
}

```

Figure 8. Distractor generation prompt to create distractors that are closer in style to correct answers.

$\mathcal{O}_{ij}^{k,M_1} \cup \mathcal{O}_{ij}^{k,M_2} \cup \mathcal{O}_{ij}^{k,M_3}$. The final distractor set is drawn by random sampling from this pool subject to a cross-model diversity constraint, ensuring that the selected distractors are not all sourced from a single generator.

Blind validation. We validate each distractor setup by measuring whether models can exploit linguistic artifacts under *blind testing*: the model receives only the question and answer options, with no video input, and must select the correct answer. Effective distractor refinement should drive blind-tested accuracy toward the random-choice baseline ($p_{\text{rand}} = 0.250$ for 3 distractors and 0.143 for 6 distractors).

We evaluate two conditions. The *off-the-shelf* (OS) condition applies the model directly under blind testing without any task-specific training, measuring whether the distractor configuration is exploitable at inference time. To assess for shortcut patterns that a zero-shot model cannot exploit but may become learnable once the model is exposed to the distractor distribution during training, the *blind fine-tuning* (FT) condition trains the same model exclusively on (question, options) pairs – no video – for the distractor variant being evaluated, then evaluates it under blind testing.

Model and training configuration. Both OS and FT

Table 8. Blind-tested accuracy under different distractor refinement strategies ($\downarrow$ is better). *Baseline*: original, unrefined distractors. *Resampled*: correct answers from other questions of the same type used as distractors. M_1 : MedGemma, M_2 : Llama, M_3 : Qwen. OS: Qwen3-VL-8B-Instruct under blind testing. FT: same model after blind fine-tuning. Avg: mean of OS and FT.

Method	RADG	Mix	3 distractors			6 distractors		
			OS ($\downarrow$)	FT ($\downarrow$)	Avg ($\downarrow$)	OS ($\downarrow$)	FT ($\downarrow$)	Avg ($\downarrow$)
Random	–	–	0.250	0.250	0.250	0.143	0.143	0.143
Baseline	–	–	0.577	0.818	0.698	–	–	–
Resampled	–	–	0.556	0.699	0.628	0.463	0.623	0.543
M_1	–	–	0.558	0.876	0.717	0.452	0.843	0.648
M_2	–	–	0.481	0.882	0.682	0.389	0.867	0.628
M_3	–	–	0.449	0.812	0.631	0.339	0.764	0.552
M_1	✓	–	0.446	0.836	0.641	0.326	0.795	0.561
M_2	✓	–	0.408	0.904	0.656	0.295	0.888	0.592
M_3	✓	–	0.415	0.783	0.599	0.302	0.732	0.517
LLM-Mix	✓	✓	0.401	0.789	0.595	0.277	0.725	0.501

conditions use the Qwen/Qwen3-VL-8B-Instruct [8] multimodal model. Blind fine-tuning is performed for one epoch with full LLM fine-tuning (no layers frozen), using a learning rate of 1×10^{-5} , a per-device batch size of 12, and 4 NVIDIA H200 GPUs.

Quantitative Results. Table 8 reports blind-tested accuracy for all configurations under both conditions and both distractor counts (3 and 6). RADG alone reduces OS blind accuracy across all three generators relative to their unaugmented counterparts, confirming that exemplar-conditioned generation narrows the distributional gap. Multi-model mixing further reduces OS accuracy in both settings, with LLM-Mix achieving the lowest blind accuracy overall (0.401 for 3 distractors; 0.277 for 6 distractors). The FT results reveal that the resampled baseline achieves competitive FT scores despite higher OS scores: resampled distractors are harder for fine-tuned models to exploit but remain vulnerable at zero-shot inference time. LLM-Mix reduces OS blind accuracy most consistently across both distractor counts and both evaluation conditions, making it the preferred configuration for the final benchmark. Importantly, RADG reduces blind accuracy across all three model families, while their mixture performs best overall, suggesting that the refinement strategy is not dependent on a single distractor generator.

Qualitative results. While semantic differences between naive distractors and correct answers—such as clinical specificity or depth of medical knowledge—are difficult to quantify directly, one surface-level characteristic stands out clearly: response length. Correct answers tend to be more explicit and precise, making them longer than naive distractors. This length asymmetry is a spurious cue that off-the-shelf models can exploit to identify the correct answer without any task-specific fine-tuning. Figure 9 quantifies this gap by plotting the word-count distributions of correct an-

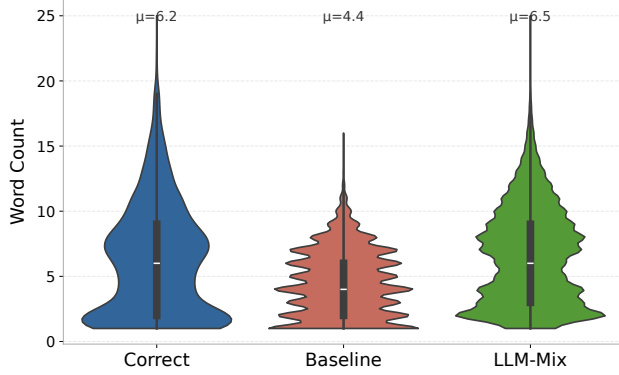


Figure 9. Word-length distributions of correct answers, baseline distractors, and LLM-Mix distractors across multiple-choice training samples.

swers, baseline distractors, and LLM-Mix distractors across all multiple-choice samples in the training set. Baseline distractors exhibit shorter sentences compared to correct answers. In contrast, LLM-Mix distractors—conditioned on retrieved exemplars via RADG—closely match the correct-answer distribution in terms of length.

B.4. Stage 3: Visual-Dependence Curation Details

As reported in Table 8, blind fine-tuning on the refined distractor set drives accuracy above 70% without any video input, confirming that residual language shortcuts persist even after Stage 2 refinement. Stage 3 removes samples solvable without visual evidence by applying a blind-testing consistency filter to both the training set (via cross-validation) and the validation set.

B.4.1. Training Set Cross-Validation Setup

The 2,211 training videos that produced at least one validated QA sample are partitioned into four folds via stratified splitting on procedure subject, ensuring that the subject distribution is balanced across folds. Figure 10 shows the resulting subject distribution per fold, Table 9 summarizes the video and QA counts per fold, and Table 10 reports the per-category QA breakdown.

Table 9. Video and QA counts per cross-validation fold (training set).

Fold	Videos	Total QAs
Fold 0	587	47,153
Fold 1	569	45,496
Fold 2	535	43,317
Fold 3	520	40,627
Total	2,211	176,593

Table 10. Per-category QA counts across the four cross-validation folds. Category proportions are consistent across folds, confirming that stratification did not introduce category-level imbalance.

Category	Fold 0		Fold 1		Fold 2		Fold 3	
	Count	%	Count	%	Count	%	Count	%
AD	7,415	15.7	7,131	15.7	6,911	16.0	6,319	15.6
EA	7,191	15.3	6,862	15.1	6,408	14.8	6,224	15.3
EE	6,222	13.2	5,884	12.9	5,716	13.2	5,309	13.1
LAR	5,111	10.8	4,867	10.7	4,685	10.8	4,288	10.6
TO	4,514	9.6	4,488	9.9	4,468	10.3	4,130	10.2
PAD	4,330	9.2	4,289	9.4	3,919	9.0	3,653	9.0
SM	3,979	8.4	3,836	8.4	3,632	8.4	3,479	8.6
IAI	3,543	7.5	3,478	7.6	3,147	7.3	2,965	7.3
SPI	2,612	5.5	2,457	5.4	2,367	5.5	2,173	5.3
F	2,236	4.7	2,204	4.8	2,064	4.8	2,087	5.1
Total	47,153	100.0	45,496	100.0	43,317	100.0	40,627	100.0

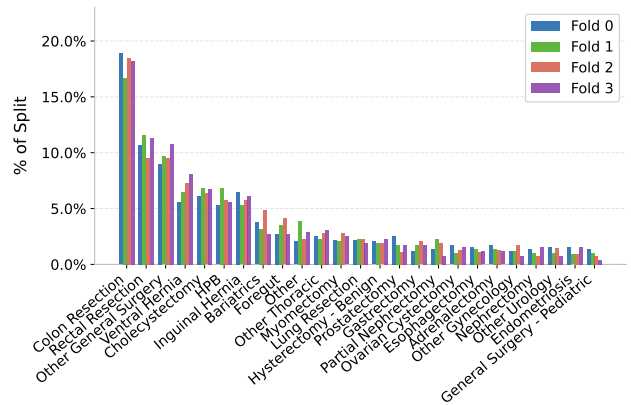


Figure 10. Video subject distribution across folds within the training set (*Included subjects with at least 1% of representation*).

B.4.2. Guesser Ensemble and Curation Protocol

Guesser composition. The ensemble $\mathcal{G}$ comprises four guesser types, all evaluated under blind testing: (i) models blind-fine-tuned on the refined distractors $\mathcal{D}_{\text{ref}}$; (ii) models blind-fine-tuned on the baseline (Stage 1) distractors; (iii) an off-the-shelf base model with no fine-tuning; and (iv) models fine-tuned with full video access but evaluated under blind testing, which expose shortcuts that video-trained models may absorb from language co-occurrences during training. Types (i)–(iii) are used for training-set curation; the full ensemble (i)–(iv) is used for validation-set curation.

Model and inference configuration. All guessers instantiate Qwen/Qwen3-VL-8B-Instruct [8], the same architecture used in Stage 2 (Sec. B.3). Ensemble diversity therefore arises from the training setup: the same model is used off-the-shelf and in fine-tuned variants to find language shortcuts (on refined distractors, on baseline distractors, and with full video access), and all guessers are queried in text-only mode.

Blind fine-tuned guessers (types (i)–(ii), and (iv) for

validation-set curation) are trained for one epoch with a learning rate of 1×10^{-5} on 4 NVIDIA H200 GPUs. Temporal Ordering; Sequence Summarization is excluded, as summarizing would require the video and no distractors are used.

Consistency score and thresholds. Each guesser $g \in \mathcal{G}$ is queried $R=5$ times at temperature $T=0.7$ with no video input, receiving only the question Q_{ij}^k and the refined options $\mathcal{O}_{ij}^{k,\text{ref}}$. Repeated queries are necessary because a single blind query carries a non-negligible probability of a correct random guess, which would cause spurious removal of visually grounded samples. The consistency score is the fraction of runs in which g recovers the correct answer:

$$S(g, x_{ij}^k) = \frac{1}{R} \sum_{r=1}^R \mathbf{1}[g_r(Q_{ij}^k, \mathcal{O}_{ij}^{k,\text{ref}}) = A_{ij}^k], \quad (5)$$

so that filtering targets only samples reliably solvable from text alone rather than those guessed correctly by chance. A sample is retained only if no guesser reaches the consistency threshold $\tau=3/5$:

$$x_{ij}^k \text{ retained} \iff \max_{g \in \mathcal{G}} S(g, x_{ij}^k) < \tau, \quad (6)$$

i.e. a sample is discarded whenever any guesser achieves $S \geq 3/5$.

Training set curation. For each fold f , guessers of types (i)–(iii) are trained on the three complementary folds and scored on the held-out fold.

Validation set curation. All guesser types (i)–(iv) are trained on the full SUREON training set and evaluated on the validation set.

SUREON-DB quality proxy. We use the out-of-fold blind-test pass rate as a proxy for text-answerable contamination in the automatically generated training data. Of the 161,667 training samples eligible for blind curation (all types except Sequence Summarization), 61,135 (37.8%) pass the criterion above, while 100,532 (62.2%) are identified as consistently answerable from text alone and removed. By construction, every sample retained in SUREON-DB passes the held-out blind-guesser criterion, providing a sample-level quality control against language shortcuts without requiring expert review of the full training corpus.

B.4.3. Distractor Hardening via Iterative Blind Testing

As a final curation step applied to the validation set, we fine-tune Qwen/Qwen3-VL-8B-Instruct [8] on the full curated training set SUREON-DB using text only (question and answer options, no video) for 10 epochs with a learning rate of 1×10^{-5} on 4 NVIDIA H200 GPUs. Training for longer than in Sec. B.4.2 produces a stronger adversary that has more thoroughly absorbed any residual shortcut patterns in the training distribution.

Algorithm 1 Distractor Hardening via Iterative Blind Testing

Require: Curated validation sample x_{ij}^k , distractor pool $\mathcal{D}_{ij}^{k,\text{pool}}$, hardened guesser $\mathcal{G}$, threshold $\tau=3/5$, max attempts $M=32$

Ensure: Hardened sample with final distractors, or $\emptyset$

