SurgCoT: Advancing Spatiotemporal Reasoning in Surgical Videos through a Chain-of-Thought Benchmark

Supplementary Material

1. Dataset Curation Details

1.1. Multi-Source Surgical Video Corpus

Our benchmark is built on a heterogeneous, multi-source surgical video corpus systematically organized across seven specialties: Colorectal, Urological, Upper Gastrointestinal (Upper GI), Ocular, Gynecologic, General Surgery (GenSurgery), and Hepatobiliary Pancreatic (HPB). The raw pool comprises **8,917** videos collected from three major sources: 1) public platforms (*e.g.*, YouTube [2] and ASVIDE [1]), which provide diverse real-world operative recordings with varying styles, camera viewpoints, and surgeon expertise; 2) ten open-source surgical video repositories, including cholecystectomy and general surgery benchmarks such as Cholec80 and related extensions [8, 13, 16], cataract and ocular surgery datasets [3, 14], and multicenter laparoscopic collections; and 3) de-identified institutional archives from collaborating hospitals, which contribute higher-resolution, consistently annotated videos under IRB-approved protocols. Across these sources, the corpus spans a wide spectrum of procedures, case complexities, recording conditions, and surgeon experience, providing a realistically challenging basis for constructing fine-grained spatiotemporal reasoning tasks in SurgCoT.

1.2. Data Curation Pipeline and Inclusion Criteria

Starting from the 8,917 raw videos, we apply a standardized multi-stage filtering pipeline to obtain **2,841** high-quality cases (**31.9%** of the original corpus). The filtering is designed to enforce clinical validity and annotation feasibility, rather than aesthetic quality alone.

1) Procedural Completeness. Videos are first screened for procedural integrity. We retain only those that (i) cover the key operative phases of the target procedure (*e.g.*, trocar placement, dissection, critical step execution, reconstruction/closure) and (ii) avoid major truncation or missing core steps. Videos consisting solely of highlights, promotional material, or partial segments (*e.g.*, only anastomosis without preceding dissection) are excluded. When multiple uploads of the same case exist (*e.g.*, different edits of a single operation), we keep the most complete version.

2) Technical and Visual Quality. We then enforce basic technical quality thresholds to ensure that fine-grained spatiotemporal reasoning is feasible. Videos with extreme motion blur, persistent out-of-focus recording, severe occlusion by instruments or smoke, or dominant overlay graphics that obscure the operative field are discarded. We also exclude cases with very low resolution or frame rate that preclude reliable phase recognition, tool tracking, or temporal localization of key events.

3) Clinical Validity and Pedagogical Clarity. To ensure that each retained case reflects standard or clearly interpretable practice, we remove videos with (i) insufficiently documented procedures (*e.g.*, unclear indication, mixed or unusual techniques without explanation), (ii) grossly atypical anatomy without adequate context, or (iii) extensive editing that disrupts temporal continuity. Preference is given to didactic recordings with stable views and consistent exposure, which support precise event annotation and question design.

4) Bilingual Narration and Textual Alignment. Because our benchmark relies on temporally aligned video–language pairs for spatiotemporal QA, we require the presence of usable narration or on-screen textual guidance in either English or Chinese. Automatic speech recognition (ASR) [11] is applied to all candidate videos, followed by language detection and basic cleanup (removal of noise, music, and non-surgical chatter). Videos without any meaningful narration, or with speech too noisy or sparse to support temporal alignment, are excluded from the final benchmark.

5) De-duplication and Metadata Harmonization. Across sources, the same surgery may appear under different titles or uploads. We perform de-duplication using a combination of frame-wise similarity checks and metadata comparison (title, duration, channel). Only one canonical copy is retained per case. For all remaining videos, we standardize metadata fields (procedure name, specialty, source, language, approximate duration) to enable consistent sampling and stratified evaluation across specialties and procedures.

1.3. Ethical De-Identification and Privacy Safeguards

All videos, including those from public platforms, are processed through a de-identification pipeline prior to annotation. Faces, patient identifiers, and any textual overlays containing names, hospital IDs, or dates of birth are automatically detected and blurred. Audio tracks are screened to remove or mask personally identifiable information.

1.4. Distribution of SurgCoT in Surgical Procedures

After filtering, the final SurgCoT dataset comprises **2,841** videos spanning seven surgical specialties. Within each

Table 1. Statistics of surgical procedures in SurgCoT.

Index	English Surgical Name	Category	Videos	Total Duration (min)
1	Abdominoperineal resection	Colorectal	27	594
2	Anterior resection	Colorectal	81	1620
3	Colectomy	Colorectal	97	1746
4	Colon cancer surgery	Colorectal	99	1782
5	Colorectal disease surgery	Colorectal	35	525
6	Hemicolectomy	Colorectal	176	2992
7	Ileostomy	Colorectal	35	490
8	Rectal cancer surgery	Colorectal	148	2812
9	Rectal prolapse repair	Colorectal	112	1792
10	Rectopexy	Colorectal	93	1581
11	Sigmoidectomy	Colorectal	72	1152
12	Hernia repair	GenSurgery	70	980
13	Ladd's procedure	GenSurgery	49	784
14	Splenectomy	GenSurgery	5	95
15	Hysterectomy	Gynecology	66	1320
16	Myomectomy	Gynecology	54	972
17	Oophorectomy	Gynecology	36	576
18	Biliary bypass surgery	HPB	70	1540
19	Cholecystectomy	HPB	77	1078
20	Choledochojejunostomy	HPB	69	1449
21	Hepaticojejunostomy	HPB	77	1617
22	Liver resection	HPB	125	2875
23	Pancreatectomy	HPB	4	96
24	Robotic_chole	HPB	57	1026
25	Whipple procedure	HPB	64	1600
26	Cataract	Oculus	294	3528
27	Cornea	Oculus	77	1155
28	Glaucoma	Oculus	82	1148
29	Esophagectomy	Upper GI	4	100
30	Gastrectomy	Upper GI	142	2982
31	Gastrojejunostomy	Upper GI	122	2318
32	Heller myotomy	Upper GI	109	1962
33	Pyloroplasty	Upper GI	60	1020
34	Cystectomy	Urology	68	1360
35	Nephrectomy	Urology	82	1558
36	Prostatectomy	Urology	3	66

