GATE: CAN VISION-LANGUAGE MODELS GROUND, AUDIT, AND TRACK SURGICAL EVENTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) could one day assist surgeons, but only if they can reliably interpret visual evidence, verify their own interpretations, and follow procedures as they unfold. How well current VLMs support these abilities, and how training can improve them, remain open questions. We introduce GATE, an evaluation framework that integrates a surgical video benchmark with systematic training interventions to assess whether models can **G**round, **A**udit, and **T**rack surgical **E**vents. We therefore establish a capability taxonomy of six categories and 15 tasks, supported by a benchmark of 13,715 human-reviewed questions over 36,264 video clips from eight datasets across six anatomical sites. We evaluate 11 leading proprietary and open-weight models, including the recently released GPT-6 Astra, which is the strongest model in our evaluation yet achieves only 49.0% overall accuracy. On average, models are stronger at recognition and spatial judgments than at workflow reasoning and joint description and workflow correction. This uneven profile makes procedural reasoning and consistent verification central priorities for advancing surgical intelligence. Training interventions offer a concrete path forward: across models adapted separately for each task from a backbone with two billion parameters, supervised fine-tuning followed by reinforcement learning raises pooled test accuracy from 2.3% to 38.9%, compared with 35.7% for supervised fine-tuning and 12.3% for direct reinforcement learning. Supervised fine-tuning accounts for most of the overall accuracy gain, with further improvement from subsequent reinforcement learning. GATE thus provides an assessment framework that quantifies current capability gaps, evaluates how training interventions address these gaps, and guides future research and model development in surgical intelligence.

1 INTRODUCTION

Surgical video processing offers a route toward surgical intelligence by enabling systems to interpret what is happening, relate events over time, and check their understanding against visual evidence. The field has expanded from phase and instrument recognition (Twinanda et al., 2017) to visual question answering with vision-language models (VLMs) (Seenivasan et al., 2023). In future systems that monitor surgery or inform operative decisions, misinterpreting visual evidence could lead to missed hazards or inappropriate guidance, with potentially irreversible harm to patients. Without rigorous assessment, deployment across clinical settings could turn undetected model failures into recurrent risks to patient safety. Establishing what models can reliably understand is therefore fundamental to the responsible development of surgical intelligence.

Existing evaluations address temporal relations (Wang et al., 2026), procedural reasoning (Li et al., 2026), and visual verification (You et al., 2026). However, performance on individual tasks provides an incomplete picture of whether visual understanding supports consistent judgments and error correction across surgical events. To address this gap, we introduce **GATE** (Figure 1), an evaluation framework integrating a surgical video benchmark with systematic training interventions to assess whether VLMs can **G**round, **A**udit, and **T**rack surgical **E**vents. The benchmark covers six capability categories: *Phase Understanding*, *Scene Understanding*, *Temporal Reasoning*, *Workflow Reasoning*, *Description Correction*, and *Description and Workflow Correction*. GATE quantifies capability gaps across these categories and evaluates how training interventions address them, providing evidence to guide future model development.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

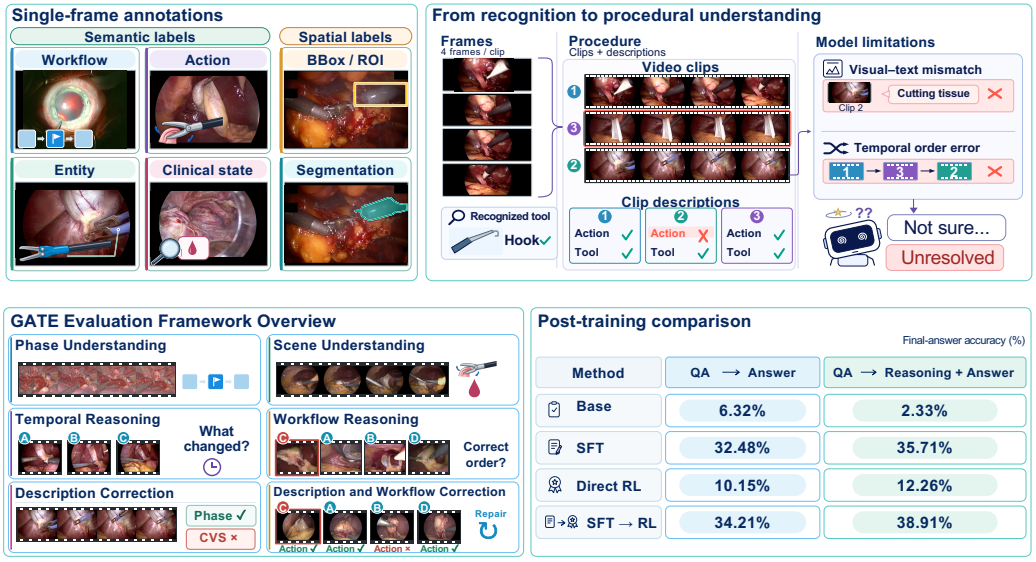


Figure 1: **Overview of the GATE evaluation framework.** GATE integrates a surgical video benchmark with model evaluation and training interventions.

The GATE framework provides the benchmark foundation for this assessment: 13,715 human-reviewed questions across 15 tasks, based on 36,264 video clips in eight datasets spanning gallbladder, eye, prostate, gynecologic, lung, and liver surgery. We construct questions and reference answers from existing surgical annotations and use human review to check that the questions are answerable from the visual evidence.

Our evaluation reveals substantial capability gaps even among leading VLMs. We assess 11 proprietary and open-weight models, including the recently released GPT-6 Astra (OpenAI, 2026b), which leads our evaluation yet achieves only 49.0% overall accuracy. The category-level results show why surgical intelligence requires progress across capabilities: improvements in recognition alone cannot establish reliable workflow reasoning or consistent verification of surgical interpretations.

Training interventions yield substantial gains on GATE. Using Qwen3-VL-2B-Instruct, we apply supervised fine-tuning (SFT), direct reinforcement learning (RL), and SFT followed by RL on the same fixed per-task partitions. Across models trained separately for each task, pooled accuracy on 2,660 test questions rises from 2.3% to 35.7% with SFT and 38.9% with SFT followed by RL. Direct RL reaches 12.3%. These results support training interventions as a practical route to improving surgical event understanding, although the additional benefit of RL remains uneven across tasks.

Our contributions are **(1)** an assessment framework integrating a surgical video benchmark with systematic training interventions, **(2)** a human-reviewed benchmark spanning 15 tasks, six capability categories, and eight datasets, **(3)** an evaluation of 11 leading models that quantifies current capability gaps, and **(4)** evidence of measurable improvement through training interventions using SFT and RL. For surgeons, GATE clarifies the strengths and limitations of current models in interpreting surgical events, informing expectations and evaluation requirements for future clinical applications. For computer science researchers and artificial intelligence (AI) model developers, it provides measurable capability targets and a common basis for evaluating training interventions, guiding progress toward more reliable surgical intelligence.

2 GATE Benchmark

2.1 Task Categories

We establish GATE into six capability categories and 15 tasks, illustrated in Figure 2. A *category* groups related capabilities, while a *task* defines a specific judgment, its visual inputs, and the

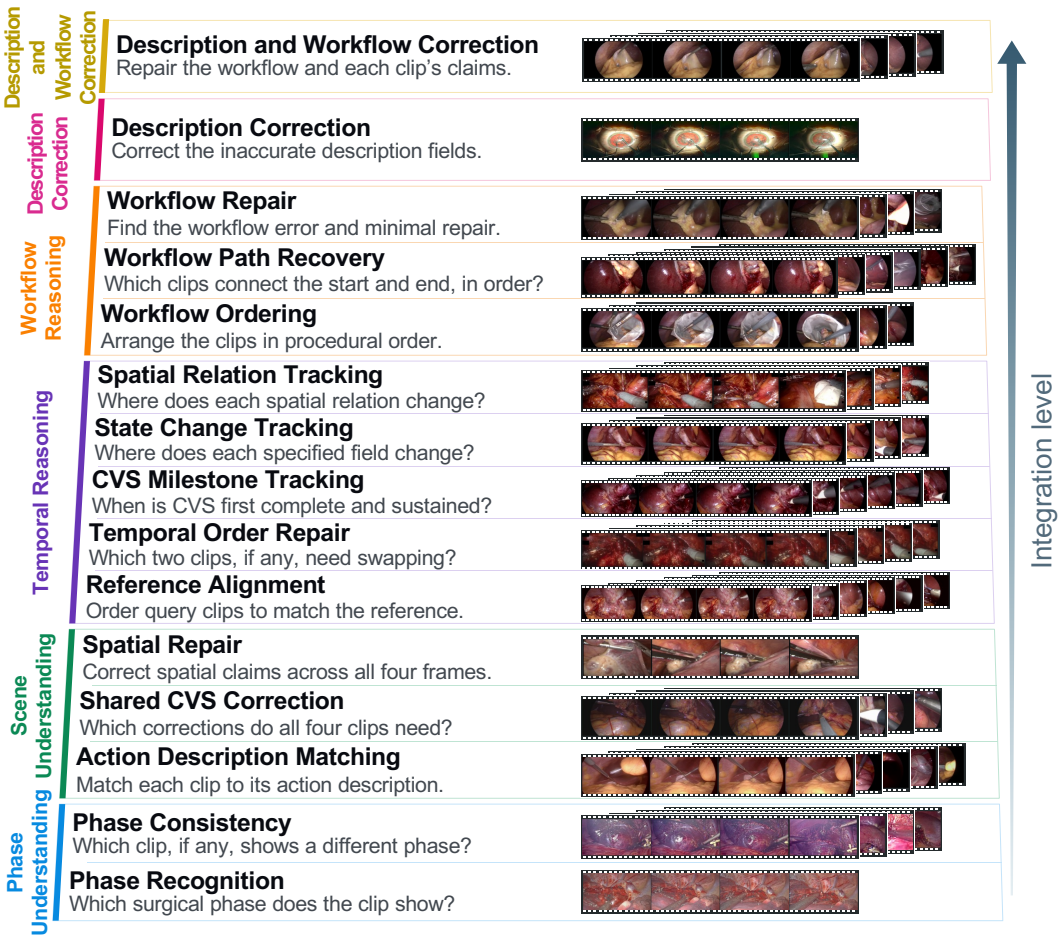


Figure 2: **GATE task overview.** Fifteen tasks span six capability categories, with example clips and simplified questions. CVS denotes Critical View of Safety.

required answer. The categories cover phase understanding, scene understanding, temporal reasoning, workflow reasoning, description correction, and joint description and workflow correction. Appendix A provides the full task specifications and dataset coverage.

Phase Understanding. This category assesses the identification and comparison of surgical phases. *Phase Recognition* asks which phase a clip depicts. *Phase Consistency* asks whether four clips share the same phase and, if not, identifies the single outlier. Together, they distinguish recognizing a phase from judging consistency across clips.

Scene Understanding. This category assesses relations within and across clips without reconstructing a workflow. *Action Description Matching* associates each clip with a description of its observed action. *Shared CVS Correction* selects the corrective operations required by all four clips from different surgeries, based on the Critical View of Safety (CVS) criteria. *Spatial Repair* identifies incorrect spatial claims and supplies the corrected relations while leaving valid claims unchanged.

Temporal Reasoning. This category assesses correspondences and changes across surgical clips. *Reference Alignment* matches action or step patterns between reference and query clips from different cases. *Temporal Order Repair* restores chronology within a phase by identifying a swapped pair, or confirms that no swap is needed. *CVS Milestone Tracking* locates the first clip with complete CVS and the later clip in the first consecutive pair of clips with complete CVS. *State Change Track-*

Table 1: VLM performance on GATE (accuracy, %). Darker and lighter purple indicate the highest and second-highest model accuracies for each task, respectively.

Model	Phase Understanding		Scene Understanding			Temporal Reasoning					Workflow Reasoning			Description Correction	Description and Workflow Correction
	Phase Recognition	Phase Consistency	Action Matching	CVS Correction	Spatial Repair	Reference Alignment	Temporal Repair	CVS Tracking	State Tracking	Spatial Relation	Workflow Ordering	Workflow Recovery	Workflow Repair	Description Correction	Description and Workflow Correction
Proprietary Models															
GPT-6 Astra	63.26	62.36	24.13	34.86	88.52	56.99	68.82	72.84	9.96	62.17	45.86	36.71	47.87	29.36	33.96
GPT-6 Luna	34.84	36.23	2.05	10.81	63.68	33.23	24.83	18.07	5.80	49.70	24.84	4.51	12.37	23.35	11.43
GPT-5.6 Sol	51.37	50.00	9.89	26.81	89.66	44.58	42.29	34.63	6.56	48.69	33.01	13.90	26.44	23.25	17.76
GPT-5.6 Luna	40.00	40.78	5.43	11.72	75.03	36.17	25.17	8.33	5.14	45.17	25.27	3.66	12.69	22.14	21.98
Claude Opus 5.5	60.74	54.23	20.27	20.29	79.19	45.85	48.87	50.43	11.71	67.10	38.13	29.39	38.81	37.27	45.17
Claude Sonnet 5	34.63	34.82	2.90	18.65	30.26	26.81	14.17	27.71	4.49	24.25	20.04	3.05	11.41	23.75	7.55
Open-Weight models															
Kimi K3	40.00	42.73	8.93	11.21	86.63	38.70	28.80	11.04	4.27	40.74	27.12	7.93	16.10	23.25	27.30
GLM-5.3-Flash	47.16	48.48	12.06	18.65	59.77	30.28	17.69	24.24	6.56	38.63	27.45	9.02	18.12	20.84	16.87
InternVL3.5	33.79	34.82	3.86	31.91	24.46	26.29	10.66	3.46	4.92	8.85	17.65	2.44	8.64	13.03	18.31
Nemotron-3 Omni	30.74	24.95	0.84	2.04	10.59	19.56	5.90	3.35	0.98	3.72	17.10	1.83	4.16	0.50	10.10
Gemma 4 31B IT	45.05	45.01	6.76	11.21	61.92	33.75	15.53	18.94	6.56	41.65	24.40	7.93	13.97	24.75	14.98
Model average	43.78	43.13	8.83	18.01	60.88	35.66	27.52	24.82	6.09	39.15	27.35	10.94	19.14	21.95	20.49
Random	23.89	19.41	0.36	12.95	10.72	14.62	9.07	4.33	0.55	5.73	18.95	1.46	3.73	0.50	1.44

ing identifies transitions in the specified phase, instrument, action, target, or step attributes. *Spatial Relation Tracking* identifies transitions in specified relations between objects.