```

1: for  $m \leftarrow 1$  to  $M$  do
2:   Sample  $\mathcal{O}_{\text{cand}}$  from  $\mathcal{D}_{ij}^{k,\text{pool}}$ 
3:   Compute  $S(\mathcal{G}, x_{ij}^k)$  using  $\mathcal{O}_{\text{cand}}$   $\triangleright R=5$  blind queries at  $T=0.7$ 
4:   if  $S(\mathcal{G}, x_{ij}^k) < \tau$  then
5:     return  $(x_{ij}^k, \mathcal{O}_{\text{cand}})$   $\triangleright$  guesser fails: configuration hardened
6:   end if
7: end for
8: return  $\emptyset$   $\triangleright$  no valid configuration found: discard

```

For each retained validation sample, we search for the distractor configuration from the pool $\mathcal{D}_{ij}^{k,\text{pool}}$ generated in Stage 2 that the hardened guesser fails to answer correctly under blind testing. Up to $M=32$ candidate distractor configurations are drawn at random (seed=42) from $\mathcal{D}_{ij}^{k,\text{pool}}$; the first configuration for which the guesser fails ($S < \tau=3/5$) is retained as the final distractor set. If no such configuration is found within the M attempts, the sample is discarded. Algorithm 1 formalizes this procedure.

Held-out guesser generalization. To test whether hardening overfits to the model family used during curation, we additionally evaluate Gemma-4-31B-IT, which is not used at any stage of blind filtering or distractor hardening, under the same text-only condition. For the Qwen3 model trained on SUREON, blind accuracy decreases from 36.5% on the pre-hardening validation set to 21.4% after hardening. Gemma-4-31B-IT remains comparably low across the same sets (23.2% to 22.9%). Thus, while hardening primarily removes shortcuts learned by the adversary used during curation, the resulting samples remain difficult for a held-out model family, providing evidence that the procedure does not simply specialize the benchmark against Qwen-specific failure modes.

Table 11 reports per-category sample counts at each stage of curation: training and validation after blind-testing filtering (Sec. B.4.2), and the validation set after distractor hardening. Sequence Summarization is excluded from curation and hardening; all SM samples are carried through unchanged from Stage 1.

Training set curation ablation. To validate that curating the training set encourages models to ground predictions in visual evidence, we compare two training configurations: a model trained on the full uncured set $\mathcal{D}_{\text{ref}}$ and a model trained on the curated set SUREON-DB, both evaluated on

Table 11. Per-category sample counts after Stage 3 blind-testing curation and validation-set distractor hardening. Survival rates are computed relative to the Stage 1 output. † SM is excluded from curation and hardening; all samples are carried through unchanged.

Category	Train (after curation)		Val (after curation)		Val (after hardening)	
	Count	Rate	Count	Rate	Count	Rate
AD	14,184	51.1%	1,534	44.7%	1,102	32.1%
EA	5,978	22.4%	582	18.1%	439	13.7%
EE	7,816	33.8%	772	26.4%	607	20.8%
LAR	5,859	30.9%	570	25.5%	377	16.8%
TO	7,889	44.8%	891	43.4%	641	31.2%
PAD	6,905	42.6%	747	36.8%	653	32.2%
SM†	14,926	100.0%	1,784	100.0%	1,784	100.0%
IAI	5,351	40.7%	580	34.6%	536	32.0%
SPI	3,870	40.3%	408	34.2%	299	25.1%
F	3,283	38.2%	335	30.7%	239	21.9%
Total	76,061	43.1%	8,203	37.9%	6,677	30.8%

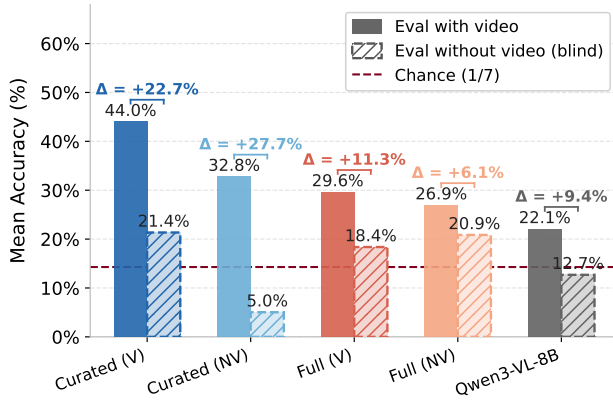


Figure 11. Training-set curation ablation. Mean accuracy on $\mathcal{D}_{val}^*$ validation set under evaluation *with* video (solid bars) and *blind*, without video (hatched bars), for models trained on the full uncured set $\mathcal{D}_{ref}$ (*Full*) vs. the curated set SUREON-DB (*Curated*), together with the off-the-shelf Qwen3-VL-8B baseline. (V) and (NV) denote video-trained vs. text-only-trained

the hardened validation set. Training on the curated set is expected to yield higher vision-enabled accuracy, as the model is no longer rewarded for exploiting language shortcuts during training. Figure 11 reports vision and blind accuracy under both training conditions.

B.5. Stage 4: Expert Verification Details

The hardened validation set $\mathcal{D}_{val}^*$ produced by Stage 3 is submitted to manual verification by expert surgeons; no additional LLM-based filtering or adjudication is applied at this stage. Only samples that pass expert review are retained in the final SUREON benchmark $\mathcal{B}_{SUREON}$.

B.5.1. Annotation Platform

Verification is conducted on a dedicated web-based annotation platform. Figure 12 shows the main verification screen, which presents: (i) the video clip and its timestamp, (ii) the

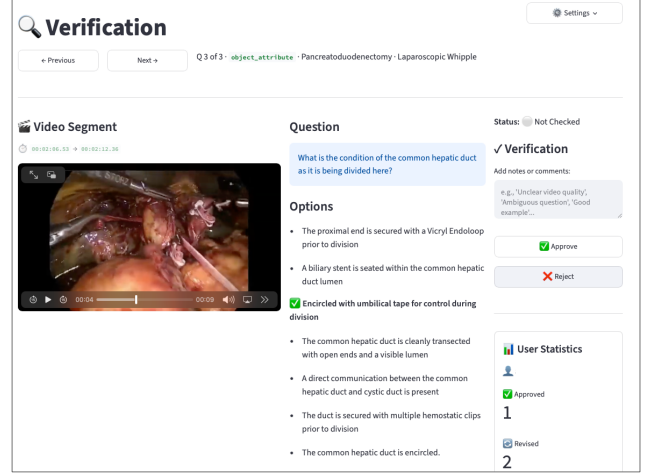


Figure 12. Main verification screen of the annotation platform. Reviewers are presented with the video clip, the question, all answer options (correct answer marked in green), and the approve/reject interface.

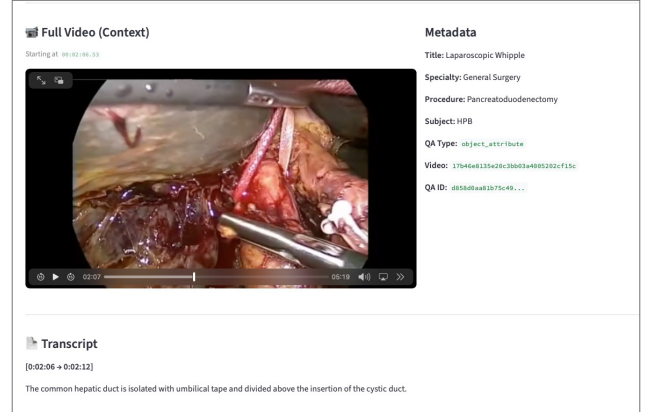


Figure 13. Extended context view, accessible by scrolling below the verification screen. Provides the full surgery video, aligned transcript, and video-level metadata.

question with all answer options (the correct answer marked in green), and (iii) an approve/reject interface requiring a written note for every rejection. Figure 13 shows the extended context view, available by scrolling below the verification screen, which provides the full surgery video (starting at the clip’s timestamp), the aligned transcript segment, and video-level metadata including title, specialty, procedure, and QA type. Reviewers are encouraged to consult the full video whenever the short clip is insufficient to make a judgement. The platform also exposes specialty, procedure, and QA type filters, allowing reviewers to focus on samples within their clinical domain.

B.5.2. Reviewer Panel and Protocol

Verification is conducted by 6 surgeons affiliated to different institutions.

Sample assignment. Each sample is assigned to a single reviewer, drawn uniformly at random from the pool of available samples across the different question types, and reviewers may skip any sample they consider outside their domain. A single reviewer per sample is sufficient because reviewers do not annotate from scratch: each candidate QA is already transcript-grounded, with its correct answer derived from a surgeon’s narration (Sec. B.2). Verification is therefore a confirmation step rather than an open-ended labeling task, in which the reviewer checks video–transcript alignment, the clinical correctness of the narration-derived answer, and distractor validity for the specific clip. Because the underlying supervision already originates from a domain expert, a single qualified reviewer suffices to confirm these conditions, and we apply a conservative retention policy under which any rejected or flagged sample is discarded, so that residual subjectivity removes samples rather than admitting incorrect ones.

Inter-rater reliability. Although each benchmark candidate was reviewed by one surgeon in the primary verification protocol, we independently dual-reviewed a subset of 74 QA samples distributed among QA types to estimate inter-rater agreement. Each sample was independently reviewed by two raters drawn from a pool of five surgeons, with reviewer pairs assigned to ensure balanced coverage across all $C(5,2)=10$ possible pairings. The two reviewers independently rendered the same binary approve/reject decision and agreed on 52 of 74 samples (70.3% agreement). The duplicate ratings were used only to estimate agreement and did not alter the primary single-review retention decision.

Review criteria. Each reviewer watches the clip and assesses the following four conditions, all of which must hold simultaneously for a sample to be approved:

- *Question soundness.* The question is clear, clinically meaningful, and answerable from the specific clip or segment provided.
- *Answer correctness.* The marked answer is clinically correct and is visible or directly supported within the clip’s time window.
- *Video–transcript alignment.* The QA reflects what is actually shown in the clip, and the clip depicts a real operative scene rather than, e.g., a slide or diagram.
- *Distractor validity.* All distractors are clinically plausible for the context but unambiguously incorrect for this specific clip. Any sample for which a second option could also be correct is rejected.

Reviewers are additionally asked to confirm that the question *requires* the video to answer and cannot be resolved from generic procedural knowledge alone, serving as a hu-

Table 12. Expert verification results per category. *Reviewed*: samples assessed. *Approved*: retained in $\mathcal{B}_{\text{SUREON}}$. *Rejected*: samples that failed review. *Rate*: approved / reviewed.

Category	Reviewed	Approved	Rejected	Rate
AD	62	35	27	56.5%
EA	63	34	29	54.0%
EE	55	37	18	67.3%
LAR	59	37	22	62.7%
TO	69	34	35	49.3%
PAD	65	29	36	44.6%
SM	51	41	10	80.4%
IAI	62	31	31	50.0%
SPI	53	31	22	58.5%
F	63	31	32	49.2%
Total	602	340	262	56.5%

man analog of the blind-testing filter applied in Stage 3. We encourage rejected samples to include a brief written note identifying the reason for rejection.

For Forecasting questions, the correct answer describes the step that occurs *after* the clip ends and therefore cannot be verified from the clip alone. Reviewers are instructed to note the clip’s end timestamp, open the full video, and confirm that the marked next step actually occurs immediately after the clip before approving.