Colorectal: colorectal surgery; **Urological**: urological surgery; **Upper GI**: upper gastrointestinal surgery; **Ocular**: ophthalmic (ocular) surgery; **Gynecologic**: gynecologic surgery; **GenSurgery**: general surgery; **HPB**: hepatobiliary–pancreatic surgery.

specialty, we include both high-volume routine operations (*e.g.*, cholecystectomy, cataract surgery, anterior resection) and technically demanding or less frequent procedures (*e.g.*, Whipple procedure, cystectomy), so that the benchmark reflects a realistic case mix rather than a narrow set of "easy" cases. Table 1 summarizes the surgical procedures in SurgCoT, listing per-procedure video counts and total duration across Colorectal, Urological, Upper GI, Ocular, Gynecologic, General Surgery, and HPB. This broad procedural and temporal coverage provides a solid basis for fine-grained, cross-specialty spatiotemporal reasoning.

Listing 1. Example of ontology-normalized temporal anchors derived from ASR transcripts

```
# Input: ASR transcript with timestamps
{
  "segments": [
    {"start": 0.0, "end": 5.2, "text": "We'll␣start␣by␣identifying␣Calot's␣triangle"},
    {"start": 5.2, "end": 12.8, "text": "Grasping␣the␣gallbladder␣fundus␣for␣retraction"},
    {"start": 12.8, "end": 18.5, "text": "Now␣dissecting␣the␣peritoneum␣over␣the␣triangle"},
    ...
  ]
}

# Output: Ontology-normalized temporal segments
{
  "temporal_anchors": [
    {
      "segment_id": "T1",
      "time_range": [0.0, 5.2],
      "raw_caption": "We'll␣start␣by␣identifying␣Calot's␣triangle",
      "normalized_terms": ["anatomical_identification", "calot_triangle"],
      "phase": "preparation"
    },
    {
      "segment_id": "T2",
      "time_range": [5.2, 12.8],
      "raw_caption": "Grasping␣the␣gallbladder␣fundus␣for␣retraction",
      "normalized_terms": ["grasp", "retraction", "gallbladder_fundus"],
      "phase": "exposure",
      "action": "grasp",
      "tool": "grasper",
      "tissue": "gallbladder_fundus"
    },
    ...
  ]
}

# Key steps:
# - Terminology normalization: "grasping" -> "grasp", "gallbladder" -> "cholecyst"
# - Phase detection: merge semantically similar neighboring segments into one phase
# - Action extraction: detect surgical verbs (grasp, dissect, clip, divide) from ASR text
```

2. Annotation Protocols and QA Generation

2.1. Evidence Mining Pipeline

This section provides implementation details of the evidence mining pipeline that was only summarized in the main text. We describe the multi-source segmentation strategy, the ASR alignment workflow, and the ontology driven normalization rulebook that standardizes textual evidence.

Multi-source Segmentation Details. We segment long surgical videos into semantically coherent clips using a hierarchical cue fusion strategy that combines: (i) visual scene-change detection (shot and illumination changes), (ii) tool/tissue transition signals (instrument entry/exit and contact-region switches), and (iii) ASR anchors from time-stamped surgical keywords. Candidate boundaries from these three sources are fused within a small temporal tolerance window and pruned by non-maximum suppression. Segmentation quality is monitored via semantic coherence (intra- vs. inter-segment embedding similarity), segment duration histograms, and boundary precision on a small expert-annotated subset.

ASR Alignment Technical Workflow. We use Whisper-large-v3 [11] with 16 kHz audio and language specific decoding, enabling millisecond-level timestamps for each transcript segment. Raw ASR output is passed through a lightweight post processing pipeline: punctuation restoration, ophthalmic term spell checking, timestamp normalization, and confidence-based quality filtering. Alignment quality is spot checked by overlaying captions on key clips to compare aligned vs. misaligned cases (*e.g.*, shifts from background noise), and minor timing offsets are corrected at the pipeline level.

Ontology-driven Normalization Rulebook. Textual evidence is normalized to a controlled vocabulary that integrates general medical ontologies (*e.g.*, MeSH, SNOMED CT) with a surgery-specific lexicon covering procedures,

Listing 2. Example of YOLOv10 tissue detections and derived tissue state evolution.

```
# Using YOLOv10
# Input: Single frame with tissue detections
{
  "frame_id": 450,
  "tissue_detections": [
    {
      "tissue_type": "gallbladder",
      "bbox": [300, 200, 400, 350],
      "confidence": 0.92,
      "anatomical_region": "fundus",
      "visibility": "partial" # full / partial / occluded
    },
    {
      "tissue_type": "cystic_artery",
      "bbox": [520, 410, 100, 80],
      "confidence": 0.78,
      "anatomical_landmark": "calot_triangle",
      "state": "intact" # intact / clipped / divided / bleeding
    },
    ...
  ]
}

# Output: Tissue state evolution over time
{
  "tissue_tracks": [
    {
      "tissue_type": "cystic_artery",
      "timeline": [
        {"time": 45.0, "state": "intact", "bbox": [520, 410, 100, 80]},
        {"time": 46.5, "state": "clipped", "bbox": [518, 408, 95, 75]},
        {"time": 60.0, "state": "divided", "bbox": [515, 405, 90, 70]}
      ],
      "events": [
        {"time": 46.5, "event": "clip_placement", "tool_id": 3},
        {"time": 60.0, "event": "division", "tool_id": 2}
      ]
    },
    ...
  ]
}
```

anatomy, tools, and actions. Surface forms from ASR and metadata are mapped to canonical labels via dictionary lookup and simple lemmatization, with ambiguous tokens deferred to context-aware disambiguation. When conflicts arise, we prioritize official ontology terms over colloquial variants and use local context (current procedure, phase, active instruments) to select the most plausible mapping, while low-confidence cases remain explicitly flagged.