Workflow Reasoning. This category assesses the organization of surgical phases under the workflow rules specified for each question. *Workflow Ordering* arranges clips in procedural order. *Workflow Path Recovery* selects and orders the candidate clips that belong between a given start and end. *Workflow Repair* corrects a displayed sequence through insertion, swapping, deletion, or replacement. These tasks require reasoning about how phases fit into a procedure beyond recognizing each phase separately.

Description Correction. This category assesses whether multiple attributes of a clip form an accurate description. *Description Correction* checks the supplied procedure, phase, instrument, action, target, anatomy, or CVS attributes, as supported by the source dataset. The model returns the complete corrected description, preserving attributes that were already correct.

Description and Workflow Correction. This category brings description correction and workflow reasoning into one assessment. *Description and Workflow Correction* presents four clips with corresponding claims and asks the model to check both the local descriptions and the displayed workflow. The answer includes only the components requiring correction. The model indicates that no correction is needed when all claims and the workflow are correct.

2.2 Data Collection and Statistics

Data collection. We use eight datasets: Cholec80 (Twinanda et al., 2017), CholecT50 (Nwoye et al., 2022), Cataract1K (Ghamsarian et al., 2024), GraSP (Ayobi et al., 2025), LungRes80 (Guo et al., 2026a), PmLR50 (Guo et al., 2025), Endoscapes2023 (Mascagni et al., 2025), and GynSurg (Nasirihaghighi et al., 2025), spanning gallbladder, eye, prostate, lung, liver, and gynecologic surgery. We additionally use the Cholec80 CVS annotations (Ríos et al., 2023). Available annotations provide phase, step, instrument, action, target, spatial, and CVS information. We select clips satisfying each task's annotation requirements and construct questions and reference answers from

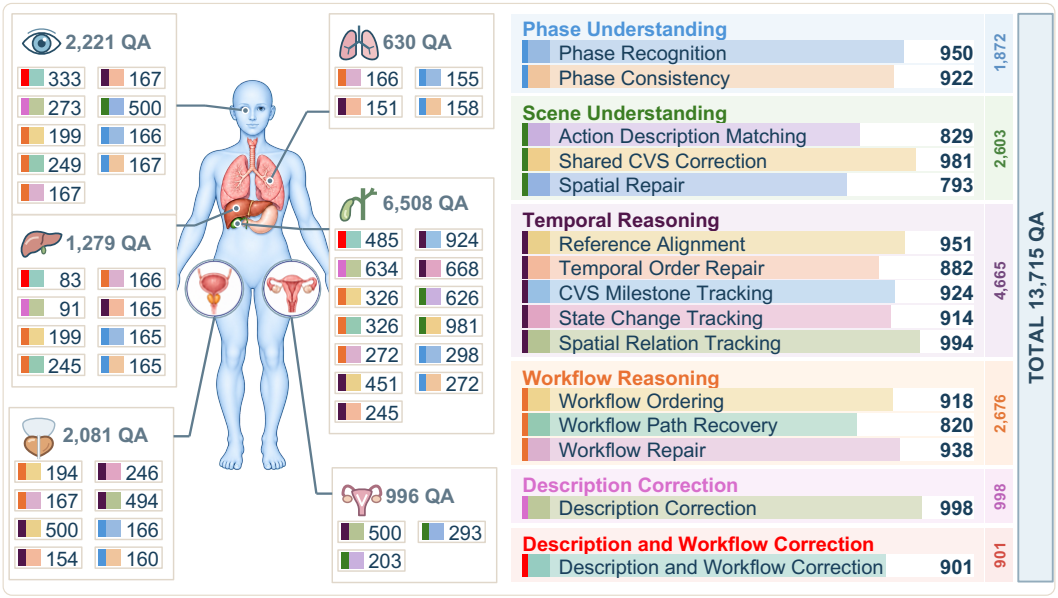


Figure 3: **GATE data statistics.** The benchmark contains 13,715 questions spanning six anatomical sites, six capability categories, and 15 tasks. The left panel shows question counts by anatomical site. The right panel shows the question distribution across tasks and capability categories.

these labels. Where required, we introduce distractors, reorder clips, or alter description attributes to create the judgments and corrections being assessed. Appendix A.1 details the source annotations.

Quality control. Automatic validation checks that the selected clips, question, and reference answer agree with the source annotations and task rules, and that the required images are available. Human reviewers then inspect each candidate's visual evidence through a dedicated interface and record a pass or fail decision. Only candidates passing both stages enter the benchmark. Human screening removes 1,048 of the 14,763 candidates. All retained questions have a recorded pass decision. Appendix A.3 describes the validation process and review interface.

Statistical Analyses. Figure 3 summarizes the distribution across anatomical sites, categories, and tasks. The final benchmark contains 13,715 questions based on 36,264 video clips across the eight source datasets. Each of the 15 tasks contributes between 793 and 998 questions, with 56 populated task–dataset combinations. The complete counts by task and source dataset appear in Table 5.

3 Experiments

3.1 Experimental Setup

Models. We evaluate 11 VLMs. The six proprietary models are GPT-6 Astra (OpenAI, 2026b), GPT-6 Luna (OpenAI, 2026c), GPT-5.6 Sol and GPT-5.6 Luna (OpenAI, 2026a), Claude Opus 5.5 (Anthropic, 2026a), and Claude Sonnet 5 (Anthropic, 2026b). The five open-weight models are Kimi K3 (Moonshot AI, 2026), GLM-5.3-Flash (Z.ai, 2026), InternVL3.5 (Wang et al., 2025b), Nemotron-3 Omni (NVIDIA et al., 2026), and Gemma 4 31B IT (Google DeepMind, 2026). Full model identifiers appear in Appendix B.1.

Protocol. Each clip is represented by four frames resized to a short side of 256 pixels and arranged in a 1×4 image. We report accuracy on all 13,715 questions. Training comparisons use 2,660 held-out test questions. Input preparation and evaluation details appear in Appendix B.

Table 2: Qwen3-VL-2B-Instruct performance on GATE after per-task training (accuracy, %). Darker and lighter purple indicate the highest and second-highest positive accuracies for each task, respectively. Ties share a color.

Model	Phase Understanding		Scene Understanding			Temporal Reasoning					Workflow Reasoning			Description Correction	Description and Workflow Correction
	Phase Recognition	Phase Consistency	Action Matching	CVS Correction	Spatial Repair	Reference Alignment	Temporal Repair	CVS Tracking	State Tracking	Spatial Relation	Workflow Ordering	Workflow Recovery	Workflow Repair	Description Correction	Description and Workflow Correction
Base model	11.58	1.09	0.00	0.00	0.63	9.47	0.00	0.00	1.09	0.50	8.15	0.00	0.00	0.50	0.00
SFT	34.21	36.96	14.86	18.46	83.02	21.05	19.32	64.86	8.74	29.15	38.59	44.51	18.09	61.50	38.89
RL	28.95	17.39	0.00	12.31	21.38	17.89	6.82	0.00	0.55	22.11	17.39	0.00	2.66	2.00	31.67
SFT + RL	34.74	44.57	14.19	41.54	77.36	12.63	23.30	64.86	13.11	44.22	38.04	48.17	20.21	62.50	44.44

3.2 Benchmark Results

Table 1 reports results across the 15 tasks. Appendix D.1 provides category-level and overall scores. We highlight three findings.

Even the strongest model leaves more than half of GATE unsolved. GPT-6 Astra leads with 49.00% overall accuracy, followed by Claude Opus 5.5 at 43.11%. The strongest evaluated open-weight model, Kimi K3, reaches 27.39%, leaving a 21.61-percentage-point gap to GPT-6 Astra. Thus, neither model tier provides consistently correct answers across the assessed surgical capabilities.

Recognition is more reliable than workflow reasoning and joint auditing. Every evaluated model scores higher on Phase Understanding than on either Workflow Reasoning or Description and Workflow Correction. For GPT-6 Astra, these category scores are 62.82%, 43.76%, and 33.96%, respectively. Recognizing surgical phases therefore does not establish comparable reliability in relating procedural steps or checking an entire episode.

Overall leadership does not imply the strongest verification ability. GPT-6 Astra leads 10 of 15 tasks, but Claude Opus 5.5 leads four, including Description Correction and Description and Workflow Correction. On these two tasks, Claude Opus 5.5 scores 37.27% and 45.17%, compared with GPT-6 Astra's 29.36% and 33.96%. GPT-5.6 Sol leads the remaining task, Spatial Repair, at 89.66%. The category and task breakdowns therefore expose strengths that an overall ranking alone would miss.

3.3 Training

Methodology. We test whether training interventions improve the surgical capabilities assessed by GATE. Starting from Qwen3-VL-2B-Instruct (Bai et al., 2025), we compare the base model with supervised fine-tuning (SFT) using low-rank adaptation (LoRA) (Hu et al., 2021), reinforcement learning (RL) using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), and SFT followed by RL. Each method trains a separate model for each of the 15 tasks, using the same fixed data splits. SFT learns reference explanations and final answers. RL rewards answer correctness and output-format compliance. Full training and evaluation details appear in Appendix C.

Supervised initialization accounts for most of the pooled improvement. Across the per-task models in Table 2, pooled test accuracy rises from 2.33% to 35.71% with SFT, compared with 12.26% for direct RL. SFT + RL reaches 38.91%, adding 3.20 percentage points over SFT. These results favor supervised initialization under the tested training budgets and show that subsequent reward-based optimization can further improve the overall score. The pooled values combine 15 separately trained models, rather than one model trained on all tasks.

The additional RL stage helps selectively rather than uniformly. Relative to SFT, SFT + RL improves 10 tasks, leaves one unchanged, and reduces accuracy on four. Shared CVS Correction rises from 18.46% to 41.54%, whereas Spatial Repair falls from 83.02% to 77.36% and Reference Alignment from 21.05% to 12.63%. Training interventions can therefore address substantial capability gaps, but their benefits must be checked at the task level. Learning curves and output-length

analyses are provided in Appendix C.7. Appendix E.2 analyzes answer changes before and after training.

3.4 Error and Reasoning Analysis

We analyze 150,865 responses from 11 models on 13,715 image-based questions, grouping them into correct, scorable but wrong, nonempty unscorable, and empty outputs (Figure 4). We examine visible reasoning and answer choices to identify recurring disagreements with the reference. Appendix E.1 provides the detailed methods and model-level results.

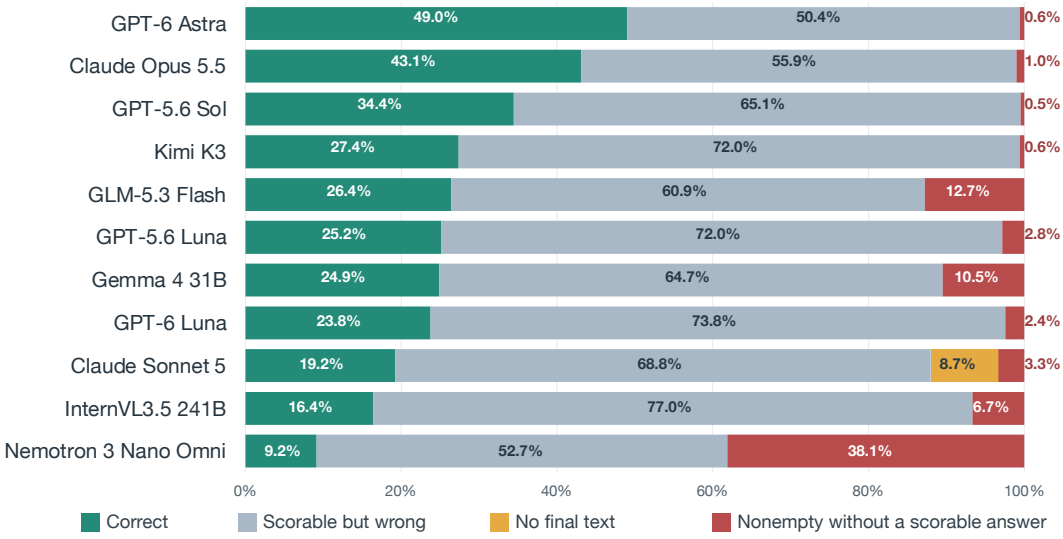


Figure 4: **Answer correctness and output status across models.** Each bar shows the proportions of a model's 13,715 responses classified as correct, scorable but wrong, no final text, or nonempty without a scorable answer. The four categories are mutually exclusive and sum to 100%.

Stronger benchmark performance tends to accompany more reliable answer generation. The three highest-accuracy models, GPT-6 Astra, Claude Opus 5.5, and GPT-5.6 Sol, yield task-valid answers on 99.44%, 99.00%, and 99.53% of questions, respectively, compared with 89.54% for Gemma 4 31B IT and 61.87% for Nemotron-3 Omni (Table 20). These rates measure whether an answer meets the task's structural requirements, regardless of whether its content matches the reference. Following task instructions and delivering usable answers are therefore part of the model capability gap, alongside the correctness of surgical judgments.

Models favor correction over preserving valid interpretations. Across six tasks with explicit no-change answers, model-mean rates of proposing a change when none is required range from 53.4% to 92.0%, whereas rates of missing a required change range from 1.0% to 23.0%. Models therefore need to improve their ability to determine whether a surgical interpretation requires correction.

3.5 Probing Analysis

Text-only evaluation: visual access alone does not ensure reliable surgical judgments. On the same 1,370-question subset sampled at approximately 10% within each task, mean overall accuracy across all 11 models is 27.29% with images versus 14.35% without them. Averaged across models, accuracy is higher with images in five of the six capability categories and 12 of the 15 tasks (Appendix D.2). Images thus help most capability groups, but their benefits remain task dependent.