Quantitative summary. Table 12 reports the number of reviewed, approved, and rejected samples per category, along with the approval rate. The experts provided rejection reasons for 178 and the breakdown is distributed as follows: 22.5% flagged as ambiguous/multiple-correct, 14.6% inaccurate, 9.6% too-short or insufficient context, and 9.6% not visible in the video. The final benchmark $\mathcal{B}_{\text{SUREON}}$ comprises all approved samples across categories. Expert verification required approximately 32 hours of cumulative surgeon time across all reviewers.

B.6. Final Dataset Configuration

We consolidate the configuration retained after all four stages; per-stage counts are in Tab. 11 and key attributes in Tab. 2. The pipeline yields two resources from a common process. SUREON-DB is the curated training database: of the 2,211 training videos with validated Stage-1 QA, 2,191 retain at least one sample after blind-testing curation (Sec. B.4), giving 76,061 QA pairs. The SUREON benchmark is the expert-verified set: the hardened validation set $\mathcal{D}_{\text{val}}^*$ is reviewed by surgeons (Sec. B.5), and the 340 approved samples, from 165 distinct source videos, form $\mathcal{B}_{\text{SUREON}}$. The two splits share no source videos, as splitting is video-level prior to generation (Sec. B). Figure 14 shows the hierarchical distribution of both resources across surgical specialties and subjects.

Table 13. Public surgical datasets transformed into $\mathcal{D}_{\text{pub}}$. *Branch* indicates whether a dataset contributes frame-level (*Img*), clip-level (*Clip*), or both (*Both*) supervision; counts are the number of source items (frames/clips) and the QA pairs produced after templating and augmentation.

Dataset	Branch	#Items	#QA
CholecT50 [41]	Img	72,815	1,919,383
CholecInstanceSeg [1]		26,830	834,398
CholecSeg8k [26]		5,600	321,268
EndoVis2017 [2]		1,800	108,282
EndoVis2018 [3]		1,639	126,343
Endoscapes [39]		6,960	124,656
LapGyn [30]		53,107	225,350
Dresden [11]		8,050	55,568
HemoSet [37]		828	8,280
GraSP [5]	Both	4,340	293,736
AutoLaparo [64]		1,336	82,524
GynSurg [40]		4,313	62,703
HeiChole [59]		2,064	57,405
HeiCo - Proctocolectomy [36]		1,681	40,281
HeiCo - Rectal Resection [36]		2,293	52,089
HeiCo Sigmoid Resection [36]		1,316	31,656
Phakir [45]		14,447	499,099
Cholec80 [58]	Clip	3,834	85,799
MultiBypass-Bern [29]		3,598	174,270
MultiBypass-Strasbourg [29]		3,907	226,995
SARRARP50 [43]		2,676	171,886
SurgicalActions160 [48]		157	1,260
Total (22 datasets)		223,591	5,503,231

C.2. Stage 2a: Image-Level QA Template Generation

Segmentation-to-instance conversion. Several datasets provide only semantic segmentation masks. For each class we apply a conversion policy: *single-instance* classes (anatomical structures known to appear once, such as the gallbladder or uterus) are treated as a single pseudo-instance by taking the union of all connected components, then fitting a tight bounding box; *connected-components* classes (blood pools, clips) are split into individual pseudo-instances via 8-connected-component labeling, sorted top-to-bottom then left-to-right by centroid; and *strict-semantic* classes (large background structures such as the abdominal wall or liver bed) are retained as presence-only annotations without bounding-box assignment. All bounding boxes are normalized to a $[0, 1000]$ coordinate space and stored in $[x_1, y_1, x_2, y_2]$ format.

QA types. Starting from the normalized frame annotations, we instantiate task-specific QA templates across two families. *Open-ended templates* produce free-text answers and cover: tool/anatomy presence (“Which surgical instruments are visible?”); counting (“How many surgical instruments are visible?”); view-region localization (“Where is

the {tool} with respect to the surgical field?”); action presence (“What action is happening in this frame?”); triplet recognition (“What is the {tool} doing?”); Critical View of Safety (three binary criteria for safe cholecystectomy); mask classification (identify structures from rendered contour overlays); and procedure-type recognition. Binary and negative variants are generated by sampling negative class labels from the dataset vocabulary.

Grounded templates. Grounded templates instead require structured JSON answers, which lets us pose localization and recognition tasks with machine-checkable outputs. They span localization (returning a JSON array of labeled, indexed bounding boxes for the visible tools or anatomy), presence (a dictionary flagging each class as present or absent), counting (the number of tools in view), phase and step recognition (the active phase or step), triplet recognition (the active instrument–verb–target relations), and action recognition (the action performed by each tool in view). In every grounded template the question explicitly enumerates the full class vocabulary as the option set, anchoring the model to a closed label space. Each QA pair is serialized as a JSONL record with the question, answer, an optional reasoning trace (omitted for non-CoT types), and provenance fields identifying the QA type, source dataset, file, video, and frame.

C.3. Stage 2b: Video Clip QA Template Generation

Clip sampling. For temporal datasets we sample variable-length clips with two strategies. By default, *strategic sampling* biases selection toward phase and step transition boundaries and stable mid-phase regions, so that clips capture both the moments of change and the steady states between them. Alternatively, *grounded clip sampling* slides a fixed 8-frame window with a stride of four and keeps a clip only if it lies entirely within a single constant-label segment across all requested tasks (phase, step, action), guaranteeing an unambiguous answer.

Language templates. For each sampled clip, a clip-description generator draws at random from several templates and fills them with the clip’s annotations. Phase and step templates describe which phase or step the clip shows, and optionally enrich this with the procedure type, the temporal position within the phase (beginning, middle, or end), or both; clips that span a boundary instead use transition templates that name the two phases and the approximate time of the switch, the procedure transitions from $\{\text{phase}_1\}$ to $\{\text{phase}_2\}$). Step templates extend this to single-, two-, and multi-step sequences, action templates to the actions present, and tool templates to which instruments appear and when they are used (e.g., “{tool} is used from {start} s to {end} s”).

Question types. The resulting questions are descriptive (“What is happening in this clip?”), binary (“Is the {phase}

phase shown in this clip?”), temporal (“When does {event} occur in this clip?”), or counting (“How many distinct phases are shown?”). As in the image branch, grounded clip questions take a structured JSON answer (e.g., the active phase). For each dataset we keep three versions of every QA pair — the direct template output, an instruction-style rephrasing, and an LLM-augmented variant (Sec. C.4) — which are pooled at training time.

C.4. Stage 3: LLM-Based Linguistic Augmentation

To increase linguistic diversity without altering factual content, we paraphrase each templated QA pair with an Qwen3.5-9B [57], adding one rephrased variant per pair to the training pool.

Image branch. We use Qwen3.5-9B [57] ($T=0.7$, $p=0.8$, $k=20$, up to 256 new tokens) to rephrase the question and answer under three constraints: surgical terminology is preserved exactly (*phase*, *step*, *procedure*, and *action* are never interchanged), the question retains explicit visual-grounding cues (e.g., “in this image”, “visible in the frame”), and any ordered or list structure in the original is kept intact. For tasks with structured answers (localization, presence, triplet, and mask classification), only the question is paraphrased and the answer is copied verbatim, so it remains valid and label-consistent. For chain-of-thought samples, the accompanying reasoning is paraphrased alongside the question and answer.

Clip branch. We use Qwen3-VL-8B-Instruct [8] in text-only mode ($T=0.7$, $p=0.9$, up to 256 new tokens) with an analogous prompt that additionally requires each question to refer to a video clip or segment and, as above, keeps *phase*, *step*, and *action* distinct.

D. Evaluation

This section specifies the inference and scoring protocols used in our experiments (Sec. 4), covering the SUREON benchmark (Tab. 3) and the public perception benchmarks for recognition and spatial grounding (Tab. 5, Tab. 19). All close-ended questions and Temporal Ordering are evaluated under accuracy. The open-ended Sequence Summarization (SM) type is scored by an LLM judge (Claude Opus 4.8) using the prompt in Fig. 15. We organize the protocols into evaluation on SUREON (Sec. D.1) and on public benchmarks (Sec. D.3).

D.1. SUREON Benchmark Evaluation

Frame sampling and inference. Every model receives the same visual input for a given question. For each clip we uniformly sample up to 8 frames across its duration at the video’s frame rate, clamped to a minimum of 4 frames. Before inference, we apply a two-stage blurring pass to the sampled frames: a YOLO-based non-surgical object detector identifies non-operative regions (faces, equipment lo-