2.2. Temporal Evidence: Action Onset Detection

Appearance-Change Indicators. Temporal evidence focuses on detecting onset frames for actions and anomalies. We utilize velocity profiles derived from tracked instrument tips (ByteTrack [21] based trajectories) to identify rapid maneuvers or contact events.

Minimal Visual Cue Definition. We define the "minimal visual cue" for an onset as the first frame in which a human annotator can reliably perceive a change (*e.g.*, the first clearly visible pixel cluster of bleeding, the initial displacement of tissue, or the first frame where an instrument contacts a new region). In the annotation interface, automatic candidates from the indicators above are shown as initial anchors on the timeline; annotators then refine these by stepping frame-by-frame around the anchor to select the earliest frame that satisfies the minimal-cue criterion, rather than the frame where the event is fully developed.

ASR-based Temporal Anchor Extraction. Starting from raw ASR segments with timestamps, we normalize key terms to the surgical ontology and aggregate semantically similar windows into phase-level temporal anchors with explicit action, tool, and tissue tags, as illustrated in listing 1 shows.

Listing 3. Example JSON-like structure for anomaly onset and evolution tracking.

```
# Input: Frame sequence with anomaly detector
{
 "frames": [645, 646, 647, ...],
 "anomaly_detections": [
  {
   "anomaly_id": "AN1",
   "onset_frame": 647,
   "onset_time": 21.57,
   "anomaly_type": "bleeding",
   "initial_bbox": [540, 420, 80, 60], # First 3mm spot
   "evolution": [
    {"frame": 647, "bbox": [540, 420, 80, 60], "size": "3mm"},
    {"frame": 650, "bbox": [535, 415, 90, 70], "size": "5mm"},
    {"frame": 655, "bbox": [530, 410, 100, 80], "size": "8mm"},
    {"frame": 680, "bbox": [530, 410, 100, 80], "size": "8mm", "state": "controlled"}
   ],
   "source": {
    "tissue": "cystic_artery_stump",
    "cause": "incomplete_clip_placement",
    "location": "calot_triangle_inferior"
   },
   "hemostasis_time": 22.67 # When bleeding stops
  }
 ]
}
```

2.3. Spatial Evidence: Tool and Tissue Detection

YOLOv10 for Tissue Detection We adopt YOLOv10 [17] as the primary detector for coarse tissue and organ localization. The model is trained on a mixed dataset comprising public laparoscopic benchmarks and in-house ophthalmic frames, totaling ∼Xk annotated images with polygonal masks converted to bounding boxes. Detection classes include key anatomical targets such as the liver, gallbladder, cystic artery, and surrounding structures, depending on the procedure family. We report mAP and per-class recall on a held-out validation set to ensure reliable coverage of clinically relevant regions, and qualitatively analyze failure modes such as heavy bleeding, smoke, or specular highlights that transiently obscure tissue boundaries.

For each frame, as shown in listing 2, YOLOv10 [17] produces tissue detections with bounding boxes, confidence, and coarse state labels (*e.g.*, intact, clipped, divided). These per-frame detections are then linked over time into tissue centric tracks, so that we obtain a temporal evolution of each structure and a list of key events (*e.g.*, clip placement, division) aligned to the video timeline.

SAM2 for Tool Segmentation and Tracking. For fine-grained tool masks, we employ SAM2 [12] with lightweight prompt engineering: point prompts on tool tips for thin instruments, and box prompts for bulkier tools. Initial prompts are placed on key frames and then propagated temporally. Cross-frame association is performed with ByteTrack, which links SAM2 masks into consistent track IDs across frames. We explicitly handle occlusion and re-identification by allowing short gaps in tracks and reassigning IDs based on spatial overlap and appearance similarity after the tool re-emerges.

2.4. Anomaly Evidence.

For anomalies such as bleeding, perforation, or tool–tissue mishandling, annotators mark two key elements: (i) the onset time, defined as the exact frame where the first abnormal sign appears, and (ii) the region-of-interest (ROI) evolution, tracking the anomaly from initial appearance to full manifestation. Practically, annotators draw an initial bounding box or mask at the onset frame and then adjust it at sparse key frames; intermediate ROIs are interpolated to obtain a dense trajectory of anomaly growth over time (*e.g.*, progressive expansion of a bleeding spot). These temporally aligned ROI sequences are later used to derive precise temporal windows and spatial masks for downstream VQA and reasoning tasks. 3 shows an example JSON-like structure for anomaly onset and evolution tracking.

2.5. VQA Generation

We integrate all relevant evidence (spatial, temporal, and anomaly information) with clinical context into a unified input for the Visual Question Answering (VQA) task. This structure in listing 4 includes video metadata, spatial tracking information of detected actions, tools, and tissues,

Listing 4. Unified JSON structure for VQA input, combining spatial, temporal, and anomaly evidence.

```
{
  "video_metadata": {
    "video_id": "CHOL_001",
    "procedure": "laparoscopic_cholecystectomy",
    "duration": 180.0,
    "surgeon_experience": "expert"
  },

  "temporal_evidence": {
    "asr_segments": [...],
    "action_onsets": [...],
    "micro_transitions": [...]
  },

  "spatial_evidence": {
    "tool_tracks": [...],
    "tissue_tracks": [...],
    "affordances": [...]
  },

  "anomaly_evidence": {
    "anomalies": [...]
  },

  "clinical_context": {
    "phase_sequence": ["preparation", "exposure", "dissection", "clipping", "division"],
    "critical_structures": ["cystic_artery", "cystic_duct", "calot_triangle"],
    "standard_workflow": {
      "clip_before_divide": true,
      "identify_before_dissect": true
    }
  }
}
```