Answer-only training: the reasoning advantage persists across training methods. Keeping the questions and training settings fixed, we remove the reasoning requirement from the prompt and supervised completion. Reasoning-mode SFT, RL, and SFT + RL reach 35.71%, 12.26%, and 38.91%, respectively, compared with 32.48%, 10.15%, and 34.21% in answer-only mode. The largest difference, 4.70 percentage points, occurs after SFT + RL (Appendix C.6). The consistent advantage

favors retaining explicit reasoning in training interventions. Appendix E.3 examines answer-only errors and repeated predictions.

Cross-task transfer: task-specific improvements do not readily generalize to other tasks. We evaluate 45 trained models on the other 14 tasks, giving 630 source–target comparisons. Across the 15 SFT + RL source models, accuracy pooled over their other-task test questions improves by 0.32–6.59 percentage points over Base, despite the much larger improvements on their own tasks (Appendix C.8, Table 10).

Category-level training: expanding supervision within one capability does not ensure transfer to others. We combine the member tasks of each of the six capability categories and evaluate SFT, RL, and SFT + RL on all task test sets. Relative to Base, SFT + RL improves within-category accuracy by 28.19–62.00 percentage points and outside-category accuracy by 1.49–8.91 percentage points (Appendix C.8, Table 14).

Joint training: one model trained on all tasks outperforms separate task-specific models in overall accuracy. Training one model on the combined training sets of all 15 tasks yields pooled accuracy of 45.75% with SFT, 16.17% with RL, and 48.23% with SFT + RL. By comparison, the task-specific SFT + RL models achieve 38.91% pooled accuracy on the same test questions, giving joint training a 9.32-percentage-point advantage (Appendix C.8, Table 18). These results suggest that joint training across capability categories can strengthen models' surgical reasoning abilities more effectively than single-task training. Appendix E.4 analyzes error patterns across cross-task, category-level, and joint training.

Image and video generation: specifying an action does not ensure that the model follows it. At 100 paired starting points, both models receive the same initial frame and action description. Human review rates 82% of the final-state images generated by Gemini 3.1 Flash Image (Google, 2026a) and 22% of the six-second videos generated by Veo 3.1 (Google, 2026b) as consistent with the recorded continuation. Adding current-state labels to these action-only prompts gives 80% and 28%, respectively. Unrequested events and incorrect action triplets account for 98 of the 150 error-labeled videos across both variants (Appendix F). Generated images and videos cannot yet serve as reliable references for surgical action outcomes.

4 Related Work

Datasets and Methods for Surgical Scene Understanding. Surgical scene understanding builds on phase and tool recognition (Twinanda et al., 2017), instrument–verb–target interactions (Nwoye et al., 2022), and joint activity recognition and segmentation (Ayobi et al., 2025). Visual question answering brings these annotations into language-based evaluation: Surgical-VQA and Surgical-GPT assess questions about surgical scenes (Seenivasan et al., 2022; 2023), while PitVQA and PitVQA++ extend coverage to pituitary surgery with categorical and open-ended answers (He et al., 2024; 2025). Beyond identifying scene contents, Surgical-VQLA++ localizes answers with bounding boxes (Bai et al., 2024), and SSG-VQA uses scene graphs to construct compositional questions about spatial and action relations (Yuan et al., 2024). These works provide a foundation for evaluating both perception and local relational understanding.

Datasets and Methods for Temporal and Procedural Reasoning. Video question answering extends scene understanding to changes over time and the relationships between surgical actions. SurgViVQA evaluates short-clip motion understanding through REAL-Colon-VQA (Drago et al., 2026), while SurgLQA addresses longer contexts through Colon-LQA (Guo et al., 2026b). SurgeryVideoQA covers procedural steps and intraoperative decisions (Thapa et al., 2025), and SurgRAW introduces SurgCoTBench for multi-task reasoning, including action prediction (Low et al., 2026). More explicit temporal evaluation includes action ordering and anomaly tracking in SurgCoT (Wang et al., 2026) and evidence-grounded temporal relations in STSG-VQA (Li and Sarikaya, 2026). SurgTEMP combines perception, surgical assessment, and procedural reasoning in CholeVidQA-32K (Li et al., 2026). SUREON evaluates forecasting, safety assessment, and decision rationale, alongside supervised and reinforcement learning (Perez et al., 2026).

Comprehensive Surgical Evaluation. Broader evaluations examine performance across tasks and clinical settings. EndoBench spans multiple endoscopic scenarios (Liu et al., 2025b), while

Table 3: Benchmark comparison. Input: T denotes text, I denotes an image, S denotes a single video, and M denotes multiple video clips. #Tasks: number of tasks. Size: number of dataset samples. Perception: recognition, counting, localization, and segmentation. Spatial: relative positions and anatomical relations. Temporal: event order, duration, motion, state changes, and prediction. Procedural: workflow reasoning, action purposes, and next-step decisions. Correction: correcting supplied claims or workflows.

Benchmark	Input	#Tasks	Size	Perception	Spatial	Temporal	Procedural	Correction
Surgical scene question answering								
EndoVis18-VQA (Seenivasan et al., 2022)	T+I	–	11,783	✓	–	–	–	–
Cholec80-VQA (Seenivasan et al., 2022)	T+I	–	∼43K	✓	–	–	–	–
PSI-AVA-VQA (Seenivasan et al., 2023)	T+I	3	10,291	✓	–	–	–	–
PitVQA (He et al., 2024)	T+I	6	884,242	✓	–	–	–	–
Open-Ended PitVQA (He et al., 2025)	T+I	6	745,972	✓	–	✓	✓	–
EndoVis17-VQLA (Bai et al., 2024)	T+I	3	708	✓	–	–	–	–
EndoVis18-VQLA (Bai et al., 2024)	T+I	4	16,561	✓	–	–	–	–
SSG-VQA (Yuan et al., 2024)	T+I	5	∼960K	✓	✓	–	–	–
Temporal and procedural reasoning								
REAL-Colon-VQA (Drago et al., 2026)	T+S	20	5,200	✓	–	✓	–	–
SurgeryVideoQA (Thapa et al., 2025)	T+S	–	2,692	✓	✓	✓	✓	–
CholeVidQA-32K (Li et al., 2026)	T+S	11	32,311	✓	✓	✓	✓	–
STSG-VQA (Li and Sarikaya, 2026)	T+S	7	18,458	✓	–	✓	–	–
SUREON (Perez et al., 2026)	T+S	12	206.8K	✓	✓	✓	✓	–
SurgCoTBench (Low et al., 2026)	T+I	5	14,256	✓	–	✓	✓	–
SurgCoT (Wang et al., 2026)	T+S	5	78,522	✓	✓	✓	✓	–
Comprehensive surgical evaluation								
EndoBench (Liu et al., 2025b)	T+I	12	6,832	✓	–	–	–	–
SurgCheck (Shin et al., 2026)	T+I	8	29,635	✓	✓	–	–	–
SurgVLM-Bench (Zeng et al., 2025)	T+I	6	–	✓	✓	–	–	–
Systematic evaluation (Rau et al., 2025)	T+I/S	17	–	✓	✓	✓	✓	–
GATE (ours)	**T+S/M**	**15**	**13,715**	✓	✓	✓	✓	✓

SurgVLM-Bench (Zeng et al., 2025) and the systematic evaluation of Rau et al. (2025) assess diverse surgical capabilities. Complementary analyses examine the evidence behind model responses: SurgCheck tests reliance on visual evidence through controlled question variants (Shin et al., 2026), and Sum-of-Checks decomposes Critical View of Safety assessment into explicit visual checks (You et al., 2026). Table 3 summarizes benchmark coverage. GATE combines recognition and relational judgments with consistency checks and correction of supplied spatial, descriptive, and workflow claims.

5 Conclusion and Limitations

GATE provides an assessment framework that connects a surgical video benchmark with diagnostic analysis and training interventions. Across 11 leading VLMs, the strongest overall accuracy is 49.0%, with recognition and spatial judgments stronger than workflow reasoning and joint description and workflow correction on average. These results establish concrete targets for improving surgical video understanding. Training interventions yield substantial gains in benchmark accuracy, although the benefits vary across tasks.

Although GATE spans multiple anatomical sites, its coverage does not capture the full diversity of procedures and clinical settings. Training interventions are evaluated on a single model backbone, so their effectiveness across model families remains to be established. Future work will broaden the benchmark's coverage and evaluate training interventions across a wider range of model architectures. A further direction is to extend GATE from offline video assessment to real-time evaluation on physical surgical robotic systems.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT and Claude, to aid and polish the writing of the paper, such as improving grammar, clarity, readability, and phrasing of text originally drafted by the authors. AI assistance also included selected code implementation, interpretation of experimental results, and feedback on the design of additional analyses. The authors made the final decisions on the research questions, core ideas, benchmark and methodology design, experiments, analyses, and interpretation of the results.

All AI-assisted edits were reviewed by the authors to ensure that they accurately reflected the intended meaning. We take responsibility for the final content of this work, including any text or code produced with the assistance of generative AI.

ETHICS STATEMENT

We obtained the necessary access permissions for all source datasets used in this study. We use existing surgical datasets for retrospective academic research in accordance with their applicable access and use conditions. For selected datasets, we additionally obtained email confirmation from the data providers permitting academic model evaluation, subject to the original data-use conditions. The source datasets and annotations are credited in Table 4. Data use remains subject to the original privacy and redistribution conditions. This work evaluates computational models and does not involve prospective clinical intervention. Generated images and videos are experimental simulations and do not establish clinical safety or suitability for patient care.

REFERENCES

Anthropic. Introducing Claude Opus 5.5. 2026a. https://www.anthropic.com/claude-opus-5-5. Accessed September 26, 2026.

Anthropic. Introducing Claude Sonnet 5. 2026b. https://www.anthropic.com/news/claude-sonnet-5. Accessed September 26, 2026.

Karl J. Åström and Richard M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, second edition, 2021. https://fbswiki.org/wiki/index.php/Main_Page.

Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, and Pablo Arbeláez. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, 106:103726, 2025. https://doi.org/10.1016/j.media.2025.103726.

Long Bai, Guankun Wang, Mobarakol Islam, Lalithkumar Seenivasan, An Wang, and Hongliang Ren. Surgical-VQLA++: Adversarial Contrastive Learning for Calibrated Robust Visual Question-Localized Answering in Robotic Surgery. *arXiv:2408.04958*, 2024. https://arxiv.org/abs/2408.04958.

Shuai Bai et al. Qwen3-VL Technical Report. *arXiv:2511.21631*, 2025. https://arxiv.org/abs/2511.21631.

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. *arXiv:2511.03325v3*, 2026. https://arxiv.org/abs/2511.03325v3.

Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. Cataract-1K Dataset for Deep-Learning-Assisted Analysis of Cataract Surgery Videos. *Scientific Data*, 11(1):373, 2024. https://doi.org/10.1038/s41597-024-03193-4.

Google. Gemini 3.1 Flash Image. Model documentation, 2026a. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-image. Accessed September 26, 2026.

Google. Veo 3.1. Model documentation, 2026b. https://ai.google.dev/gemini-api/docs/models/veo-3.1-generate-preview. Accessed September 26, 2026.

Google DeepMind. Gemma 4 31B IT. Model card, 2026. https://huggingface.co/google/gemma-4-31B-it. Accessed September 26, 2026.

Diandian Guo, Weixin Si, Zhixi Li, Jialun Pei, and Pheng-Ann Heng. Surgical Workflow Recognition and Blocking Effectiveness Detection in Laparoscopic Liver Resection with Pringle Maneuver. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):3220–3228, 2025. https://doi.org/10.1609/aaai.v39i3.32332.

Diandian Guo, Shu Yang, Jialun Pei, Jiaao Li, Yanhui Wan, Hao Chen, and Pheng-Ann Heng. LungRes80: Towards tangled surgical workflow recognition in video-assisted thoracoscopic surgery. *Medical Image Analysis*, 114:104237, 2026a. https://doi.org/10.1016/j.media.2026.104237.

Diandian Guo, Xikai Yang, Ruiyang Li, Jialun Pei, and Pheng-Ann Heng. SurgLQA: Scalable Long-Horizon Surgical Video Question Answering. *arXiv:2605.17915*, 2026b. https://arxiv.org/abs/2605.17915.

David Ha and Jürgen Schmidhuber. World Models. *arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.

Runlong He, Mengya Xu, Adrito Das, Danyal Z. Khan, Sophia Bano, Hani J. Marcus, Danail Stoyanov, Matthew J. Clarkson, and Mobarakol Islam. PitVQA: Image-grounded Text Embedding LLM for Visual Question Answering in Pituitary Surgery. *arXiv:2405.13949*, 2024. https://arxiv.org/abs/2405.13949.

Runlong He, Danyal Z. Khan, Evangelos B. Mazomenos, Hani J. Marcus, Danail Stoyanov, Matthew J. Clarkson, and Mobarakol Islam. PitVQA++: Vector Matrix-Low-Rank Adaptation for Open-Ended Visual Question Answering in Pituitary Surgery. *arXiv:2502.14149*, 2025. https://arxiv.org/abs/2502.14149.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*, 2021. https://arxiv.org/abs/2106.09685.

Jing Li and Duygu Sarikaya. STSG-VQA: Evidence-Grounded Temporal Question Answering from Surgical Spatio-Temporal Scene Graphs. *arXiv:2609.08543v1*, 2026. https://arxiv.org/abs/2609.08543v1.

Shi Li, Vinkle Srivastav, Shih-Min Yin, Nicolas Chanel, Saurav Sharma, Nabani Banik, Lorenzo Arboit, Kun Yuan, Pietro Mascagni, and Nicolas Padoy. SurgTEMP: Temporal-Aware Surgical Video Question Answering with Text-guided Visual Memory for Laparoscopic Cholecystectomy. *arXiv:2603.29962v5*, 2026. https://arxiv.org/abs/2603.29962v5.

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. *arXiv:2505.23601*, 2025b. https://arxiv.org/abs/2505.23601.

Chang Han Low, Ziyue Wang, Tianyi Zhang, Zhu Zhuo, Zhitao Zeng, Evangelos B. Mazomenos, and Yueming Jin. SurgRAW: Multi-Agent Workflow with Chain of Thought Reasoning for Robotic Surgical Video Analysis. *IEEE Robotics and Automation Letters*, 11(4):4857–4864, 2026. https://doi.org/10.1109/LRA.2026.3665443.