```

You are grading a model's summary of a (surgical/procedural) video clip against a
reference summary. You see only the two texts, NOT the video, so you cannot verify
any claim that goes beyond the reference.

Question: {question}
Ground Truth Answer: {ground_truth}
Predicted Answer: {predicted}

Your task:
1. Decide if the predicted answer is CORRECT or INCORRECT relative to the ground truth.

Judge on two things, in this priority order:

a) SEQUENCE & COVERAGE (primary): Do the predicted steps/structures/actions match
the reference in content and order?
- Be semantically flexible: accept paraphrases, reordered lists where order is
not meaningful, and minor wording differences if the meaning matches.
- Mark INCORRECT if any key step, structure, or item from the reference is
missing, or if the order is wrong when order matters.

b) FAITHFULNESS (secondary): Does the prediction stay consistent with the reference?
- Detail present in the prediction but ABSENT from the reference (extra anatomy,
named techniques, planes, landmarks) is UNVERIFIABLE, not automatically right
or wrong. Do not reward it as richness and do not penalize it as hallucination
on its own — but note it in the assessment.
- Treat surgical synonyms for the SAME action as matching: identify / expose /
protect / preserve / safeguard a structure are equivalent framings of one step.
- Mark INCORRECT only if added detail CONTRADICTS the reference, or if the
prediction substitutes a DIFFERENT action on a structure — e.g. reference says
a structure is identified/preserved but prediction says it is divided, transected,
ligated, or removed. Safeguarding vs. cutting is a real difference: identifying
vs. preserving is not.
- A fluent, textbook-style elaboration that matches how the procedure is usually
done is a red flag for ungrounded generation, but is grounds for INCORRECT only
when it conflicts with or substitutes for what the reference states.