Listing 5. Video-level event detection.

```
"Q1": {
  "question": "... clinically meaningful question about action ordering ...",
  "options": {
    "A": "... first plausible ordering ...",
    "B": "... second plausible ordering ...",
    "C": "... third plausible ordering ...",
    "D": "... fourth plausible ordering ..."
  },
  "knowledge": "... clinical rationale about standard workflow and required ordering ...",
  "clue": "... video-grounded evidence indicating which actions occur, without exact timing ...",
  "answer": "A" | "B" | "C" | "D"
}
% \end{verbatim}
```

anomaly detection data, and clinical workflows.

LLM-Based Realization and Three-Stage Chains. To convert structured specifications into natural-language VQA items, we feed the bound templates and their evidence tuples into a large language model (GPT-5 [10]) using carefully designed system and user prompts. For each target event with interval $[t_s, t_e]$, we construct a three-step chain (Q1→Q2→Q3) by progressively narrowing the temporal and spatial scope: (i) Q1 uses the full clip or a broad window $[t_s - \Delta_1, t_e + \Delta_1]$ (typically 10–20 s) with no explicit numeric clue, probing coarse recognition or global context; (ii) Q2 restricts the clue to a medium window $[t_s - \Delta_2, t_e + \Delta_2]$ (4–8 s) or a coarse phase anchor (*e.g.*, "during mesorectal dissection"), requiring the

Listing 6. Temporal localization.

```
"Q2": {
  "question": "... ask for the temporal segment of the FIRST action, given Q1 ...",
  "context_from_Q1": "... explicitly restate the chosen Q1 answer ...",
  "options": {
    "A": "40-45s",
    "B": "45-50s",
    "C": "50-55s",
    "D": "55-60s"
  },
  "knowledge": "... typical clinical markers of when this action occurs ...",
  "clue": "... temporal evidence from ASR or visual changes, e.g. tool appears at 45s ...",
  "answer": "A" | "B" | "C" | "D"
}
```

Listing 7. Frame-level precision.

```
"Q3": {
  "question": "... ask for the precise frame/timestamp of the key milestone, given Q2 ...",
  "context_from_Q2": "... explicitly restate the time segment from Q2 ...",
  "options": {
    "A": "46.0s",
    "B": "46.5s",
    "C": "47.0s",
    "D": "47.5s"
  },
  "knowledge": "... fine-grained visual definition of the exact moment ...",
  "clue": "... frame-level evidence with spatial localization, e.g. 'frame 1395 at (520,410)' ..."
      ,
  "answer": "A" | "B" | "C" | "D"
}
```

model to reuse Q1 knowledge under tighter temporal guidance; and (iii) Q3 focuses on a narrow window (1–3 s) or an explicit time reference (*e.g.*, "around 645–648 s") and, when applicable, a spatial hint (*e.g.*, "near the right rectopexy stitch line"), demanding fine-grained localization or causal reasoning. Temporal hints are derived directly from event onset/offset timestamps and constrained so that the correct answer lies strictly within the specified interval. The LLM is instructed to generate concise questions and answers while strictly adhering to this structured evidence.

Answer and Distractor Construction. For each instantiated template, the semantic ground truth (phase label, action order, anomaly onset, tool tissue pair, etc.) is derived deterministically from $\mathcal{E}(c)$ and passed to the LLM as the only admissible correct answer. We then use ontology-aware prompts to ask the LLM to propose three distractor options under explicit constraints (*e.g.*, must be clinically plausible, mutually exclusive, and distinct from the ground truth). Semantically, distractors are drawn from: (i) *intra-procedure negatives*, i.e., phases, tools, or actions from the same procedure family; (ii) *intra-task negatives*, i.e., wrong timestamps within the same clip length range or alternative tool–tissue pairs from other clips of the same task type; and, when needed, (iii) *generic negatives* from other specialties for additional diversity. All LLM outputs are post-processed with lexical de-duplication, semantic constraints (*e.g.*, distractor timestamps must lie outside the ground-truth event window for temporal questions), and random shuffling of option order to avoid position bias.

Rule-Based Validation and Expert Spot-Check. Every generated QA triplet (or Q1–Q2–Q3 chain) passes through automated consistency checks before inclusion. These cover: (i) *temporal validity* (the correct option must overlap the annotated event with IoU > 0.7, while temporal distractors must not); (ii) *spatial validity* (for region questions, the referenced tool/tissue region must have sufficient overlap with tracked masks or bounding boxes); (iii) *ontology consistency* (all mentioned entities must exist in the clip's normalized label set and tool/tissue–action combinations must be clinically plausible); and (iv) *chain coherence* (Q1, Q2, Q3 of a chain must share the same event ID, with strictly increasing informational specificity). QA items failing any rule are discarded or regenerated via a new LLM call.

Finally, we perform human verification on a stratified

10% sample per task and specialty. Three board-certified surgeons independently review shared QA chains to assess clinical correctness, clarity, and difficulty; disagreements are resolved by consensus, and the resulting revision guidelines are fed back into the prompt design and post-processing rules. This hybrid pipeline, structured evidence, LLM-based realization, rule-based filtering, and expert spot-checking, ensures that SurgCoT's VQA items are both tightly grounded in the mined spatiotemporal annotations and aligned with realistic surgical reasoning patterns.