Pietro Mascagni, Deepak Alapatt, Aditya Murali, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12(1):331, 2025. https://doi.org/10.1038/s41597-025-04642-4.

Moonshot AI. Kimi K3. Model card, 2026. https://huggingface.co/moonshotai/Kimi-K3. Accessed September 26, 2026.

Aditya Murali, Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. The Endoscapes Dataset for Surgical Scene Segmentation, Object Detection, and Critical View of Safety Assessment: Official Splits and Benchmark. *arXiv:2312.12429*, 2023. https://arxiv.org/abs/2312.12429.

Aditya Murali, Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Didier Mutter, and Nicolas Padoy. Latent Graph Representations for Critical View of Safety Assessment. *IEEE Transactions on Medical Imaging*, 43(3):1247–1258, 2024. https://doi.org/10.1109/TMI.2023.3333034.

Sahar Nasirihaghighi, Negin Ghamsarian, Leonie Peschek, Matteo Munari, Heinrich Husslein, Raphael Sznitman, and Klaus Schoeffmann. GynSurg: A Comprehensive Gynecology Laparoscopic Surgery Dataset. *arXiv:2506.11356*, 2025. https://arxiv.org/abs/2506.11356.

NVIDIA et al. Nemotron 3 Nano Omni: Efficient and Open Multimodal Intelligence. *arXiv:2604.24954*, 2026. https://arxiv.org/abs/2604.24954.

Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. https://doi.org/10.1016/j.media.2022.102433.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. 2026a. https://openai.com/index/gpt-5-6/. Accessed September 26, 2026.

OpenAI. GPT-6 Astra: A new generation of intelligence. 2026b. https://openai.com/index/gpt-6-astra/. Accessed September 26, 2026.

OpenAI. Introducing GPT-6 Sol and Luna. 2026c. https://openai.com/index/introducing-gpt-6-sol-and-luna/. Accessed September 26, 2026.

Alejandra Perez, Anita Rau, Lee White, Busisiwe Mlambo, Chinedu Nwoye, Muhammad Abdullah Jamal, and Omid Mohareri. SUREON: A Benchmark and Vision-Language-Model for Surgical Reasoning. *arXiv:2603.06570*, 2026. https://arxiv.org/abs/2603.06570.

Anita Rau, Mark Endo, Josiah Aklilu, Jaewoo Heo, Khaled Saab, Alberto Paderno, Jeffrey Jopling, F. Christopher Holsinger, and Serena Yeung-Levy. Systematic Evaluation of Large Vision-Language Models for Surgical Artificial Intelligence. *arXiv:2504.02799*, 2025. https://arxiv.org/abs/2504.02799.

Manuel Sebastián Ríos, María Alejandra Molina-Rodriguez, Daniella Londoño, Camilo Andrés Guillén, Sebastián Sierra, Felipe Zapata, and Luis Felipe Giraldo. Cholec80-CVS: An open dataset with an evaluation of Strasberg's critical view of safety for AI. *Scientific Data*, 10(1):194, 2023. https://doi.org/10.1038/s41597-023-02073-7.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-VQA: Visual Question Answering in Surgical Scenes using Transformer. *arXiv:2206.11053*, 2022. https://arxiv.org/abs/2206.11053.

Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. SurgicalGPT: End-to-End Language-Vision GPT for Visual Question Answering in Surgery. *arXiv:2304.09974*, 2023. https://arxiv.org/abs/2304.09974.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv:2402.03300*, 2024. https://arxiv.org/abs/2402.03300.

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? *arXiv:2605.01911v2*, 2026. https://arxiv.org/abs/2605.01911v2.

Rahul Thapa, Andrew Li, Qingyang Wu, Bryan He, Yuki Sahashi, Christina Binder, Angela Zhang, Ben Athiwaratkun, Shuaiwen Leon Song, David Ouyang, and James Zou. How Well Can General Vision-Language Models Learn Medicine By Watching Public Educational Videos? *arXiv:2504.14391*, 2025. https://arxiv.org/abs/2504.14391.

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. https://doi.org/10.1109/TMI.2016.2593957.

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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. https://arxiv.org/abs/2604.20319v1.

Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, Lijuan Wang, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, Yejin Choi, and Manling Li. VAGEN: Reinforcing World Model Reasoning for Multi-Turn VLM Agents. *arXiv:2510.16907*, 2025a. https://arxiv.org/abs/2510.16907.

Weiyun Wang et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv:2508.18265*, 2025b. https://arxiv.org/abs/2508.18265.

Weiqiu You, Cassandra Goldberg, Amin Madani, Daniel A. Hashimoto, and Eric Wong. Sum-of-Checks: structured reasoning for surgical safety with large vision-language models. *International Journal of Computer Assisted Radiology and Surgery*, 21(6):1433–1438, 2026. https://doi.org/10.1007/s11548-026-03643-x.

Kun Yuan, Manasi Kattel, Joel L. Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing Surgical VQA with Scene Graph Knowledge. *arXiv:2312.10251v3*, 2024. https://arxiv.org/abs/2312.10251v3.

Z.ai. GLM-5.3-Flash. Model card, 2026. https://huggingface.co/zai-org/GLM-5.3-Flash. Accessed September 26, 2026.

Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, Xiaochun Cao, Yutong Ban, Qi Dou, Yang Liu, and Yueming Jin. SurgVLM: A Large Vision-Language Model and Systematic Evaluation Benchmark for Surgical Intelligence. *arXiv:2506.02555v1*, 2025. https://arxiv.org/abs/2506.02555v1.

APPENDIX CONTENTS

A GATE Benchmark Details **16**

A.1 Benchmark Overview and Data Sources 16

A.2 Task Coverage 16

A.3 Question Construction and Quality Control 16

A.4 Phase Understanding 17

A.5 Scene Understanding 17

A.6 Temporal Reasoning 18

A.7 Workflow Reasoning 19

A.8 Description Correction 19

A.9 Description and Workflow Correction 20

A.10 Benchmark Statistics and Data Partitions 20

B Evaluation Protocols **21**

B.1 Models, Visual Inputs, and Prompts 21

B.2 Answer Extraction and Scoring 21

B.3 Random Baseline 22

C Training Methods and Results **22**

C.1 Shared Training Settings 22

C.2 Supervised Fine-Tuning 23

C.3 Reinforcement Learning 23

C.4 Checkpoint Selection and Reporting 23

C.5 Response Conditions and Training Scope 24

C.6 Per-Task Training and Answer-Only Comparison 24

C.7 Learning Dynamics and Output Length 24

C.8 Transfer across Training Scopes 31

D Additional Benchmark Results **40**

D.1 Results by Capability and Anatomical Site 40

D.2 Text-Only Evaluation 40

E Reasoning and Error Analysis **41**

E.1 Benchmark Evaluation 41

E.2 Errors Before and After Training 45

E.3 Answer-Only Errors and Repeated Answers 46

E.4 Errors across Training Scopes 46

F Action-Conditioned Image and Video Generation **46**

F.1 State Transitions and Rollouts 46

F.2 Task and Conditioning 47
F.3 Generation and Evaluation Protocol 47
F.4 Results 48
F.5 Action-Level Differences 49

G Prompt Templates and Task Definitions **50**
G.1 Complete Prompt Examples 50

A GATE BENCHMARK DETAILS

A.1 BENCHMARK OVERVIEW AND DATA SOURCES

GATE contains 15 tasks covering phase understanding, scene understanding, temporal reasoning, workflow reasoning, description correction, and joint description and workflow correction. Each question provides sampled surgical clips and asks for a structured answer, such as a phase label, an ordered sequence, or a correction. Reference answers follow source annotations and task-specific rules. Figure 2 illustrates the tasks.

We use eight datasets spanning gallbladder, eye, prostate, lung, liver, and gynecologic surgery. Table 4 lists the annotations used from each source. Cholec80-CVS adds Critical View of Safety (CVS) annotations to Cholec80 and is not counted as a separate video dataset. Cataract1K questions use its phase-recognition and segmentation subsets: phase questions use frames sampled at 1 frame per second (fps), while object-based questions use directly annotated images. GynSurg action questions use its annotated action clips. Source data remain subject to the original release conditions.

Table 4: Source datasets and annotations used to construct GATE questions. Dataset names link to the official release instructions. Citations identify the corresponding source publications. The descriptions concern the inputs required by GATE, not an exhaustive account of each dataset.

Source	Required annotation or media input
Cholec80 (Twinanda et al., 2017)	Official 1-fps frames, phase labels, and tool-presence labels.
Cholec80-CVS (Ríos et al., 2023)	CVS assessments aligned with Cholec80 frames as an annotation supplement for the same videos.
CholecT50 (Nwoye et al., 2022)	1-fps images with phase and instrument–action–target annotations.
Cataract1K (Ghamsarian et al., 2024)	Videos with temporal phase annotations and directly annotated images with object segmentations.
Endoscapes2023 (Mascagni et al., 2025; Murali et al., 2023; 2024)	CVS annotations, surrounding video frames, and frames with spatial annotations.
GraSP (Ayobi et al., 2025)	1-fps frames with dense phase/step labels and sparse instrument/action or spatial annotations.
GynSurg (Nasirihaghighi et al., 2025)	Three-second action clips and the instrument/anatomy segmentation subset.
LungRes80 (Guo et al., 2026a)	Video frames, phase annotations, and author-provided data partitions.
PmLR50 (Guo et al., 2025)	Video frames, phase labels, and ineffective-blocking annotations.

A.2 TASK COVERAGE

Table 5 reports the 13,715 questions retained after quality control across 56 task–dataset combinations. The six categories contain two, three, five, three, one, and one tasks, respectively. Coverage depends on the annotations needed for each task. For example, State Change Tracking uses Cholec80, CholecT50, and GraSP. Spatial Relation Tracking uses GraSP and GynSurg. Counts refer to retained questions, rather than the full size of each source dataset.

A.3 QUESTION CONSTRUCTION AND QUALITY CONTROL

Each clip contains four distinct frames in chronological order. When a task requires a stable clip, the queried attributes must agree across those frames. For clip $C_i = (x_{i1}, \ldots, x_{i4})$ and task-relevant

Table 5: Questions approved through quality control (QC) in each task–dataset combination. The full benchmark contains 13,715 questions across 56 populated combinations. The symbol – denotes no questions in the benchmark.

Task	Cholec80	CholecT50	Cataract1K	Endoscapes2023	GraSP	GynSurg	LungRes80	PmLR50
Phase Recognition	145	153	166	–	166	–	155	165
Phase Consistency	129	143	167	–	160	–	158	165
Action Description Matching	–	626	–	–	–	203	–	–
Shared CVS Correction	489	–	–	492	–	–	–	–
Spatial Repair	–	–	500	–	–	293	–	–
Reference Alignment	–	451	–	–	500	–	–	–
Temporal Order Repair	112	133	167	–	154	–	151	165
CVS Milestone Tracking	481	–	–	443	–	–	–	–
State Change Tracking	191	477	–	–	246	–	–	–
Spatial Relation Tracking	–	–	–	–	494	500	–	–
Workflow Ordering	152	174	199	–	194	–	–	199
Workflow Path Recovery	141	185	249	–	–	–	–	245
Workflow Repair	127	145	167	–	167	–	166	166
Description Correction	273	361	273	–	–	–	–	91
Description and Workflow Correction	189	296	333	–	–	–	–	83
Total	2,429	3,144	2,221	935	2,081	996	630	1,279

annotated state s_τ,

$$s_\tau(x_{i1}) = s_\tau(x_{i2}) = s_\tau(x_{i3}) = s_\tau(x_{i4}). \tag{1}$$

CVS stability applies to the full three-criterion vector. Chronological clips are non-overlapping where required, but separate questions may reuse frames or surgical cases.

Sampling. Frame spacing follows the source annotations. Cholec80 and CholecT50 state-change clips use 1-fps frames. Endoscapes2023 CVS clips use reviewed keyframes five seconds apart. Cataract1K description-correction and joint-correction clips use successive annotated images approximately five seconds apart. Spatial Repair can use nonconsecutive annotated images. PmLR50 sequence indices specify order, not elapsed seconds. Consequently, clip duration varies, and observed stability does not establish stability during unobserved gaps.

Construction and review. Generators select eligible clips, construct questions, and derive answers from their annotations. Depending on the task, they introduce distractors, reorder clips, or corrupt descriptions. Candidate quotas are allocated across eligible datasets, with deterministic sampling over seeds 0–9,999. Spatial Repair assigns a fixed clip to each random seed. These quotas do not enforce balanced answers or surgical cases.

Automatic checks verify annotation–question–answer consistency, image availability, and duplicate removal. Reviewers then inspect the visual evidence and record a pass or fail decision through the interface in Figure 5. Human screening removes 1,048 of 14,763 candidates, leaving 13,715 questions with recorded pass decisions. Each retained record stores its dataset, task, question, images, and reference answer.

A.4 Phase Understanding

Phase Recognition. One clip is shown with four phase options from the source dataset's vocabulary. Source phase annotations determine frame selection and the correct option. The model returns one letter, A–D.

Phase Consistency. Four clips A–D either share one phase or contain three clips from one phase and a single outlier. The model returns the outlier's label, or E if all four are consistent. This task tests phase comparison across clips.