Decision rule: CORRECT if (a) holds and (b) shows no contradiction or
meaning-changing substitution. Otherwise INCORRECT.

2. Write a brief assessment (1-3 sentences) explaining your decision. If you pass a
prediction that contains unverifiable added detail, say so explicitly.

Respond in EXACTLY this format (no other text):
Verdict: <correct|incorrect>
Assessment: <your explanation>

```

Figure 15. LLM judge prompt to evaluate answers to the open-ended sequence summarization question type.

gos, and other non-surgical content) trained for YouTube videos [12], and optical character recognition detects on-screen textual overlays such as procedure labels, instrument names in the user interface, and equipment text; both sets of detected regions are blurred. Surgical academic videos can contain these overlays, which modern VLMs may exploit through OCR to infer answers without interpreting the underlying surgical scene; blurring them prevents this source of leakage and better isolates visual surgical understanding. Each frame in the prompt is tagged with its absolute timestamp, preceded by the clip’s time range, followed by the question and the option list (prompt template in Fig. 17). The question text, option set, and option order are held fixed across all evaluated models.

Option format. Multiple-choice questions are presented with seven options giving a random baseline of $1/7 = 14.3\%$. Temporal Ordering (TO) requires ordering three events, for which the random baseline over the $3! = 6$ permutations is $1/6 = 16.7\%$. Distractor selection and option shuffling are seeded per sample for reproducibility.

Blind condition and multimodal gain. The blind condition removes the visual input entirely: the model receives a text-only payload containing the identical question and option list, with no frames. We report multimodal gain $MG = S_v - S_{wv}$ as the accuracy difference between the visual (S_v) and blind (S_{wv}) passes.

Scoring. The eight multiple-choice types are scored by exact-match accuracy against the correct option. For TO, we extract the predicted sequence of event labels from the

free-form output and require exact, order-sensitive equality with the reference sequence. SM is scored by the LLM judge described below. Reported μ is the pooled accuracy over the 299 accuracy-scored items; SM is reported separately and excluded from μ .

Human baseline protocol. The human baseline is collected from 5 surgeons, a subset of the 6-surgeon panel used for Stage 4 expert verification; Sequence Summarization is excluded from the baseline due to the time cost of open-ended grading, leaving 299 accuracy-scored items. Baseline collection uses a separate interface from the verification platform: each surgeon is shown only the video clip, the question, and the answer options (no blind testing mode for expert baseline). Items are distributed uniformly at random across QA types for each surgeon, without matching raters to their own specialty, as our reviewer pool is not large enough to guarantee specialty coverage per item. Each item is answered by a single surgeon, and the reported score aggregates across raters. To avoid trivial recall, no surgeon is shown a question they had already adjudicated during expert verification. After each answer, surgeons self-report confidence (Confident / Somewhat / Not confident); across recorded responses the breakdown is 31.4% / 37.5% / 17.7%, with 13.4% not recorded. Total collection required approximately 8.5 hours of cumulative surgeon time. Several surgeons reported difficulty answering questions outside their own specialty, consistent with the per-category gap between near-ceiling performance on Procedure-Agnostic Action Description and comparatively weaker performance on categories that require specialty-specific understanding, which also motivates a surgical VLM capable of breadth across specialties, which no single human specialist reviewer possesses.

Uncertainty and statistical comparisons. We quantify uncertainty in the SUREON results using 95% confidence intervals. For the overall accuracy, we verify our non-parametric bootstrap intervals (10,000 resamples over benchmark items) against the analytic Wilson score interval. The two estimates closely agree, with a mean absolute discrepancy of 0.25 percentage points and a maximum of 0.54 points across all model estimates. We additionally bootstrap accuracy separately for each question category and for the perception and reasoning aggregates, and we report a surgeon baseline collected under the same video-based protocol as the model evaluations. Because model (and surgeon) predictions are paired on the same benchmark samples, pairwise differences are tested using McNemar’s test with Holm–Bonferroni correction across all $\binom{12}{2} = 66$ model pairs (11 evaluated models and the human baseline). Tables 14, 15, 16, and 17 report the corresponding confidence intervals; Table 18 reports the full pairwise significance matrix underlying the claims above.

Category-level estimates are necessarily less precise be-

Table 14. Overall accuracy and bootstrap 95% confidence intervals on the 299 accuracy-scored SUREON benchmark samples.

Model	n	Accuracy (%)	Bootstrap 95% CI (%)
Human	299	64.9	[59.5, 70.2]
Gemini 3.1 Pro	299	52.2	[46.8, 57.9]
SUREON-VLM	299	48.2	[42.5, 53.8]
GPT-5.1 (high)	299	46.8	[41.1, 52.5]
SUREON-VLM (SUREON-only)	299	44.1	[38.5, 49.8]
Gemma 4	299	35.8	[30.4, 41.1]
Qwen3VL-32B-T	299	34.1	[28.8, 39.5]
Qwen3VL-8B-I	299	23.1	[18.4, 27.8]
SUREON-VLM (Public-only)	299	20.7	[16.4, 25.4]
MedGRPO	299	16.7	[12.7, 21.1]
Molmo2	299	15.0	[11.0, 19.1]
MedGemma 1.5	299	10.7	[7.4, 14.4]

Table 15. Accuracy and bootstrap 95% confidence intervals for the perception and reasoning subsets of SUREON. Perception contains $n = 166$ samples and reasoning contains $n = 133$.

Model	Perception (%)	Reasoning (%)
Human	72.9 [66.3, 79.5]	54.9 [46.6, 63.2]
Gemini 3.1 Pro	53.0 [45.2, 60.8]	51.1 [42.9, 59.4]
SUREON-VLM	48.8 [41.6, 56.0]	47.4 [39.1, 55.6]
GPT-5.1 (high)	50.0 [42.2, 57.8]	42.9 [34.6, 51.1]
SUREON-VLM (SUREON-only)	42.8 [35.5, 50.6]	45.9 [37.6, 54.1]
Gemma 4	39.8 [32.5, 47.0]	30.8 [23.3, 39.1]
Qwen3VL-32B-T	42.2 [34.3, 49.4]	24.1 [17.3, 31.6]
Qwen3VL-8B-I	26.5 [19.9, 33.1]	18.8 [12.0, 25.6]
SUREON-VLM (Public-only)	27.1 [20.5, 33.7]	12.8 [7.5, 18.8]
MedGRPO	19.3 [13.2, 25.3]	13.5 [8.3, 19.6]
Molmo2	20.5 [14.5, 27.1]	8.3 [3.8, 13.5]
MedGemma 1.5	12.0 [7.2, 17.5]	9.0 [4.5, 14.3]

cause each accuracy-based category contains only 29–37 samples, compared with 299 samples for overall μ . Across the 12 models, the average bootstrap 95% CI half-width is 14.3 percentage points, compared with approximately 5.0 percentage points overall. We therefore use category-level results primarily to characterize task-specific strengths and weaknesses rather than to establish fine-grained model rankings. Perception ($n = 166$) and reasoning ($n = 133$) aggregates provide intermediate precision. Notably, the surgeon baseline attains near-ceiling accuracy on PAD (96.5%) but is markedly lower on EE (51.3%, statistically indistinguishable from several models), suggesting category difficulty is not uniform even for expert humans.

LLM-judge for Sequence Summarization. The SM judge scores each prediction against a fixed transcript-derived reference, not the judge’s own surgical knowledge. Seeing only the two texts, it is prompted to be strict (Fig. 15): a prediction is incorrect if any reference step is missing, if meaningful ordering is violated, or if an action is substituted (e.g. a structure the reference preserves is reported as divided), while paraphrases and surgical synonyms count as matching.

LLM Judge Validation. To validate the judge, three anno-

Table 16. Per-category accuracy and bootstrap 95% confidence intervals for perception-oriented SUREON question types. Entries report accuracy [95% CI], in percent. AD: Action Description; IAI: Instrument–Action Interaction; EA: Entity (Object) Attribute; EE: Entity (Object) Existence; PAD: Procedure-Agnostic Action Description.