2.6. LLM Prompt Design

We design a structured prompting scheme to feed GPT-5 [10] with rich but tightly constrained evidence for VQA generation. The core principles are: (i) *structured input*, where multi-source evidence is organized into a clear JSON schema; (ii) *explicit task specification*, indicating the target VQA family (CAO, CAA, AM, MTL, AOT); (iii) *hard constraints*, enforcing a five-tuple (`question/options/knowledge/clue/answer`) and three stage (Q1–Q3) reasoning logic; and (iv) *few-shot guidance*, where small task-specific exemplars are provided to stabilize output format and style.

Stage 1 (Q1): Video-Level Event Detection. Q1 targets *what* causal relationship holds in the video at a global level. The model must produce a five-tuple as listed 5. Options must be mutually exclusive (*e.g.*, "A before B" vs. "B before A" vs. "Simultaneous"), include plausible distractors reflecting alternative techniques or workflows, and avoid trivial choices such as "neither action occurs".

Stage 2 (Q2): Temporal Localization. Q2 asks *when* the validated action sequence occurs at clip-segment granularity, and is explicitly conditioned on Q1. Temporal listing 6 distractors are designed as (i) adjacent time windows, (ii) preparatory vs. execution intervals (tool approaching vs. acting), or (iii) windows of related but distinct actions.

Stage 3 (Q3): Frame-Level Precision. Q3 refines localization to the *exact* onset frame or timestamp of the critical transition, conditioned on Q2 as shown in listing 8. Frame-level distractors are chosen within 0.5–1.0s of the true onset and include preparatory frames, completion frames, or visually similar but non-onset moments.

Global Constraints. We enforce a strict separation between *knowledge* and *clue*: the `knowledge` field must state general, reusable clinical facts (*e.g.*, "Clip placement follows vessel isolation"), while the `clue` field must contain only video-specific evidence (*e.g.*, "At 46.5s, the clip applicator jaws close at the cystic artery stump"). These two must never be mixed. Progressive conditioning is mandatory: Q2 explicitly references the selected answer from Q1 ("Given that [Q1_answer], when does…"), and Q3 explicitly references the time window from Q2 ("Within [Q2_segment], at what exact moment…"). All questions must remain clinically valid, avoiding nonsensical orderings such as "divide before identifying anatomy," and every `clue` must contain clear temporal information (timestamp or range) plus, when applicable, spatial grounding via coordinates, regions, or anatomical landmarks. This design keeps GPT-5 [10] generations well-structured, clinically meaningful, and tightly anchored to the underlying spatiotemporal evidence.

Three-Stage VQA Chain Generation We then instruct GPT-5 [10] to use the consolidated spatiotemporal evidence to generate a full three stage CAO VQA chain (Q1–Q2–Q3) that strictly follows the above five-tuple format and constraints: *Knowledge* must include only generalizable clinical facts, while *Clue* must include only video-specific spatiotemporal evidence; Q2 must be explicitly conditioned on the selected answer to Q1, and Q3 must be explicitly conditioned on the chosen time window in Q2. The model is further required to design mutually exclusive, clinically plausible distractors that genuinely test understanding of causal ordering, rather than allowing random guessing. In practice, a few similar-shot blocks are provided for each task family (CMD, AFF, MT, AOT), all using the same output format and grounded exclusively in structured evidence. This design keeps LLM generations tightly constrained, while still leveraging its strength in producing fluent, clinically realistic questions and rationales.

2.7. VQA Distribution in SurgCoT Across Surgical Procedures

Fig. 1 visualizes the distribution of SurgCoT across surgical specialties and procedure types for both main questions and sub-questions. The inner ring groups all items into seven specialties (Colorectal, HPB, Upper GI, Ocular, Urology, Gynecology, and General Surgery), while the outer ring further breaks each specialty down into concrete procedures such as cholecystectomy, rectopexy, gastrectomy, cataract surgery, and nephrectomy. This sunburst layout makes it clear that most questions are concentrated in high-volume domains such as colorectal, HPB, and upper gastrointestinal surgery, while still maintaining substantial coverage of ocular, urologic, gynecologic, and general surgery procedures.

Importantly, the two panels (MAIN QUESTION vs. SUB-QUESTION) exhibit closely matched patterns, indicating that we do not only sample diverse procedures at the case level, but also generate rich chains of follow-up questions for each operation. As a result, SurgCoT provides balanced supervision across common and complex procedures, enabling evaluation of surgical MLLMs on both broad cross-specialty coverage and fine-grained, procedure-specific spatiotemporal reasoning.