A.5 Scene Understanding

Action Description Matching. Five clips A–E are paired with four numbered descriptions. CholecT50 descriptions specify an instrument–action–target triplet: the instrument, its action, and the anatomical or surgical target. GynSurg uses its native action descriptions. Every clip has a

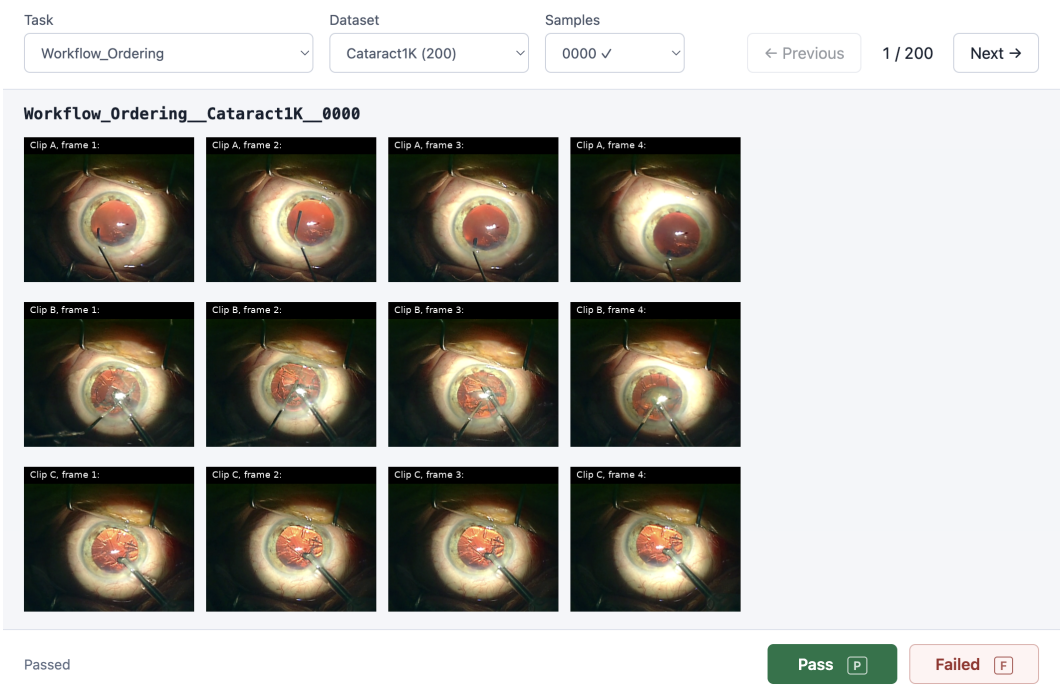


Figure 5: **Human image quality-control interface.** Reviewers inspect the sampled visual evidence and record a pass or fail decision.

match, and descriptions can be reused. Either all four description numbers are used or exactly one is unused. The answer lists five description numbers in clip order.

Shared CVS Correction. Four clips from different surgeries are compared using the three CVS criteria: a cleared hepatocystic triangle, separation of the lower gallbladder to expose the cystic plate, and identification of only two structures entering the gallbladder. Cholec80 and Endoscapes2023 annotations provide a stable criterion vector for each clip. The model selects the corrective operations required by *all* four clips, corresponding to the intersection of their required-operation sets. It lists operation numbers in ascending order without duplicates, or indicates that none is required by all four clips. This includes cases where every clip already shows complete CVS.

Spatial Repair. A clip is paired with three spatial claims whose object names are correct but whose relations may be corrupted. Every relation in a valid claim must hold in all four frames. For each incorrect claim, the answer lists only the relations that need repair, in their displayed order. Corrections stay within the corresponding pair: left of or right of, above or below, and overlapping or non-overlapping.

Questions use directly annotated Cataract1K and GynSurg images. To exclude ambiguous geometry, directional relations require bounding-box center displacement of at least 15% of the corresponding image dimension, and each object must occupy at least 0.1% of image area. Polygon-based overlap requires interior intersection. Touching or missing polygons are rejected. Disjoint pairs require no mask intersection and a bounding-box gap of at least 2% of the shorter image side.

A.6 Temporal Reasoning

Reference Alignment. Three reference clips R1–R3 show an ordered pattern, and three query clips A–C from another case show matching events in a shuffled order. The model returns each query clip label exactly once in reference order. CholecT50 questions match three distinct action–target pairs within a common phase. Instruments may differ across cases, but phase and triplet labels are stable within each clip. GraSP questions match three distinct steps within a phase, using four

consecutive frames with stable phase and step labels. Matched events need not be adjacent in the original video.

Temporal Order Repair. Five clips from one case and phase are either chronological or corrupted by swapping one pair, which may be nonadjacent. The model lists the two clip labels to swap in alphabetical order, or indicates that no repair is needed. Because all clips share a phase, phase identity alone does not determine their chronology.

CVS Milestone Tracking. Six chronological, non-overlapping clips A–F have stable CVS criterion vectors. The model identifies the first complete-CVS clip and the later clip in the first consecutive pair of complete-CVS clips. With $z_i = \prod_{j=1}^{3} c_{ij}$ indicating complete CVS, these positions are

$$i_{\text{first}} = \min\{i : z_i = 1\}, \qquad i_{\text{pair}} = \min\{i \in \{2, \ldots, 6\} : z_{i-1} = z_i = 1\}. \tag{2}$$

The answer lists the two corresponding clip labels, indicating when either milestone is absent. Persistence refers to consecutive displayed clips, not continuous attainment during unobserved intervals.

State Change Tracking. Four chronological clips A–D from one case have stable values for the requested attributes. Cholec80 questions compare phase and instrument set. CholecT50 questions compare phase and instrument, action, and target sets. GraSP questions compare phase and surgical step. Any addition or removal changes an annotated set. For each requested attribute, the answer lists all later clip labels where it changes from the preceding clip: B for A-to-B, C for B-to-C, and D for C-to-D. If it does not change, the answer states this.

Spatial Relation Tracking. Four chronological clips show a specified subject and object. The model tracks the requested relation types, such as horizontal and vertical, using GraSP or GynSurg spatial annotations. Relations are stable within each clip. As in State Change Tracking, the answer lists, for each requested relation type, all later clip labels at which it changes, or indicates that it remains unchanged. Transition labels refer to displayed clips without assuming a fixed interval between clips.

A.7 Workflow Reasoning

Workflow Ordering. Three clips from one case are shuffled. The model lists all clip labels from earliest to latest. Questions use Cholec80, CholecT50, Cataract1K, GraSP, and PmLR50. Source chronology and dataset-specific admissible phase sequences determine the reference order.

Workflow Path Recovery. Clips A and F define the start and end of a workflow interval. From candidates B–E, the model selects every clip that belongs between the endpoints and orders the selection procedurally. The answer lists up to four selected clips in order, or indicates that none qualifies.

Workflow Repair. Four clips contain one constructed error: missing stages, an adjacent inversion, an extra duplicate or unrelated clip, or an incorrect phase. The model describes the minimum repair by specifying the change, the affected clip positions, and any phases to insert or substitute. Clip positions are numbered from one. Insertion specifies the two positions around a gap and lists the missing phases in order. Swapping specifies two positions in ascending order without adding phases. Deletion specifies one position without adding phases. Replacement specifies one position and the expected phase. For adjacent duplicated phases, the later clip is the extra occurrence. Workflow rules describe the selected case and task construction rather than a universal order for every valid surgery.

A.8 Description Correction

A stable clip is paired with a structured description containing correct and potentially corrupted values. The model returns the complete corrected description, preserving its structure and all valid values. The described attributes follow the source annotations. Cholec80 uses procedure, phase, instruments, and CVS state. CholecT50 uses procedure, phase, instrument, action, and target. Cataract1K

uses procedure, phase, instruments, and anatomy. PmLR50 uses procedure and phase. The prompt specifies legal values and ordering rules for attributes containing multiple categories.

A.9 Description and Workflow Correction

Four clips are paired with Claims A–D. The model checks their workflow and local descriptions together. Claim attributes are stable within each clip, and the prompt selects attributes supported by Cholec80, CholecT50, Cataract1K, or PmLR50. Cataract1K claims may also contain spatial relations. Their subject, object, and relation type are fixed, while the relation value may require correction.

The answer specifies only necessary corrections to the workflow and Claims A–D. Workflow edits follow Section A.7. Each incorrect claim is returned in full with all correct values preserved. Correct components are omitted. If no repair is needed, the answer proposes no changes. Thus, Description Correction returns the entire description, while this task returns complete claims only when they require repair.

A.10 Benchmark Statistics and Data Partitions

The benchmark contains 13,715 questions based on 36,264 clips, with 793–998 questions per task. Figure 3 summarizes the distribution, and Table 5 gives source-level counts. External-model evaluation uses the full pool. Training comparisons use 9,335 training, 1,336 validation, and 2,660 test questions (Table 6). Within each task, questions and the image files they reference are disjoint across partitions. Surgical cases can overlap.

Table 6: Full benchmark size and partitions used for per-task training comparisons. Validation is used for checkpoint selection. Each task's partitions have distinct questions and referenced image files. The full pool is retained for external-model evaluation.

Task	Full pool	Train	Validation	Test	Excluded
Phase Recognition	950	665	95	190	0
Phase Consistency	922	646	92	184	0
Action Description Matching	829	564	97	148	20
Shared CVS Correction	981	437	50	130	364
Spatial Repair	793	555	79	159	0
Reference Alignment	951	666	95	190	0
Temporal Order Repair	882	618	88	176	0
CVS Milestone Tracking	924	647	92	185	0
State Change Tracking	914	640	91	183	0
Spatial Relation Tracking	994	696	99	199	0
Workflow Ordering	918	642	92	184	0
Workflow Path Recovery	820	574	82	164	0
Workflow Repair	938	656	94	188	0
Description Correction	998	698	100	200	0
Description and Workflow Correction	901	631	90	180	0
Total	13,715	9,335	1,336	2,660	384

To separate overlapping images within tasks, Shared CVS Correction retains 617 of 981 questions: Cholec80 contributes 208/26/77 and Endoscapes2023 229/24/53 training/validation/test questions. Action Description Matching retains 809 of 829 questions, excluding 20 GynSurg questions. Other tasks use their full pools. All 384 excluded questions remain in the external-model benchmark with unchanged reference answers.

Partitions are established per task. When pooled for cross-task, category, or joint training, training and validation share 3,232 distinct image files, training and test share 6,239, and validation and test share 1,366. These counts concern image files rather than questions. Such experiments evaluate held-out questions with possible prior exposure to their images through another task. They do not establish image-disjoint transfer.

B Evaluation Protocols

B.1 Models, Visual Inputs, and Prompts

Table 7 lists the evaluated models and their identifiers. All 11 VLMs use their default reasoning settings.

Table 7: Details of vision-language models and generative models. Full Name gives the corresponding model identifier.

Organization	Model Name	Full Name
Proprietary Models		
OpenAI	GPT-6 Astra (OpenAI, 2026b)	gpt-6-astra
OpenAI	GPT-6 Luna (OpenAI, 2026c)	gpt-6-luna
OpenAI	GPT-5.6 Sol (OpenAI, 2026a)	gpt-5.6-sol
OpenAI	GPT-5.6 Luna (OpenAI, 2026a)	gpt-5.6-luna
Anthropic	Claude Opus 5.5 (Anthropic, 2026a)	claude-opus-5-5
Anthropic	Claude Sonnet 5 (Anthropic, 2026b)	claude-sonnet-5
Open-Weight models		
Moonshot AI	Kimi K3 (Moonshot AI, 2026)	moonshotai/kimi-k3
Z.ai	GLM-5.3-Flash (Z.ai, 2026)	z-ai/glm-5.3-flash
OpenGVLab	InternVL3.5-241B-A28B (Wang et al., 2025b)	internvl3.5-241b-a28b
NVIDIA	Nemotron-3-Nano-Omni-30B-A3B-Reasoning (NVIDIA et al., 2026)	nvidia/nemotron-3-nano-omni-30b-a3b-reasoning
Google	Gemma 4 31B IT (Google DeepMind, 2026)	gemma-4-31b-it
Image Generative Models		
Google	Gemini 3.1 Flash Image (Google, 2026a)	gemini-3.1-flash-image
Video Generative Models		
Google	Veo 3.1 (Google, 2026b)	veo-3.1-generate-preview

Each frame is resized to a short side of 256 pixels while preserving aspect ratio. A clip's four frames are arranged left to right in one image with 2-pixel borders and 8-pixel gaps, giving one to six clip images per question. Model-specific image processors may resize these images further. Clip and frame labels replace source filenames and case identifiers.

Prompts contain a task description, labeled visual evidence, rules, a question, and the required answer format. The standard reasoning condition requests an explanation followed by a final answer, with separate opening and closing markers for each. Depending on the task, the answer is a single value, an ordered list, or a set of named components. Answer-only training requests only the final-answer section.

B.2 Answer Extraction and Scoring

Answers are extracted and normalized using task-specific rules. The order of named answer components does not affect scoring, and equivalent labels are treated alike under each task's answer requirements.

Full responses, extracted answers, correctness, response length, and available information about why generation stopped are retained. Supervised-model evaluation and RL rollouts allow 1,536 output tokens. Reaching this limit does not invalidate an already complete answer. An unfinished answer remains invalid. Transient request failures are retried and do not enter the response-outcome counts.

Missing information about why generation stopped is not taken as evidence that the response was cut short.

B.3 Random Baseline

The random baseline makes one prediction per question with seed 20260919, without observing the target. It samples uniformly from displayed single-choice options, draws legal permutations for permutation tasks, and samples each position or attribute from its allowed values for other structured answers. Multi-label and repair tasks use their legal subsets, operations, positions, and values. CVS Milestone Tracking samples structurally valid pairs of first-complete and first-consecutive-complete clips. Overall random accuracy is 8.61%.

C Training Methods and Results

C.1 Shared Training Settings

All runs start from Qwen3-VL-2B-Instruct and use one NVIDIA H200. SFT uses Transformers Reinforcement Learning (TRL) with Parameter-Efficient Fine-Tuning (PEFT) adapters. RL uses VAGEN-Lite (Wang et al., 2025a), Volcano Engine Reinforcement Learning (VERL) with Fully Sharded Data Parallel (FSDP) updates, and vLLM generation. vLLM also evaluates saved models. Each rollout consists of one question, one response, and one terminal reward. Table 8 lists the optimization, decoding, and checkpoint settings.

Table 8: Shared settings for per-task training comparisons. SFT and RL use different optimization objectives but the same visual inputs, partitions, and answer extraction procedure.