Model	AD ($n = 35$)	IAI ($n = 31$)	EA ($n = 34$)	EE ($n = 37$)	PAD ($n = 29$)
Human	74.3 [60.0, 88.6]	74.2 [58.1, 87.1]	73.5 [58.8, 88.2]	51.3 [35.1, 67.6]	96.5 [89.7, 100.0]
Gemini 3.1 Pro	54.3 [37.1, 71.4]	54.8 [38.7, 71.0]	41.2 [23.5, 58.8]	48.6 [32.4, 64.9]	69.0 [51.7, 86.2]
SUREON-VLM	65.7 [48.6, 80.0]	38.7 [22.6, 54.8]	47.1 [29.4, 64.7]	48.6 [32.4, 64.9]	41.4 [24.1, 58.6]
GPT-5.1 (high)	57.1 [40.0, 74.3]	61.3 [45.2, 77.4]	38.2 [23.5, 55.9]	54.0 [37.8, 70.3]	37.9 [20.7, 55.2]
SUREON-VLM (SUREON-only)	51.4 [34.3, 68.6]	25.8 [9.7, 41.9]	44.1 [26.5, 61.8]	48.6 [32.4, 64.9]	41.4 [24.1, 58.6]
Gemma 4	48.6 [31.4, 65.7]	41.9 [25.8, 58.1]	32.4 [17.6, 50.0]	27.0 [13.5, 40.5]	51.7 [34.5, 69.0]
Qwen3VL-32B-T	40.0 [22.9, 57.1]	45.2 [29.0, 61.3]	41.2 [23.5, 58.8]	27.0 [13.5, 43.2]	62.1 [44.8, 79.3]
Qwen3VL-8B-I	28.6 [14.3, 42.9]	38.7 [22.6, 54.8]	17.6 [5.9, 32.4]	16.2 [5.4, 29.7]	34.5 [17.2, 51.7]
SUREON-VLM (Public-only)	42.9 [25.7, 60.0]	29.0 [12.9, 45.2]	14.7 [2.9, 26.5]	16.2 [5.4, 29.7]	34.5 [17.2, 51.7]
MedGRPO	25.7 [11.4, 40.0]	25.8 [12.9, 41.9]	17.6 [5.9, 32.4]	16.2 [5.4, 29.7]	10.3 [0.0, 24.1]
Molmo2	31.4 [17.1, 48.6]	16.1 [3.2, 29.0]	20.6 [8.8, 35.3]	10.8 [2.7, 21.6]	24.1 [10.3, 41.4]
MedGemma 1.5	11.4 [2.9, 22.9]	6.5 [0.0, 16.1]	8.8 [0.0, 20.6]	10.8 [2.7, 21.6]	24.1 [10.3, 41.4]

Table 17. Per-category accuracy and bootstrap 95% confidence intervals for the accuracy-scored reasoning SUREON question types. Entries report accuracy [95% CI], in percent. LAR: Local Action Reasoning; SPI: Safety Practice Identification; F: Forecasting; TO: Temporal Ordering. Sequence Summarization is open-ended and LLM-judge scored, and is excluded here.

Model	LAR ($n = 37$)	SPI ($n = 31$)	F ($n = 31$)	TO ($n = 34$)
Human	35.1 [18.9, 51.3]	71.0 [54.8, 87.1]	54.8 [38.7, 71.0]	61.8 [44.1, 76.5]
Gemini 3.1 Pro	56.8 [40.5, 73.0]	61.3 [45.2, 77.4]	19.4 [6.5, 32.3]	64.7 [50.0, 79.4]
SUREON-VLM	43.2 [27.0, 59.5]	58.1 [41.9, 74.2]	35.5 [19.4, 51.6]	52.9 [35.3, 70.6]
GPT-5.1 (high)	46.0 [29.7, 62.2]	45.2 [29.0, 64.5]	25.8 [9.7, 41.9]	52.9 [35.3, 70.6]
SUREON-VLM (SUREON-only)	40.5 [24.3, 56.8]	61.3 [45.2, 77.4]	38.7 [22.6, 54.8]	44.1 [26.5, 61.8]
Gemma 4	29.7 [16.2, 46.0]	35.5 [19.4, 51.6]	16.1 [3.2, 29.0]	41.2 [23.5, 58.8]
Qwen3VL-32B-T	27.0 [13.5, 40.5]	19.4 [6.5, 35.5]	12.9 [3.2, 25.8]	35.3 [20.6, 52.9]
Qwen3VL-8B-I	18.9 [8.1, 32.4]	16.1 [3.2, 29.0]	12.9 [3.2, 25.8]	26.5 [11.8, 41.2]
SUREON-VLM (Public-only)	10.8 [2.7, 21.6]	19.4 [6.5, 35.5]	6.5 [0.0, 16.1]	14.7 [2.9, 26.5]
MedGRPO	13.5 [2.7, 24.3]	12.9 [3.2, 25.8]	6.5 [0.0, 16.1]	20.6 [8.8, 35.3]
Molmo2	8.1 [0.0, 18.9]	12.9 [3.2, 25.8]	3.2 [0.0, 9.7]	8.8 [0.0, 17.6]
MedGemma 1.5	10.8 [2.7, 21.6]	6.5 [0.0, 16.1]	6.5 [0.0, 16.1]	11.8 [2.9, 23.5]

tators independently rated 44 stratified predictions sampled across all 11 models (4 per model, balanced by verdict). Raters were shown only the question, reference, and extracted prediction and applied the same two-criterion rubric. Inter-rater agreement was moderate (mean pairwise raw agreement 64.5%), consistent with the subjectivity of surgical sequence evaluation. Majority-vote human verdicts agreed with the LLM judge on 79.1% of samples. Disagreements were asymmetric: in 6 of the 10 discordant cases the judge marked a prediction incorrect that humans accepted, versus 3 in the opposite direction, indicating that the judge applies the rubric more strictly than human raters.

D.2. Qualitative Error Analysis

Figure 16 presents representative SUREON-VLM failures across four perception and four reasoning categories, revealing two distinct error modes.

Perception errors are near-misses, not random guesses. Predictions consistently land in the correct local neighborhood of the ground truth rather than on implausible alternatives. In Action Description, the model correctly locates clip application in Calot’s triangle but confuses the cystic *artery* for the cystic *duct*. In Entity Existence, “Toldt space” is predicted for “retroperitoneal space,” an adjacent landmark. These suggest coarse scene understanding without the fine-grained discrimination the benchmark targets. The Instrument–Action example is a sharper failure: the model predicts the suction irrigator’s *canonical* function (clearing blood) over its atypical use in this clip (blunt dissection), which is a language prior-driven error rather than a visual one, and a reminder that dataset-level shortcut removal (RADG, blind-testing) still has limitations in preventing a trained model from defaulting to real-world instrument priors when visual attention is weak.

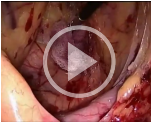
Perception			
			
Q: What action is performed by the suction irrigator in this clip?	Q: What space is visualized after plane entry using gauze and insufflation in this clip?	Q: What is the surgeon doing in this segment?	Q: Which procedure-agnostic surgical action is being performed in this clip?
GT: Bluntly dissect tissue planes	GT: Retroperitoneal space	GT: Applying clips to the cystic duct	GT: Dissecting tissue
Pred: Suction blood to clear the field	Pred: Toldt space	Pred: Applying clips to the cystic artery	Pred: Transecting
Instrument-Action Interaction	Entity Existence	Action Description	Procedure-agnostic Action Description
Reasoning			
			
Q: Determine the chronological arrangement of these intraoperative actions: A: 'Ureter is dissected and shifted medially', B: 'Paravesical space is created', C: 'Inferior mesenteric artery is clipped and divided'	Q: In this scene, how is the team reducing risk of bleeding?	Q: Considering the surgical progress up to this point, what is the next step the team is likely to take?	Q: What is the immediate purpose of keeping the saphenous vein at the base of the dissection in this segment?
GT: C, A, B	GT: Applying large vascular clips to secure the pedicle and prevent back bleeding	GT: Begin the duct-to-limb anastomosis with a running suture	GT: To maintain the correct plane and avoid getting too close to the skin
Pred: B, C, A	Pred: Elevating the prostate to create a right angle and then dividing the vascular pedicle	Pred: Place initial corner sutures to start the anastomosis at the head end	Pred: To maintain orientation within the surgical field
Temporal Ordering	Safety Practice Identification	Forecasting	Local Action Reasoning

Figure 16. **Representative SUREON-VLM failure cases** across four perception categories (Instrument-Action Interaction, Entity Existence, Action Description, Procedure-Agnostic Action Description) and four reasoning categories (Temporal Ordering, Safety Practice Identification, Forecasting, Local Action Reasoning). For each example we show the question (Q), SUREON-VLM’s prediction (Pred), and the ground-truth answer (GT).

Reasoning errors tend toward generic, plausible-sounding answers. In Safety Practice Identification, the model describes a real but different aspect of the scene (a dissection step) rather than the specific hemostatic action asked about. Local Action Reasoning shows this most clearly: “maintain orientation within the surgical field” is a generic justification that could apply to nearly any dissection, where the ground truth requires reasoning about a specific plane and proximity to skin. These are fluent, surgically literate answers that fail on specificity. Forecasting errors are comparatively softer, correctly anticipating the coarse next phase while erring on technical detail.

D.3. Public Perception Benchmark Evaluation

Inputs and clip construction. Public benchmarks are evaluated with three input modes depending on the task. Frame-level tasks (tool localization and frame-level phase/action recognition) use a single frame. Clip-level recognition uses a fixed window of 8 frames at a stride of 4 (spanning up to ~28 s at 1 fps extraction) anchored on the keyframe. We evaluate all clip tasks in the online setting, where the keyframe is the final frame of the window with preceding frames providing causal context. SUREON-VLM is evalu-

```

You are viewing frames from a surgical video (time range: {start} to {end}).