Listing 8. Prompt template for cue-action alignment (CAA)

```
## OBJECTIVE
Generate a three stage VQA chain that:
1. Q1: Does a pre-action cue exist for the target action?
2. Q2: When (which segment) does the pre-action cue appear?
3. Q3: At what exact moment does visible execution begin?
## KEY DIFFERENCES FROM CAO
### Pre-Action Cue Definition
A **pre-action cue** is a visual or behavioral indicator that precedes the onset of action
    execution:
- **Tool approaching** tissue (motion toward target)
- **Surgeon commentary** indicating intent (ASR: "Now␣I'll␣dissect...")
- **Camera zoom** focusing on target area
- **Tool positioning** without yet engaging tissue
### Execution Onset Definition
The **execution onset** is the first frame where the action produces a visible effect:
- **Dissection**: First visible tissue deformation
- **Grasping**: First jaw-tissue contact with compression
- **Clipping**: First jaw closure movement
- **Cutting**: First blade-tissue contact with visible cut
## EXAMPLE OUTPUT
'''json
{
 "task": "CAA",
 "video_id": "CHOL_001",
 "Q1": {
  "question": "Is␣there␣a␣visible␣pre-action␣cue␣before␣peritoneal␣dissection␣over␣Calot's␣
      triangle?",
  "options": {
   "A": "Yes,␣a␣pre-action␣cue␣is␣visible",
   "B": "No,␣execution␣begins␣immediately␣without␣warning",
   "C": "The␣action␣does␣not␣occur␣in␣this␣video",
   "D": "Multiple␣cues␣appear␣simultaneously"
  },
  "knowledge": "In␣laparoscopic␣surgery,␣instruments␣are␣usually␣positioned␣near␣the␣target␣
      tissue␣before␣dissection␣to␣allow␣visual␣confirmation␣and␣preparation.",
  "clue": "Temporal:␣14-16s,␣Spatial:␣the␣hook␣electrode␣moves␣toward␣the␣peritoneum␣without␣
      contact.",
  "answer": "A"
 },
 "Q2": {
  "question": "When␣does␣the␣pre-action␣cue␣identified␣in␣Q1␣appear?",
  "context_from_Q1": "A␣pre-action␣cue␣exists",
  "options": {
   "A": "12-13␣s", "B": "13-14␣s", "C": "14-15␣s", "D": "15-16␣s"
  },
  "knowledge": "Pre-action␣cues␣typically␣occur␣0.5-2␣seconds␣before␣execution,␣during␣final␣
      positioning.",
  "clue": "Temporal:␣14.0-15s,␣Spatial:␣hook␣electrode␣approaches␣the␣peritoneum",
  "answer": "C"
 },
 "Q3": {
  "question": "Within␣14-15s␣(from␣Q2),␣when␣does␣visible␣execution␣start?",
  "context_from_Q2": "Pre-action␣cue␣appears␣in␣14-15s␣segment",
  "options": {
   "A": "14.8␣s", "B": "15.0␣s", "C": "15.2␣s", "D": "15.4␣s"
  },
  "knowledge": "Peritoneal␣dissection␣execution␣is␣marked␣by␣tissue␣tenting:␣the␣peritoneum␣
      elevates␣at␣the␣moment␣of␣first␣contact␣with␣cutting␣current.",
  "clue": "Temporal:15s,␣Spatial:␣␣hook␣tip␣contacts␣peritoneum␣at␣(540,␣420)␣with␣tissue␣tenting
      .",
  "answer": "B"
 }
}
```

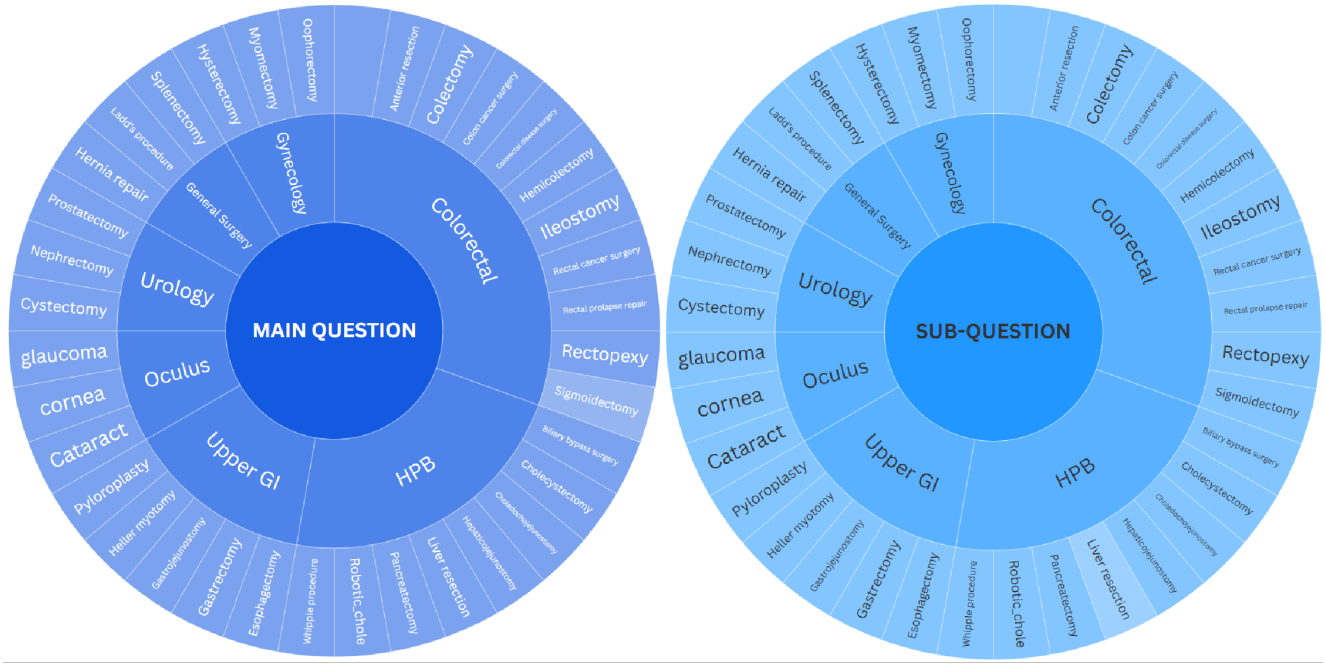


Figure 1. Distribution of SurgCoT main questions and sub-questions across surgical procedures.

2.8. Statistics of SurgCoT Across Five Critical Dimensions.

Fig. 2 summarizes the distribution of SurgCoT across seven surgical specialties and five spatiotemporal task dimensions. The upper panel reports, for each specialty, the number of curated videos and the total number of three-step question chains (Q1/Q2/Q3), together with the resulting main clinical questions. Colorectal, HPB, and Ocular surgery contribute the largest case volumes, while Upper GI, Urology, Gynecology, and General Surgery still provide hundreds of videos and thousands of questions, ensuring that every specialty is represented on a meaningful scale. The lower panel visualizes the composition of question chains across the five task families, showing that each specialty contains a rich mixture of Causal Ordering, Command/Action, Affordance, Micro-Transition and Anomaly Tracking items. Taken together, these statistics demonstrate that SurgCoT achieves both broad procedural coverage and a balanced distribution of fine-grained reasoning tasks across specialties.