Setting	Value
Model / hardware	Qwen3-VL-2B-Instruct / one NVIDIA H200 per run
Random seed	20260705
Visual input	Four 256-short-side frames per clip, 1×4 montage, at most six clip images
Token limits	Prompt 12,000, completion 1,536
SFT adaptation	LoRA rank 32, scaling 64, dropout 0.05, all linear modules
SFT optimization	Three epochs, batch 5, microbatch 1, learning rate 10^{-4}, cosine decay, warm-up 3%
RL adaptation	Full parameters, vision tower unfrozen
RL optimization	One epoch, 5 questions per batch, 8 responses per question, learning rate 10^{-6}, actor microbatch 2
RL regularization	proximal policy optimization (PPO) clip 0.2, Kullback–Leibler (KL) divergence loss coefficient 0.001, entropy coefficient 0
Optimizer	AdamW, betas $(0.9, 0.999)$, weight decay 0.01, gradient norm limit 1
Rollout generation	Temperature 0.8, top-p 0.95, at most 16 concurrent sequences
Evaluation generation	Greedy, temperature 0, top-p 1
Reward	Correct final answer $+5$, output format $+0.5$
Checkpoint selection	Lowest validation loss for SFT, highest validation accuracy for RL
Validation interval	Every 20 updates and final update for per-task training

Visual inputs use the same four-frame montages as evaluation. The Qwen processor allows 65,536–1,048,576 pixels per image. Prompts are limited to 12,000 tokens and completions to 1,536. Over-length supervised examples are excluded. SFT sequences are right-padded, with prompt and padding tokens masked from loss. Rollout prompts are left-padded, and the response mask excludes padding.

The recorded SFT environment uses PyTorch 2.8.0, Transformers 5.10.0.dev0, TRL 0.26.2, PEFT 0.19.1, vLLM 0.11.0, and FlashAttention 2.8.3. Training uses bfloat16 (BF16) precision and gradient checkpointing. FlashAttention and the inference key–value (KV) cache are enabled. RL also

offloads FSDP parameters and optimizer state to the central processing unit (CPU) between relevant computations.

C.2 Supervised Fine-Tuning

SFT learns a reference explanation followed by the answer in the requested format. Loss covers the entire model-generated completion, including the markers separating its parts and indicating its end. Image, prompt, assistant-prefix, and padding tokens are masked. Validation uses completion loss, while test evaluation generates answers without access to the reference explanation.

Reference explanations are deterministic templates derived from construction annotations. Phase tasks state annotated phases. Action tasks verbalize triplets or action labels. CVS and tracking tasks describe the annotated states and required comparison. Reference Alignment states the annotated correspondence. Temporal Order Repair supplies the swap or no-swap decision without claiming that phase labels explain within-phase chronology. The amount of intermediate supervision therefore differs across tasks.

SFT runs for three epochs with LoRA adapters on all linear modules selected by PEFT. Base parameters remain frozen. Adapter rank, learning rate, batch size, and other settings appear in Table 8.

C.3 Reinforcement Learning

Direct RL starts from Base. SFT + RL starts from the base model merged with the SFT adapter selected by validation loss. Both use full-parameter GRPO training with the vision tower unfrozen. Each update samples five questions and eight responses per question, computes within-question relative advantages, and optimizes all generated tokens.

Training runs for one epoch with one PPO epoch per rollout batch. The KL term is added to the optimization loss relative to the initialization policy. It is separate from the terminal task reward:

$$R(o, y) = 5\, m_\tau(o, y) + 0.5\, F(o), \tag{3}$$

where $m_\tau(o, y)$ indicates answer correctness and F indicates format compliance. Reasoning mode requires a reasoning section containing text and an answer that follows the task's rules. Answer-only mode requires exactly one final-answer section containing text. Answer correctness contributes 5 to the reward, format compliance adds 0.5, and a correct formatted answer earns 5.5. Explanation content is not directly scored. Equal rewards within a group provide no relative correctness signal. The final incomplete prompt batch is retained. Rollout sampling and greedy evaluation settings are listed in Table 8.

C.4 Checkpoint Selection and Reporting

Per-task runs validate every 20 optimizer updates and at the final update. SFT selects the lowest validation loss. RL selects the highest validation accuracy, favoring later checkpoints on ties. Best and final weights are retained, with one copy when they coincide. SFT tests its best model. RL tests both best and final. Primary results always use validation-selected checkpoints.

Base evaluation precedes SFT and direct RL. Sequential SFT + RL separately evaluates its merged SFT initialization under the same test settings. The result tables use one shared Base score for SFT and direct RL, and one shared SFT score for SFT and the SFT + RL reference. Transfer, category, and joint tables reuse the per-task Base values in Table 2. The SFT reporting rule also applies to category and joint models. When cross-task evaluation has no second SFT record, the available SFT result is used.

Paired error analyses compare the original response records from each evaluation. Their pre-training accuracy can differ from the shared-reference score because the underlying evaluation records differ. Pooled scores weight questions equally.

C.5 Response Conditions and Training Scope

Answer-only runs preserve images, questions, options, answers, splits, and optimization settings while removing the reasoning block from the prompt and supervised completion. We apply this condition to SFT, RL, and SFT + RL.

Cross-task, category-level, and joint experiments use reasoning mode. For cross-task transfer, each of the 45 per-task models is evaluated on the other 14 task test sets, giving 630 source–target evaluations. Category training fits six models per method on the combined member-task partitions. The two singleton categories reuse the corresponding per-task models. Joint training fits one model per method on all task partitions. Pooling concatenates questions without task or dataset balancing.

Category validation intervals are 40, 50, 100, 60, 20, and 20 updates for categories A–F. Joint runs validate every 300 updates. All also validate at the final update. Training scopes share hyperparameters but differ in data volume and possible image exposure across tasks (Section A.10).

C.6 Per-Task Training and Answer-Only Comparison

Tables 2 and 9 compare reasoning and answer-only training on the same 2,660 test questions. Each trained entry comes from a separate task-specific model selected on validation data.

Table 9: Answer-only per-task training results for Qwen3-VL-2B-Instruct (accuracy, %). Base and SFT use shared reference scores defined in Section C.4. Each trained entry uses a separate model trained for that task and its validation-selected best checkpoint. Darker and lighter purple indicate the best and second-best accuracies within each task, respectively. Tied best results share the darker shade.

Model	**Phase Understanding**		**Scene Understanding**			**Temporal Reasoning**					**Workflow Reasoning**			**Description Correction**	**Description and Workflow Correction**
	Phase Recognition	Phase Consistency	Action Matching	CVS Correction	Spatial Repair	Reference Alignment	Temporal Repair	CVS Tracking	State Tracking	Spatial Relation	Workflow Ordering	Workflow Recovery	Workflow Repair	Description Correction	Description and Workflow Correction
Base model	28.42	17.39	0.00	4.62	0.00	16.32	0.00	0.00	1.64	0.00	19.57	0.61	2.13	0.50	0.00
SFT	52.11	33.15	11.49	16.92	89.31	20.53	11.93	64.86	1.64	29.65	16.30	28.05	13.83	60.00	32.78
RL	31.05	29.35	0.00	4.62	0.00	16.84	0.00	0.00	0.00	3.02	14.67	0.61	13.83	1.00	31.67
SFT + RL	52.11	29.35	13.51	41.54	87.42	21.58	11.36	64.86	2.73	31.16	16.30	30.49	13.83	66.00	32.22

The response analysis in Section E.2 shows that SFT also sharply reduces invalid answers. Its benefit therefore includes learning how to express a task-valid response, a prerequisite for receiving correctness reward, as well as changing the decision itself.

The benefit of explicit reasoning depends on the task. Answer-only SFT performs better on Spatial Repair, while reasoning-mode SFT performs better on Workflow Ordering.

C.7 Learning Dynamics and Output Length

The learning curves use recorded training and validation histories. Training loss is shown as a trailing mean over 2% of each run's updates. Validation measurements are unsmoothed. Aggregate curves align runs by fractional training progress and weight tasks equally. These averages summarize different tasks rather than repeated seeds. Individual RL panels use recorded steps. Stars, where shown, mark selected post-update checkpoints. The step-zero model is shown only as a reference.

Continued fitting of training data can outlast validation improvement. Figure 6 shows a rapid initial drop in SFT loss, followed by slower changes in validation loss while training loss continues to fall. Figures 7 and 8 show this pattern at the task level. Reference Alignment reaches its best validation loss relatively early, while its training loss keeps decreasing. This supports selecting checkpoints using validation data. Absolute losses across response conditions are not directly comparable because their supervised tokens differ.

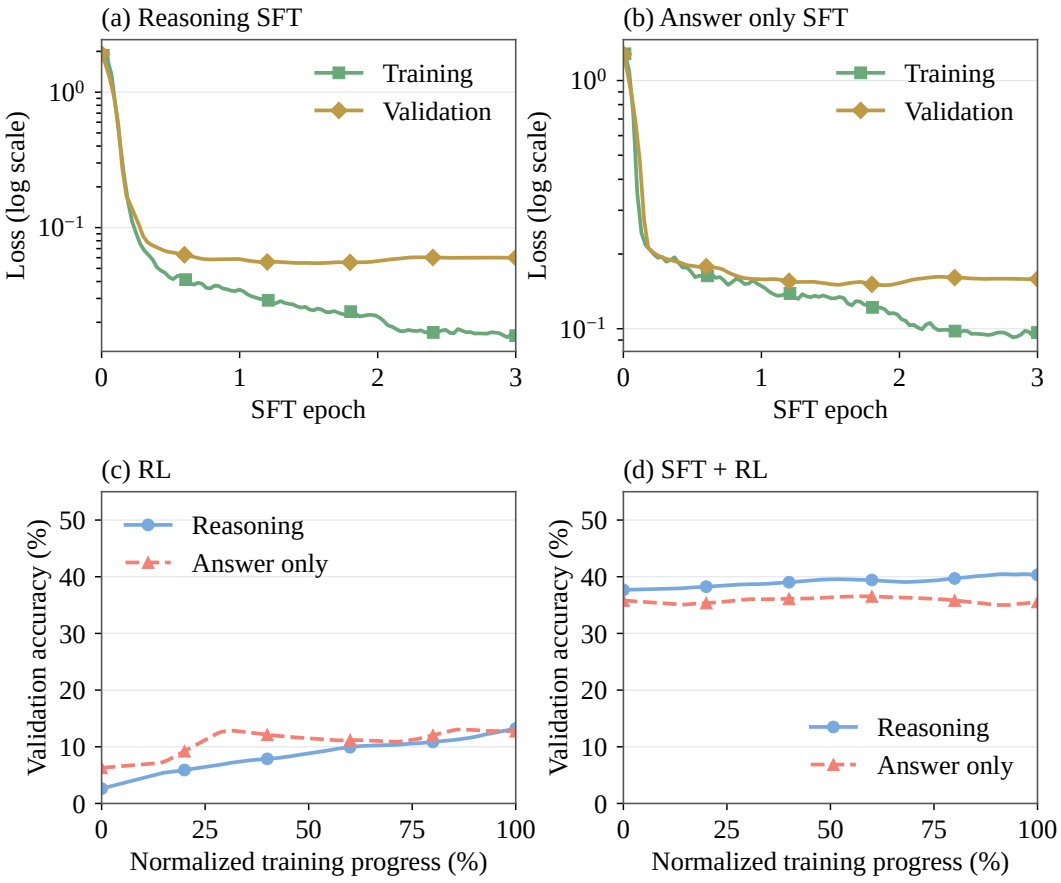


Figure 6: Per-task learning dynamics averaged over the 15 tasks. The top panels show SFT loss separately for reasoning and answer-only supervision. Their loss magnitudes are not directly comparable because the supervised tokens differ. The bottom panels compare validation accuracy during direct RL and SFT + RL. Normalized progress covers the RL stage only.

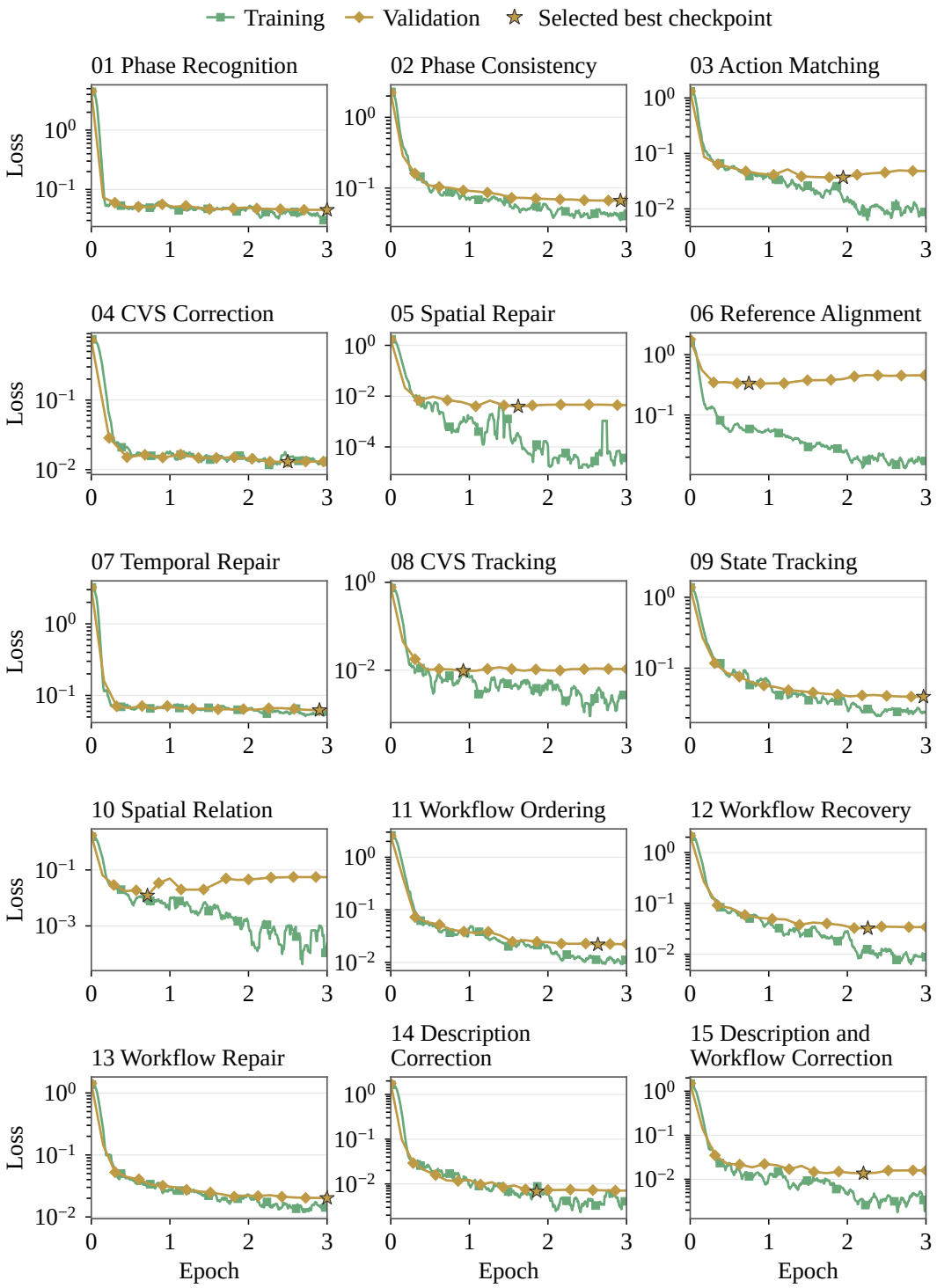


Figure 7: Reasoning-mode SFT convergence for every task. Each panel shows training loss and the measured validation loss, with the selected best checkpoint marked by a star. Loss axes are logarithmic.