(timestamp: HH:MM:SS.mmm)
[frame 1]
(timestamp: HH:MM:SS.mmm)
[frame 2]
...
(timestamp: HH:MM:SS.mmm)
[frame N]

(question)

Select from these options:
- (option 1)
- (option 2)
- (option 3)
- (option 4)
- (option 5)
- (option 6)
- (option 7)

```

Figure 17. Prompt template used to query all models on the SUREON benchmark. Frames are interleaved with their timestamps, followed by the question and seven answer options.

ated in both frame and clip variants where applicable. Commercial and open-source baselines are evaluated in frame-only mode on public benchmarks—commercial models due to inference costs and surgical models due to unavailable weights—using numbers reported in their original papers.

Recognition and classification metrics. Phase and action recognition are scored by video-averaged macro-F1 (per-video macro-F1 averaged over videos), which avoids long videos dominating the aggregate. For Cholec80 we

Table 18. Full pairwise significance matrix. McNemar’s test on paired predictions, Holm–Bonferroni corrected across all $\binom{12}{2} = 66$ model pairs. Rows are sorted by corrected p -value (most significant first); * denotes significance at the Holm-corrected 0.05 level.

Model A	Model B	Δ Acc. (pts)	p (raw)	P_{holm}
Human	MedGemma 1.5	54.18	<0.001	<0.001*
Human	Molmo2	49.83	<0.001	<0.001*
Human	MedGRPO	48.16	<0.001	<0.001*
Human	SUREON-VLM (Public-only)	44.15	<0.001	<0.001*
Human	Qwen3VL-8B-I	41.81	<0.001	<0.001*
Gemini 3.1 Pro	MedGemma 1.5	41.47	<0.001	<0.001*
MedGemma 1.5	SUREON-VLM	-37.46	<0.001	<0.001*
Gemini 3.1 Pro	Molmo2	37.12	<0.001	<0.001*
GPT-5.1 (high)	MedGemma 1.5	36.12	<0.001	<0.001*
Gemini 3.1 Pro	MedGRPO	35.45	<0.001	<0.001*
MedGemma 1.5	SUREON-VLM (SUREON-only)	-33.44	<0.001	<0.001*
Molmo2	SUREON-VLM	-33.11	<0.001	<0.001*
GPT-5.1 (high)	Molmo2	31.77	<0.001	<0.001*
Gemini 3.1 Pro	SUREON-VLM (Public-only)	31.44	<0.001	<0.001*
MedGRPO	SUREON-VLM	-31.44	<0.001	<0.001*
Human	Qwen3VL-32B-T	30.77	<0.001	<0.001*
GPT-5.1 (high)	MedGRPO	30.10	<0.001	<0.001*
Human	Gemma 4	29.10	<0.001	<0.001*
Gemini 3.1 Pro	Qwen3VL-8B-I	29.10	<0.001	<0.001*
Molmo2	SUREON-VLM (SUREON-only)	-29.10	<0.001	<0.001*
MedGRPO	SUREON-VLM (SUREON-only)	-27.42	<0.001	<0.001*
SUREON-VLM (Public-only)	SUREON-VLM	-27.42	<0.001	<0.001*
GPT-5.1 (high)	SUREON-VLM (Public-only)	26.09	<0.001	<0.001*
Qwen3VL-8B-I	SUREON-VLM	-25.08	<0.001	<0.001*
Gemma 4	MedGemma 1.5	25.08	<0.001	<0.001*
GPT-5.1 (high)	Qwen3VL-8B-I	23.75	<0.001	<0.001*
Qwen3VL-32B-T	MedGemma 1.5	23.41	<0.001	<0.001*
SUREON-VLM (SUREON-only)	SUREON-VLM (Public-only)	23.41	<0.001	<0.001*
Qwen3VL-8B-I	SUREON-VLM (SUREON-only)	-21.07	<0.001	<0.001*
Human	SUREON-VLM (SUREON-only)	20.74	<0.001	<0.001*
Gemma 4	Molmo2	20.74	<0.001	<0.001*
Gemma 4	MedGRPO	19.06	<0.001	<0.001*
Qwen3VL-32B-T	Molmo2	19.06	<0.001	<0.001*
Human	GPT-5.1 (high)	18.06	<0.001	<0.001*
Gemini 3.1 Pro	Qwen3VL-32B-T	18.06	<0.001	<0.001*
Qwen3VL-32B-T	MedGRPO	17.39	<0.001	<0.001*
Gemini 3.1 Pro	Gemma 4	16.39	<0.001	<0.001*
Gemma 4	SUREON-VLM (Public-only)	15.05	<0.001	<0.001*
Qwen3VL-8B-I	MedGemma 1.5	12.37	<0.001	<0.001*
Human	SUREON-VLM	16.72	<0.001	0.0027*
Qwen3VL-32B-T	SUREON-VLM (Public-only)	13.38	<0.001	0.0027*
Qwen3VL-8B-I	Gemma 4	-12.71	<0.001	0.0027*
Qwen3VL-32B-T	SUREON-VLM	-14.05	<0.001	0.0072*
Qwen3VL-32B-T	Qwen3VL-8B-I	11.04	<0.001	0.0092*
GPT-5.1 (high)	Qwen3VL-32B-T	12.71	<0.001	0.0110*
MedGemma 1.5	SUREON-VLM (Public-only)	-10.03	<0.001	0.0126*
Gemma 4	SUREON-VLM	-12.37	<0.001	0.0140*
Human	Gemini 3.1 Pro	12.71	0.0013	0.0247*
GPT-5.1 (high)	Gemma 4	11.04	0.0013	0.0247*
Qwen3VL-8B-I	Molmo2	8.03	0.0022	0.0374*
Qwen3VL-32B-T	SUREON-VLM (SUREON-only)	-10.03	0.0064	0.1024
Gemma 4	SUREON-VLM (SUREON-only)	-8.36	0.0261	0.3915
Gemini 3.1 Pro	SUREON-VLM (SUREON-only)	8.03	0.0323	0.4522
MedGemma 1.5	MedGRPO	-6.02	0.0414	0.5382
Qwen3VL-8B-I	MedGRPO	6.35	0.0448	0.5382
Molmo2	SUREON-VLM (Public-only)	-5.69	0.0533	0.5863
Molmo2	MedGemma 1.5	4.35	0.1175	1.0000
GPT-5.1 (high)	Gemini 3.1 Pro	-5.35	0.1448	1.0000
MedGRPO	SUREON-VLM (Public-only)	-4.01	0.2007	1.0000
SUREON-VLM (SUREON-only)	SUREON-VLM	-4.01	0.2185	1.0000
Gemini 3.1 Pro	SUREON-VLM	4.01	0.3232	1.0000
Qwen3VL-8B-I	SUREON-VLM (Public-only)	2.34	0.4944	1.0000
GPT-5.1 (high)	SUREON-VLM (SUREON-only)	2.68	0.5363	1.0000
Molmo2	MedGRPO	-1.67	0.6305	1.0000
Qwen3VL-32B-T	Gemma 4	-1.67	0.6718	1.0000
GPT-5.1 (high)	SUREON-VLM	-1.34	0.7926	1.0000

additionally report accuracy, recall, and precision, all video-averaged. For Endoscapes Critical View of Safety (CVS), each of the three criteria (C1–C3) is posed as an independent binary (Yes/No) query (prompts in Fig. 18). We report per-criterion accuracy and the macro-average over C1–C3.

Localization metrics. Tool localization is scored by mAP₅₀, the COCO bounding-box average precision at an IoU threshold of 0.50. Because generative models emit boxes without confidence scores, all predicted boxes are assigned a constant score, so every prediction participates in the precision–recall computation. The category universe is

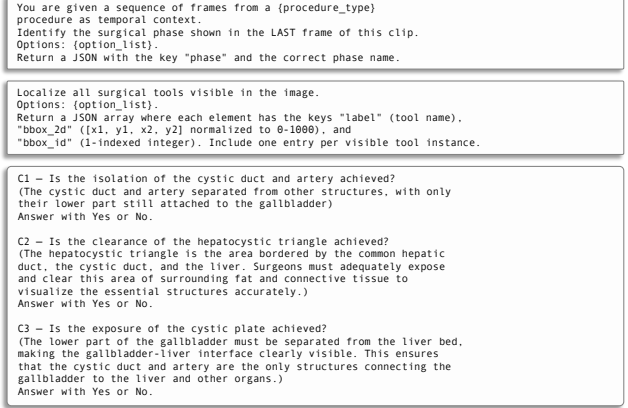


Figure 18. Prompts used for the public perception benchmarks: phase/step recognition, tool localization, and the three Critical View of Safety criteria (C1–C3).

defined by the ground-truth class names; predicted labels are canonicalized (lowercased, normalized for spacing and trivial pluralization) before matching.

Test splits and exclusions. Each benchmark is evaluated on its official test split. For Cholec80 clip evaluation we exclude 17 videos that overlap with training sources from Endoscapes and CholecT50 to prevent train/test leakage¹.

Downstream task prompts. The prompts used to query models for phase recognition, tool localization, and CVS assessment are shown in Fig. 18. Each prompt enumerates the closed label set as the option list and specifies a structured (JSON) answer format to enable checkable scoring.

E. SUREON-VLM training details

A common approach to domain adaptation is vanilla supervised fine-tuning (SFT), which we take as a representative baseline to study how data composition and staging order affect both reasoning and grounding on the Qwen3-VL-8B-Instruct backbone. Our adaptation combines two complementary resources: public datasets supplying annotated images and clips for spatial grounding and workflow recognition, and SUREON-DB supplying narration-derived VQA for reasoning over actions, states, and procedural progression. We evaluate on SUREON for visual reasoning alongside four public grounding benchmarks: CholecInstanceSeg and GraSP for tool localization, Cholec80 for phase recognition, and SARRARP-50 for action recognition.