3. Quality Control and Expert Validation

To ensure that SurgCoT provides reliable supervision for surgical reasoning, we adopt a multi-layer quality control pipeline that combines rule-based validation with expert review.

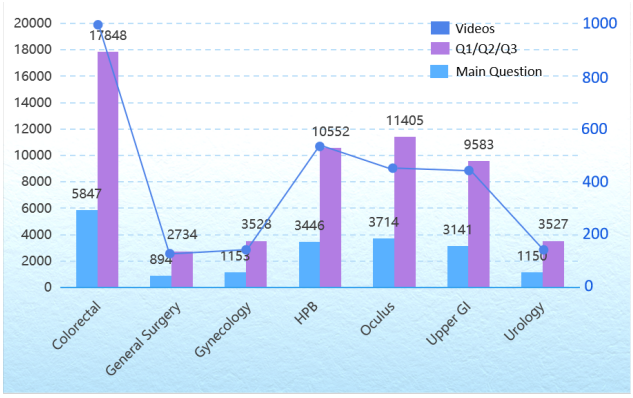
Automated Consistency Checks. After QA generation, every item (and each Q1→Q2→Q3 chain) is screened by a battery of automatic checks. We verify (i) *temporal consistency*, enforcing sufficient overlap between the answer timestamp and the annotated event while ensuring distractor timestamps lie outside the target window; (ii) *spatial validity*, requiring that referenced tools and tissues overlap corresponding masks or boxes across frames; (iii) *ontology consistency*, checking that all entities and tool–tissue–action combinations exist in the normalized label set and are clinically plausible; and (iv) *chain coherence*, ensuring that all steps in a chain refer to the same underlying event and that clues become strictly more informative from Q1 to Q3. Items failing any rule are automatically discarded.

Expert Spot-Checking and Adjudication. We further conduct human verification on a stratified 10% sample per task family and specialty. Three board-certified surgeons independently assess whether each question, answer, and distractor set is clinically correct, unambiguous, and appropriately challenging. Disagreements are resolved by consensus, and the resulting revision rules (*e.g.*, tightening temporal windows, refining phrasing, removing borderline clips) are fed back into the generation pipeline for another iteration. Across this sample, automated checks required no expert correction for 94.3% of items, and inter-rater agreement measured by Cohen's κ reached 0.87 for disease/phase identification, 0.82 for severity and action-related grading, and 0.79 for clinical decision items; 98.1% of questions were judged to fully address all components in the prompt.

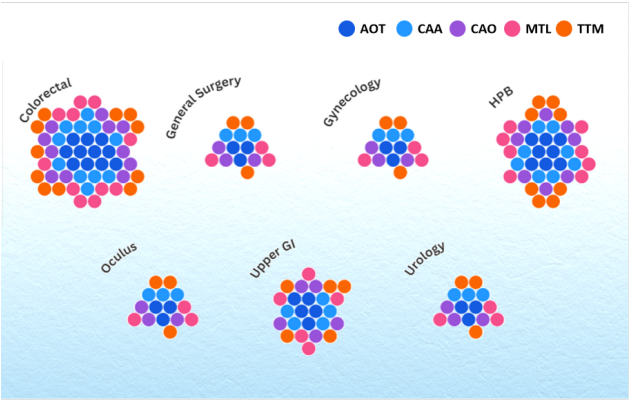
Expert Spot-Checking and Adjudication. We further conduct human verification on a stratified 10% sample per

Table 2. Chain completion accuracy (%) of 10 MLLMs across five clinical reasoning tasks under three information settings (**BL**, **KE**, **FC**). Values denote the percentage of cases where *all sub-questions and the main question* are answered correctly. Best results in **bold**, second-best underlined.

	Model	CAO. (%) ↑			CAA. (%) ↑			AM. (%) ↑			MTL. (%) ↑			AOT. (%) ↑			Avg. (%) ↑		
		BL	KE	FC	BL	KE	FC	BL	KE	FC	BL	KE	FC	BL	KE	FC	BL	KE	FC
Com.	GPT-5 [10]	14.92	17.46	**20.00**	**15.82**	14.58	17.25	**14.19**	11.68	14.69	12.77	13.16	15.69	11.52	11.01	13.09	13.84	13.58	16.14
	Gemini-2.5-Pro [7]	17.35	**20.00**	**20.00**	14.40	12.15	15.24	12.35	11.71	15.86	12.55	14.86	15.71	10.02	**15.06**	13.58	13.34	14.76	16.08
	claude-sonnect-4.5 [4]	18.27	**20.00**	**20.00**	15.69	15.62	16.31	11.35	11.29	**17.08**	13.62	12.08	15.35	7.58	12.36	14.39	13.30	14.27	**16.63**
Medical	MedGemma-27B-IT [15]	**19.01**	15.00	**20.00**	11.14	11.39	14.68	13.77	5.96	9.38	**13.98**	6.27	16.96	8.75	8.75	13.15	13.33	9.47	14.83
	Lingshu-7B [20]	16.50	18.62	16.25	9.16	8.57	7.48	10.87	5.39	5.76	10.80	11.31	5.44	10.36	12.75	7.62	11.54	11.33	8.51
	LLaVA-Med-7B [9]	17.10	**20.00**	**20.00**	13.60	11.28	12.08	10.72	14.09	13.23	12.72	13.74	16.27	10.02	10.00	11.62	12.83	13.82	14.64
	HuatuoGPT-Vision-7B [5]	14.72	9.33	19.99	12.09	**18.74**	6.61	12.73	5.70	10.16	10.84	7.85	**17.69**	10.00	10.84	7.62	12.08	10.49	12.42
Open.	InternVL3.5-8B [6]	18.51	19.49	19.69	15.12	7.42	13.38	6.15	7.77	12.79	10.89	10.47	12.30	10.38	9.57	12.23	12.21	10.95	14.08
	Qwen2.5-VL-7B [18]	14.09	13.42	**20.00**	13.27	14.98	13.72	10.12	10.14	10.38	13.30	**16.11**	12.62	9.38	10.16	11.26	12.03	12.96	13.60
	Qwen3-VL-8B [19]	18.93	19.84	19.92	13.41	13.26	**18.62**	10.79	**16.60**	6.05	13.72	13.65	11.04	**15.03**	12.99	**17.36**	**14.38**	**15.27**	14.60