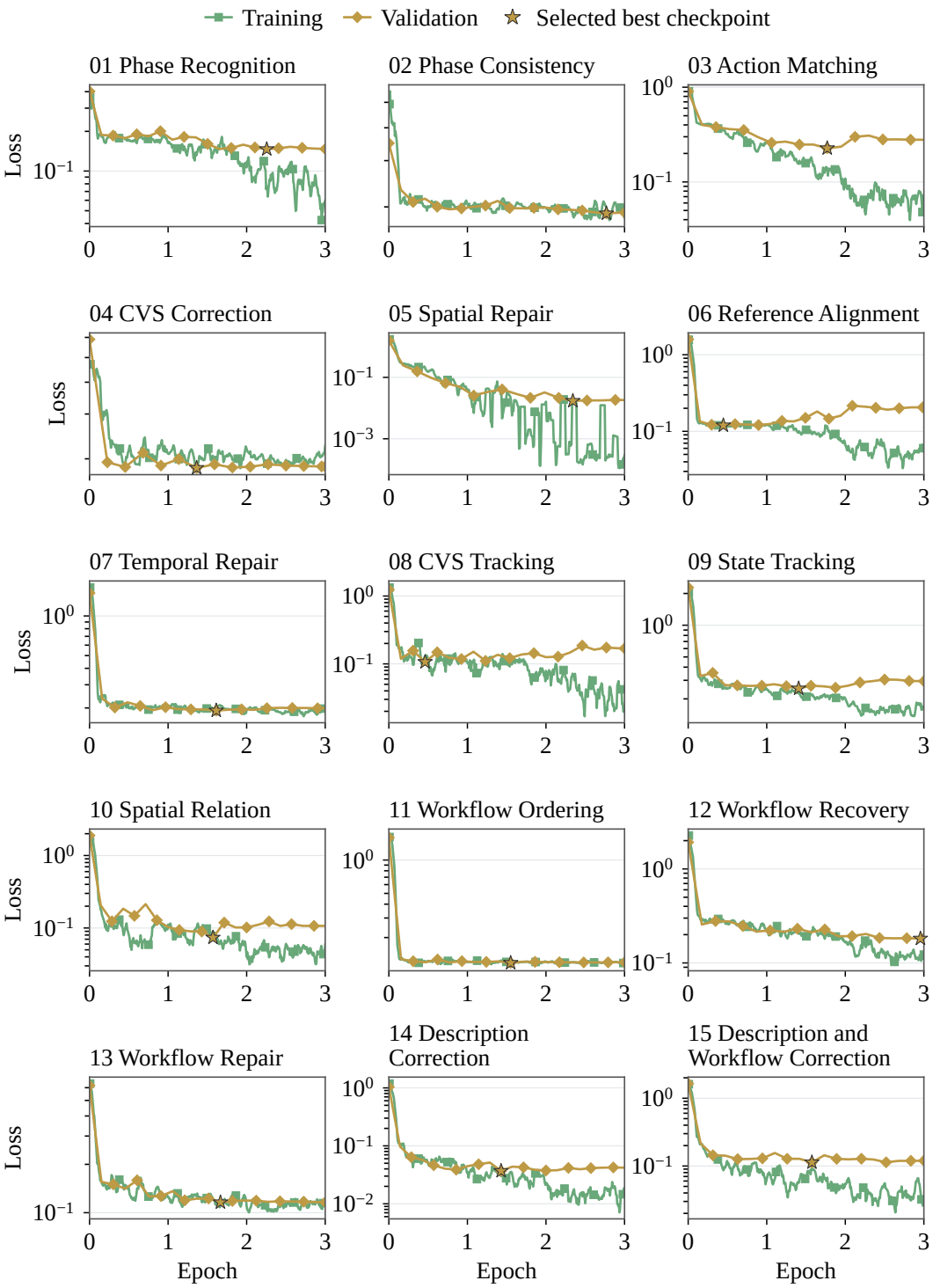


Figure 8: Answer-only SFT convergence for every task, using the same presentation as Figure 7. Training and validation loss are shown on logarithmic axes. Stars identify the selected best checkpoints.

The average RL curve hides substantial task differences. Figure 9 compares direct RL and SFT + RL under both response conditions. Direct RL stays near its starting accuracy on Action Description Matching and CVS Milestone Tracking, whereas SFT-initialized runs retain a higher level. Other

tasks improve or fluctuate during RL. These validation trajectories complement the test results. A smooth average does not imply consistent progress on every task.

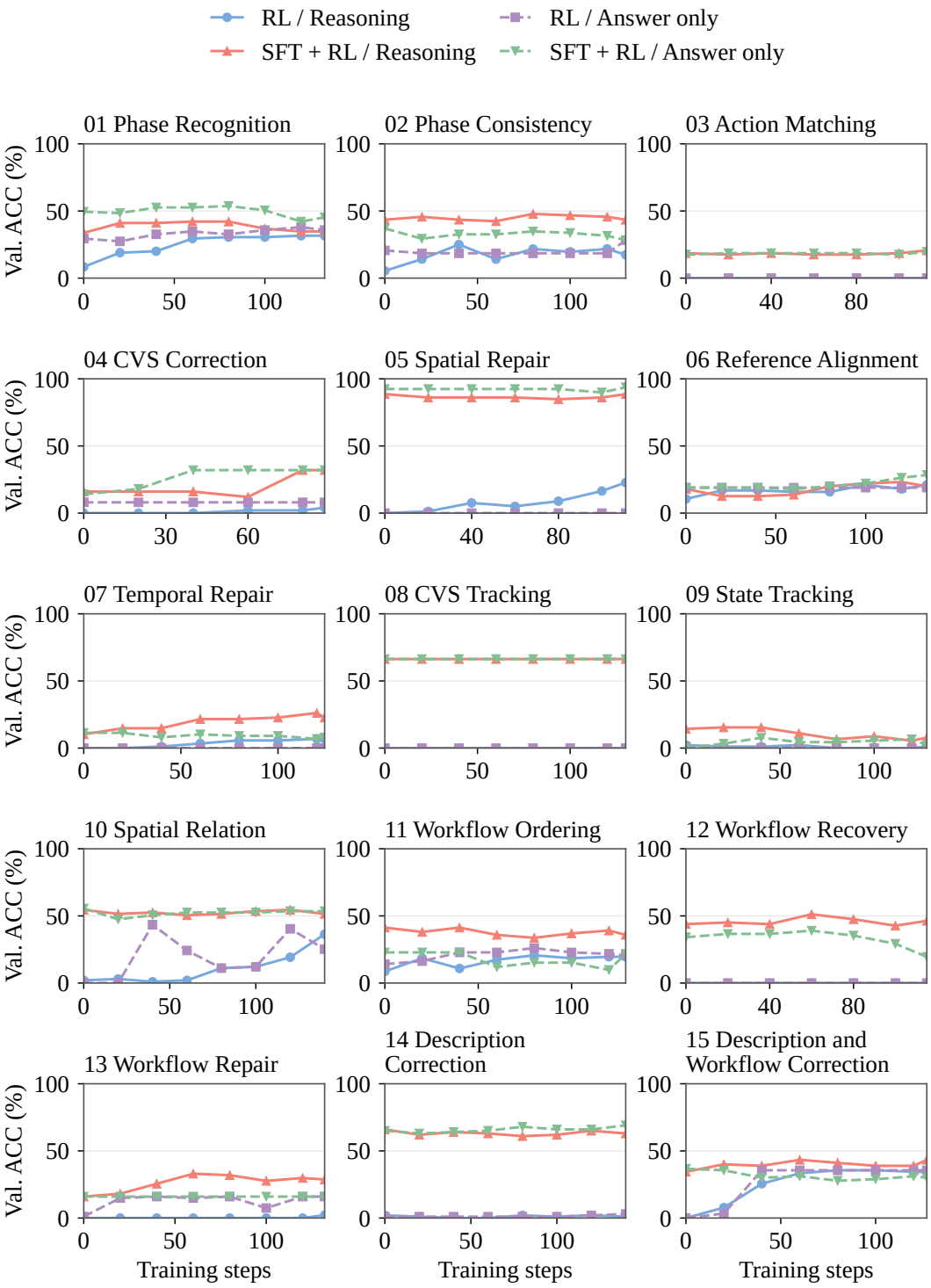


Figure 9: Task-level RL learning curves for validation accuracy (Val. ACC). Each panel compares direct RL and SFT + RL under reasoning and answer-only output conditions. Markers show actual validation evaluations, connected by straight lines. Horizontal axes show recorded training steps from the start of RL.

Longer responses do not consistently accompany better answers. Figure 10 tracks response length and validation accuracy together. Reasoning responses are longer, but accuracy does not increase monotonically with length. Direct answer-only RL improves while its responses become shorter. Additional generated tokens are therefore not a prerequisite for higher task accuracy in this setting. Length counts all generated tokens and cannot distinguish useful intermediate reasoning from repetition.

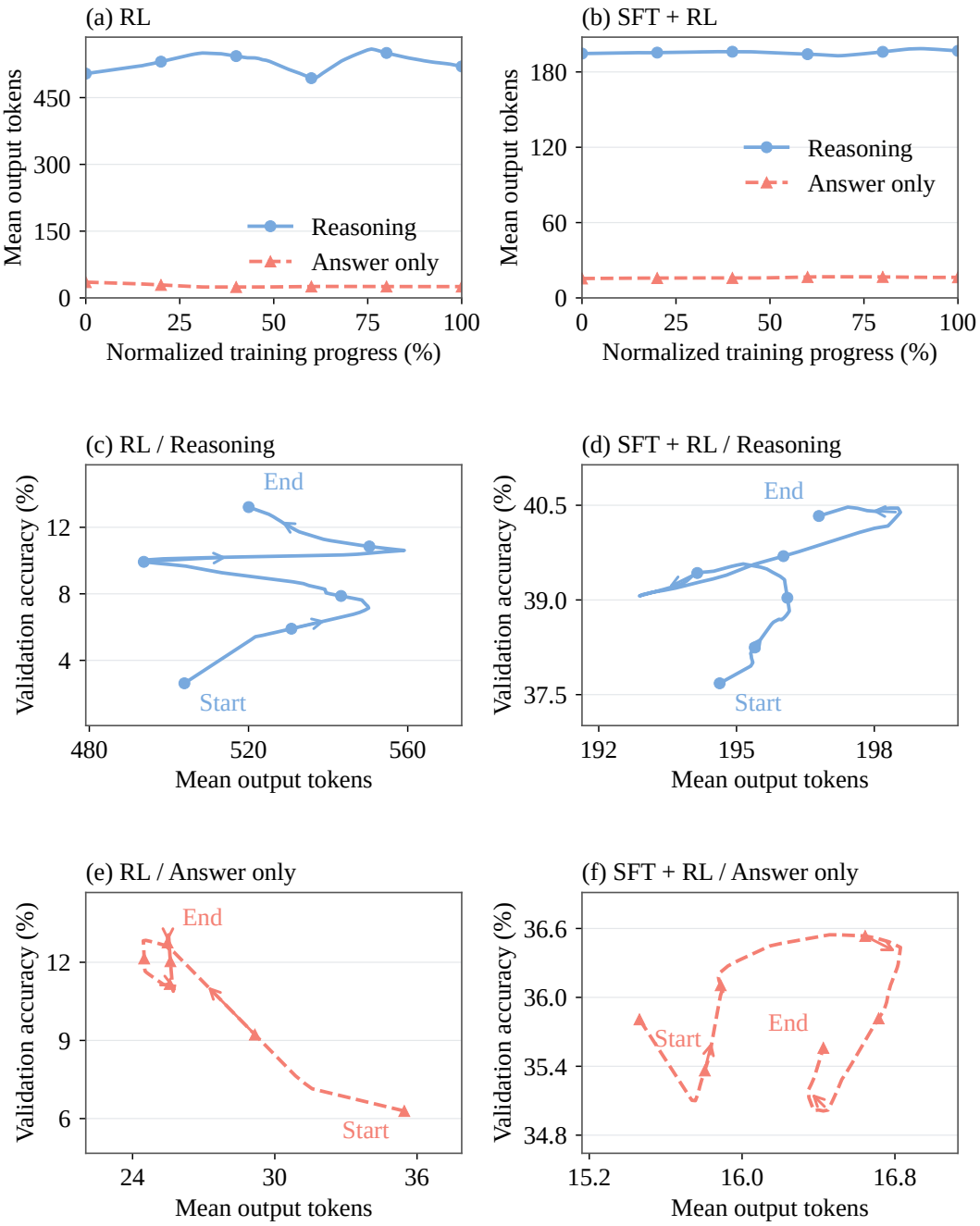


Figure 10: Output length and accuracy during per-task RL. Top: mean output length on fixed validation questions as training progresses. The middle and bottom rows plot accuracy against output length for reasoning and answer-only outputs, respectively. Each trajectory uses its own axis ranges to show its variation. Arrows indicate training direction. Curves average the 15 tasks after aligning normalized training progress. Length counts the complete response and does not establish the quality of its reasoning.

C.8 Transfer across Training Scopes

C.8.1 Cross-Task Transfer

We evaluate each source-task model on the other 14 task test sets. Table 10 pools those questions for each source. Tables 11, 12, and 13 report SFT, RL, and SFT + RL, with test tasks as rows and training tasks as columns. Diagonal entries are excluded. Base uses the same target questions, and arrows indicate numerical changes from Base rather than statistical significance.