Setup. We vary the data—**P** (public images), **S** (SUREON-DB), **P + S** (full corpus)—and exposure order. *Single* trains in one stage; $P \rightarrow \text{mix}$ and $S \rightarrow \text{mix}$ warm-start on public images or SUREON video before a second stage on the full

¹As suggested in https://github.com/CAMMA-public/camma_dataset_overlaps

Table 19. Training data and staging ablation, evaluated on SUREON and public benchmarks covering spatial grounding (tool recognition) and temporal tasks (phase and action recognition). Data sources: **P** (public images), **S** (SUREON-DB), **P+S** (full corpus of public images, public clips, and SUREON). Staging schedules: *single* trains in one stage; $P \rightarrow \text{mix} / S \rightarrow \text{mix}$ warm-start on public images / SUREON video, then train on the full mixture; $\text{mix}_1 \rightarrow \text{mix}_2$ splits the full mixture into an image-heavy stage (mix_1) followed by a balanced stage (mix_2). **C80**: Cholec80, **CISeg**: CholecInstanceSeg, **SP50**: SAR-RARP50. The backbone (Qwen3-VL-8B) is a zero-shot reference. Best in **bold**, second best underlined.

Data	Stage	SUREON (Acc.)	C80 [58] (F1)	SP50 [43] (F1)	CISeg [1] (mAP ₅₀)	GraSP [5] (mAP ₅₀)
Qwen3-VL-8B (Base model)		24.71	17.18	14.46	9.92	0.86
P	single	14.72	53.72	8.35	75.92	60.66
S	single	43.10	6.67	7.38	3.34	3.54
P+S	P $\rightarrow$ mix	43.97	74.87	23.19	67.09	<u>59.93</u>
P+S	S $\rightarrow$ mix	<u>46.02</u>	70.87	20.51	61.53	42.86
P+S	single	46.67	<u>76.90</u>	<u>24.50</u>	77.70	59.38
P+S	$\text{mix}_1 \rightarrow \text{mix}_2$	45.10	78.86	25.21	<u>77.45</u>	59.56

mixture; $\text{mix}_1 \rightarrow \text{mix}_2$ splits the mixture into two stages, image-heavy first, then balanced.

Single-source fine-tuning erodes existing capabilities.

We read the zero-shot backbone as a reference for *forgetting*: when training on source *A* drops task *B* below the untuned base, training has overwritten capability the base already had. Both single-source runs do this. Public-only training raises grounding but drops SUREON (24.7 $\rightarrow$ 14.7); SUREON-only raises SUREON but collapses tool localization on CISeg (9.9 $\rightarrow$ 3.3). Action recognition (SP50) falls below the base under *both* (14.5 $\rightarrow$ 8.4 public, $\rightarrow$ 7.4 SUREON), and SUREON also degrades phase recognition on C80 (17.2 $\rightarrow$ 6.7). Single-source supervision thus sacrifices abilities the model already possessed rather than specializing it.

Staging order determines what is retained. Exposure order matters in the mixed corpus. Public-then-video (P $\rightarrow$ mix) recovers SUREON but partly forgets the grounding learned first (CISeg 75.9 $\rightarrow$ 67.1; GraSP 60.7 $\rightarrow$ 59.9). The reverse is more revealing: adding public data after SUREON video lifts SUREON by over three points (43.1 $\rightarrow$ 46.0), yet grounding never recovers (GraSP only 42.9, $\sim$ 17 points below schedules that see images early). Grounding installed late is harder to add than reasoning layered onto a perception-first model—an asymmetry favoring early image exposure.

Strong downstream scores do not mean the model understands the scene. The combined schedules look like clear wins on the public benchmarks, but these reward fitting a fixed label set, not seeing the scene; adaptation moves capability around more than it adds understanding. Even the balanced schedules, which avoid the worst forgetting, fall

short on SUREON and stay sensitive to training order. The open problem is how to adapt a model so it reasons over the surgical scene *without* forgetting how to see it—which SUREON, requiring perception to answer, is built to track.

Final SUREON-VLM Configuration. Based on this ablation, SUREON-VLM uses the P+S ($\text{mix}_1 \rightarrow \text{mix}_2$) staging schedule, as it achieves the best average performance across SUREON and downstream perception tasks. This model fine-tunes Qwen3-VL-8B-Instruct on the consolidated corpus $\mathcal{D}_{\text{all}}$ in two stages: Stage 1 with image-heavy composition, Stage 2 with balanced data mixture, as described above.

Training implementation. All runs fully fine-tune Qwen3-VL-8B-Instruct (vision encoder, MLP projector, and LLM) for 10 epochs in bfloat16, using AdamW with a learning rate of 2×10^{-6} , a cosine schedule with 10% linear warmup, and no weight decay. We use a per-device batch size of 8 with gradient accumulation of 4 over 6 H200 GPUs (effective batch size 192) under DeepSpeed ZeRO-3, a context length of 8,192 tokens, and 8 frames per clip. For each SUREON-DB multiple-choice item, the distractor set is resampled at random from its Stage-2 candidate pool (Sec. B.3) each time the item is loaded, so option content and ordering vary across epochs as augmentation. Both stages of the best two-stage schedule ($\text{mix}_1 \rightarrow \text{mix}_2$) share these hyperparameters: Stage 1 initializes from the pretrained base with an image-heavy mixture (image 72.7% / clip 14.9% / SUREON-DB 12.4%), and Stage 2 warm-starts from the Stage-1 checkpoint with an equal-thirds mixture to re-balance the sources. Given the relative corpus sizes, the SUREON-DB video data is effectively seen $\sim 20\times$ over the schedule.

F. Ethics

Ethical use of data. SUREON is constructed from publicly available surgical lecture videos and is intended for research use only. SUREON-DB is automatically pseudo-labeled and is not expert-verified; models trained on it are research artifacts and must not be used for clinical decision-making or surgical practice in their current form. The expert-verified benchmark is intended for evaluation, not as clinical ground truth.

Data licensing and usage. SUREON is derived from SurgLaVi- β [42], which is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). In accordance with the ShareAlike provision, we release both SUREON-DB and the SUREON benchmark under the same CC BY-NC-SA 4.0 license, restricting use to non-commercial research and requiring attribution and identical redistribution terms. We release the derived QA annotations, temporal windows, and metadata; the underlying lecture videos remain governed by their original sources and are not redis-

tributed.

De-identification. The source material consists of publicly available educational lecture videos that are edited and narrated for pedagogical purposes; they do not contain patient names, medical record numbers, or other protected health identifiers. Laparoscopic and robot-assisted footage depicts internal anatomy rather than external patient features, inherently limiting re-identification risk.