(a) Numbers of videos/mian question/sub-question Across Surgical Categories



(b) Density Map of Five AOT Task Types Across Surgical Categories

Figure 2. Statistics of SurgCoT include 2,841 videos, 19,345 main questions, and 59,177 sub-questions across five critical dimensions.

task family and specialty. Three board-certified surgeons independently assess whether each question, answer, and distractor set is clinically correct, unambiguous, and appropriately challenging. Disagreements are resolved by consensus, and the resulting revision rules (*e.g.*, tightening temporal windows, refining phrasing, removing borderline clips) are fed back into the generation pipeline for another iteration. Across this sample, automated checks required no expert correction for 94.3% of items, and inter-rater agreement measured by Cohen's κ reached 0.87 for disease/phase identification, 0.82 for severity and action-related grading, and 0.79 for clinical decision items; 98.1% of questions were judged to fully address all components in the prompt.

4. Supplementary Results: Chain Completion Accuracy Across Settings

Performance Trends. Due to space constraints, the full chain completion accuracy of 10 MLLMs across five tasks is presented in Table 2 instead of the main manuscript. Across all tasks and models, a consistent performance improvement is observed as the experimental setting advances from the baseline (**BL**) to the knowledge-enhanced (**KE**) and full-context (**FC**) configurations.

- Under the **BL** setting, where models only receive the surgical video and the main question, chain completion rates remain modest (typically 10%–15%), highlighting the inherent difficulty of end-to-end multi-step reasoning without structured guidance.
- Introducing clinical knowledge (**KE**) yields noticeable gains, particularly in tasks such as causal action ordering and cue–action alignment, indicating that explicit domain priors help models stabilize reasoning logic beyond pure visual pattern matching.
- The **FC** setting, which incorporates both knowledge and spatio-temporal clues, leads to the highest chain completion accuracy across nearly all models. Commercial systems (*e.g.*, GPT-5 [10], Gemini-2.5-Pro [7], Claude-Sonnet-4.5 [4]) exhibit the most substantial gains, with average completion rates rising from 15% under **BL** to over 16% under **FC**, demonstrating effective utilization of contextual scaffolding.

Impact of `Knowledge` and `Clue` on Complex Reasoning Tasks. The contribution of `Knowledge` and `Clue` is most critical for tasks with high temporal and causal de-

mands, such as Micro-transition Localization and Anomaly Onset Tracking. Under the baseline (**BL**) video-only setting, models frequently exhibit uncertainty in identifying fine-grained phase boundaries or anomaly initiation points, resulting in breaks in the reasoning chain. In contrast, the full-context (**FC**) setting, which provides explicit anatomical knowledge and spatio-temporal clues, enables models to anchor their reasoning to clinically meaningful cues, leading to significant improvements in chain completion accuracy. These results indicate that while tasks with inherent structural regularity, such as Causal Action Ordering , can be partially addressed from visual sequences alone, temporally complex tasks like MTL and AOT strongly depend on external scaffolding to resolve ambiguity and ensure reasoning reliability.

Med-Specialized vs. General Models. Although medically fine-tuned models (*e.g.*, MedGemma-27B-IT [15], Lingshu-7B [20]) show measurable gains under enhanced settings, their improvements are generally less pronounced and stable compared to leading commercial and open-source MLLMs. This indicates that current medical domain adaptation methods are insufficient for embedding robust, multi-step clinical reasoning capabilities. These findings underscore the importance of integrating explicit `Knowledge` and contextual `Clues` to bridge the gap toward clinically reliable chain completion.

References

[1] Asvide: Ame surgical video database. https://www.asvide.com/. Online ISSN 2412-270X. Accessed: 2025-11-18. 1

[2] Youtube. https://www.youtube.com. Accessed: 2025-11-18. 1

[3] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Béatrice Cochener, and Gwenolé Quellec. Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical Image Analysis*, 52:24–41, 2019. 1

[4] Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025. Accessed: 2025-11-09. 11

[5] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 11

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23904–23915, 2024. 11

[7] Google DeepMind. Gemini 1.5: Scaling up multimodal reasoning, 2024. https://deepmind.google/technologies/gemini/. 11

[8] Wen-Yen Hong, Chia-Lung Kao, Yi-Hung Kuo, Jui-Ru Wang, Wen-Lin Chang, and Chi-Sheng Shih. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80, 2020. 1

[9] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 11

[10] OpenAI. Gpt-5: Multimodal large language models for understanding complex visual inputs. *OpenAI Technical Report*, 2025. 6, 8, 11

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2023. 1, 3

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5

[13] Dominik Rivoir, Sebastian Bodenstedt, Felix von Bechtolsheim, Marius Distler, Jürgen Weitz, and Stefanie Speidel. Unsupervised temporal video segmentation as an auxiliary task for predicting the remaining surgery duration. *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging (OR/MLCN@MICCAI)*, 2019. Workshop paper on surgical workflow analysis using Cholec80. 1

[14] Klaus Schoeffmann, Manfred J. Primus, Stefan Petscharnig, Mario Taschwer, and Stefan Sarny. Cataract-101: Video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM International Conference on Multimedia Systems (MMSys)*, pages 421–425, 2018. 1

[15] 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. 11, 12

[16] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36 (1):86–97, 2017. 1

[17] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 5

[18] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 11

[19] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 11

[20] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. 11, 12

[21] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 4