Table 10: Cross-task transfer: pooled results (accuracy, %). Each row identifies the training task. Accuracy pools all test questions from the other 14 tasks. Base pools the same target questions using the per-task baselines in Table 2. Arrows indicate higher (↑) or lower (↓) accuracy than Base. Equal values have no marker.

Training task	Base	SFT	RL	SFT + RL
Phase Recognition	1.62	4.05↑	3.89↑	4.49↑
Phase Consistency	2.42	5.94↑	2.26↓	6.02↑
Action Description Matching	2.47	10.43↑	2.67↑	9.00↑
Shared CVS Correction	2.45	5.57↑	4.07↑	5.22↑
Spatial Repair	2.44	5.60↑	3.88↑	5.76↑
Reference Alignment	1.78	1.94↑	3.81↑	2.67↑
Temporal Order Repair	2.50	5.76↑	3.86↑	5.92↑
CVS Milestone Tracking	2.51	3.19↑	2.63↑	2.83↑
State Change Tracking	2.42	5.21↑	3.47↑	5.57↑
Spatial Relation Tracking	2.48	4.47↑	4.02↑	4.47↑
Workflow Ordering	1.90	6.14↑	2.46↑	6.02↑
Workflow Path Recovery	2.48	5.85↑	2.64↑	7.17↑
Workflow Repair	2.51	8.66↑	3.72↑	9.10↑
Description Correction	2.48	7.20↑	3.54↑	6.71↑
Description and Workflow Correction	2.50	8.91↑	3.43↑	8.47↑

Figure 11 compares gains on the source task with pooled gains elsewhere. Spatial Repair and CVS Milestone Tracking show large improvements on their own tasks after supervised initialization, yet much smaller improvements on other tasks. This pattern is consistent with learning task-specific answer rules and response patterns rather than a uniformly transferable surgical-state representation. The transfer scores establish the limited reach of the adaptation, but do not uniquely identify its internal representation. The constant CVS answers documented in Section E.3 give a concrete example of high own-task accuracy without input-dependent tracking.

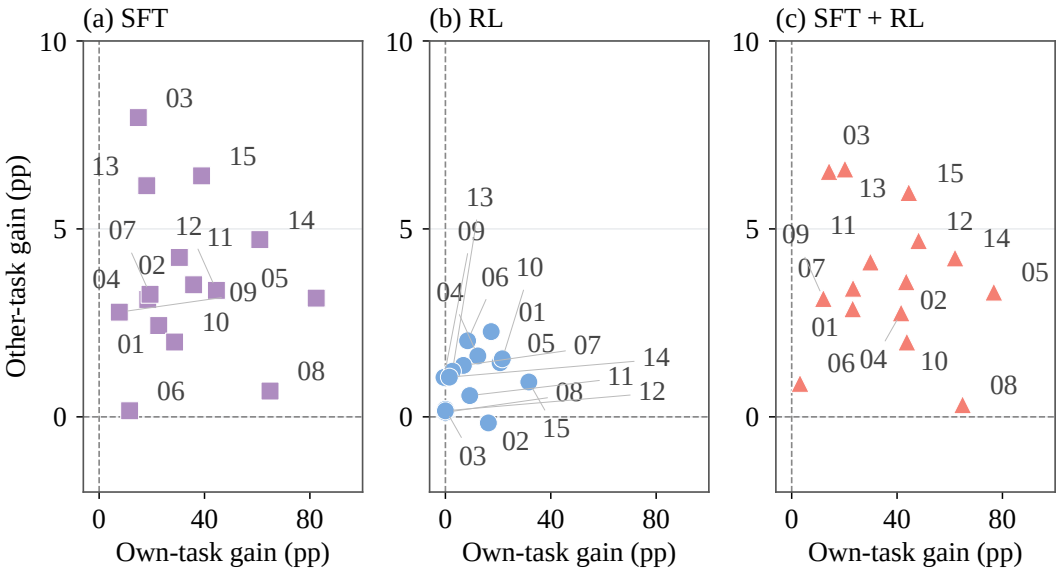


Figure 11: Cross-task transfer relative to Base. Each numbered point is a source task, comparing its own-task accuracy gain with its pooled gain on the other 14 tasks, where pp denotes percentage points. Task numbers use the order in Table 1. Full task names and corresponding scores are given in Tables 11–13.

Table 11: Cross-task transfer after SFT (accuracy, %). Rows identify test tasks. Columns identify training tasks. Task numbers are shared across both axes. Base follows Table 2. Arrows indicate higher (↑) or lower (↓) accuracy than Base for the same test task. Equal values have no marker. The diagonal is excluded.

Test task	Base	Training task														
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01 Phase Recognition	11.58	—	36.32↑	25.79↑	20.53↑	24.74↑	6.32↓	27.89↑	4.21↓	31.58↑	24.74↑	41.58↑	35.79↑	29.47↑	30.53↑	32.63↑
02 Phase Consistency	1.09	22.83↑	—	23.91↑	15.22↑	18.48↑	17.39↑	15.22↑	14.67↑	27.72↑	24.46↑	25.54↑	11.96↑	26.09↑	18.48↑	20.11↑
03 Action Description Matching	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68↑	0.00
04 Shared CVS Correction	0.00	2.31↑	3.85↑	41.54↑	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.46↑	1.54↑	6.92↑
05 Spatial Repair	0.63	0.00↓	2.52↑	0.00↓	0.00↓	—	0.63	0.00↓	0.00↓	0.00↓	0.00↓	0.00↓	0.00↓	0.63	0.00↓	0.00↓
06 Reference Alignment	9.47	0.00↓	7.89↓	22.63↑	16.32↑	8.42↓	—	12.63↑	14.21↑	0.53↓	0.00↓	6.84↓	2.11↓	14.21↑	15.79↑	13.16↑
07 Temporal Order Repair	0.00	5.11↑	7.95↑	0.57↑	3.98↑	3.98↑	0.00	—	2.27↑	4.55↑	3.41↑	6.82↑	1.14↑	2.84↑	7.39↑	7.95↑
08 CVS Milestone Tracking	0.00	0.00	1.08↑	1.08↑	0.00	0.00	0.00	0.00	—	1.08↑	0.00	0.00	0.00	1.62↑	1.08↑	2.16↑
09 State Change Tracking	1.09	0.00↓	0.55↓	0.00↓	0.00↓	1.09	0.00↓	0.00↓	0.00↓	—	0.55↓	0.00↓	0.00↓	0.55↓	1.09	0.55↓
10 Spatial Relation Tracking	0.50	3.02↑	0.00↓	0.00↓	0.00↓	3.02↑	0.00↓	0.00↓	0.00↓	0.50	—	0.00↓	0.00↓	2.01↑	0.00↓	4.02↑
11 Workflow Ordering	8.15	14.67↑	10.87↑	18.48↑	18.48↑	9.24↑	1.63↓	18.48↑	5.98↓	2.17↓	4.35↓	—	27.17↑	15.76↑	15.76↑	15.76↑
12 Workflow Path Recovery	0.00	5.49↑	9.15↑	20.73↑	1.22↑	6.10↑	0.00	2.44↑	0.00	0.61↑	0.61↑	0.00	—	1.83↑	3.66↑	10.37↑
13 Workflow Repair	0.00	1.60↑	1.06↑	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.06↑	0.00	0.00	—	0.00	2.66↑
14 Description Correction	0.50	0.50	0.00↓	0.50	0.00↓	0.50	0.00↓	0.00↓	0.00↓	0.50	0.00↓	0.00↓	0.00↓	0.00↓	—	5.00↑
15 Description and Workflow Correction	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.11↑	0.00	0.00	0.56↑	0.00	0.00	0.00	—

Table 12: Cross-task transfer after RL (accuracy, %). Rows identify test tasks. Columns identify training tasks. Task numbers are shared across both axes. Base follows Table 2. Arrows indicate higher (↑) or lower (↓) accuracy than Base for the same test task. Equal values have no marker. The diagonal is excluded.

Test task	Base	Training task														
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01 Phase Recognition	11.58	—	20.00↑	11.05↓	18.42↑	13.68↑	7.89↓	10.53↓	14.74↑	14.74↑	4.21↓	7.37↓	11.05↓	25.79↑	11.05↓	1.05↓
02 Phase Consistency	1.09	15.22↑	—	1.09	9.78↑	1.09	0.54↓	1.09	1.09	0.00↓	0.54↓	0.54↓	1.09	19.02↑	3.80↑	3.26↑
03 Action Description Matching	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68↑
04 Shared CVS Correction	0.00	3.08↑	2.31↑	0.00	—	5.38↑	3.85↑	1.54↑	0.00	0.00	2.31↑	3.85↑	0.00	0.00	4.62↑	1.54↑
05 Spatial Repair	0.63	5.03↑	1.26↑	1.26↑	0.00↓	—	2.52↑	2.52↑	0.63	3.14↑	1.89↑	1.89↑	1.26↑	1.89↑	3.14↑	0.63
06 Reference Alignment	9.47	13.16↑	2.11↓	11.58↑	13.68↑	13.16↑	—	15.26↑	12.11↑	14.74↑	17.89↑	16.84↑	11.05↑	0.00↓	11.58↑	14.74↑
07 Temporal Order Repair	0.00	1.14↑	0.57↑	0.00	3.98↑	0.57↑	6.25↑	—	0.00	2.27↑	7.39↑	0.57↑	0.00	0.00	4.55↑	6.82↑
08 CVS Milestone Tracking	0.00	10.27↑	0.00	0.00	0.00	1.62↑	8.65↑	1.62↑	—	0.00	1.62↑	0.00	0.00	0.00	1.08↑	0.54↑
09 State Change Tracking	1.09	1.09	1.64↑	1.64↑	0.00↓	2.19↑	1.09	2.19↑	1.09	—	1.64↑	0.00↓	2.73↑	1.09	1.64↑	0.55↓
10 Spatial Relation Tracking	0.50	1.51↑	0.50	0.50	2.51↑	4.52↑	3.52↑	0.00↓	0.50	0.50	—	1.51↑	0.50	1.51↑	1.01↑	0.00↓
11 Workflow Ordering	8.15	2.17↓	0.00↓	8.15	6.52↓	8.70↑	14.67↑	16.85↑	4.35↓	10.87↑	14.13↑	—	7.07↓	0.00↓	5.43↓	15.76↑
12 Workflow Path Recovery	0.00	0.61↑	0.00	0.00	0.00	0.61↑	0.61↑	0.61↑	0.00	0.00	1.22↑	0.61↑	—	0.00	0.61↑	0.61↑
13 Workflow Repair	0.00	0.00	0.53↑	0.00	0.00	0.53↑	1.06↑	0.00	0.00	0.00	1.60↑	0.53↑	0.00	—	0.00	0.53↑
14 Description Correction	0.50	0.00↓	1.50↑	0.50	0.00↓	1.00↑	1.50↑	0.00↓	0.00↓	0.00↓	0.00↓	0.00↓	0.50	0.00↓	—	0.00↓
15 Description and Workflow Correction	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	—

Table 13: Cross-task transfer after SFT + RL (accuracy, %). Rows identify test tasks. Columns identify training tasks. Task numbers are shared across both axes. Base follows Table 2. Arrows indicate higher (↑) or lower (↓) accuracy than Base for the same test task. Equal values have no marker. The diagonal is excluded.

Test task	Base	Training task														
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01 Phase Recognition	11.58	—	37.37↑	26.84↑	21.58↑	26.32↑	7.89↓	25.79↑	3.68↓	34.21↑	24.74↑	36.84↑	40.53↑	31.05↑	28.95↑	33.16↑
02 Phase Consistency	1.09	22.28↑	—	21.74↑	14.13↑	18.48↑	18.48↑	15.22↑	9.78↑	29.35↑	20.11↑	26.09↑	15.22↑	25.00↑	19.57↑	19.02↑
03 Action Description Matching	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.03↑	1.35↑
04 Shared CVS Correction	0.00	2.31↑	6.92↑	15.38↑	—	0.77↑	0.00	0.00	0.00	0.77↑	0.77↑	0.00	0.00	34.62↑	3.08↑	5.38↑
05 Spatial Repair	0.63	0.00↓	2.52↑	0.00↓	0.00↓	—	0.00↓	0.00↓	0.00↓	0.00↓	0.00↓	0.00↓	0.00↓	0.63	0.00↓	0.63
06 Reference Alignment	9.47	0.00↓	8.95↓	24.21↑	14.21↑	7.89↓	—	11.58↑	16.32↑	1.58↓	1.05↓	8.95↓	6.32↓	13.68↑	12.63↑	15.26↑
07 Temporal Order Repair	0.00	6.82↑	7.39↑	5.11↑	2.84↑	6.25↑	2.27↑	—	2.27↑	5.11↑	2.84↑	7.39↑	2.84↑	3.41↑	3.98↑	8.52↑
08 CVS Milestone Tracking	0.00	2.16↑	0.54↑	1.08↑	0.00	0.00	0.00	0.54↑	—	0.54↑	0.00	0.00	0.00	0.54↑	1.08↑	4.32↑
09 State Change Tracking	1.09	0.00↓	0.00↓	0.00↓	0.00↓	1.09	0.00↓	0.00↓	0.00↓	—	1.09	0.00↓	0.00↓	1.09	0.00↓	0.55↓
10 Spatial Relation Tracking	0.50	7.54↑	0.00↓	0.00↓	0.00↓	0.50	0.00↓	0.00↓	0.00↓	0.00↓	—	0.00↓	0.00↓	1.51↑	0.00↓	3.02↑
11 Workflow Ordering	8.15	13.59↑	10.87↑	17.93↑	17.39↑	9.24↑	7.07↓	21.74↑	4.89↓	1.63↓	6.52↓	—	30.98↑	15.22↑	15.22↑	15.76↑
12 Workflow Path Recovery	0.00	4.27↑	4.27↑	14.63↑	0.61↑	7.32↑	0.00	1.83↑	0.00	0.61↑	1.22↑	0.00	—	4.88↑	3.66↑	2.44↑
13 Workflow Repair	0.00	2.13↑	3.72↑	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.06↑	0.00	0.00	—	0.00	2.13↑
14 Description Correction	0.50	0.00↓	0.00↓	0.50	0.00↓	0.50	0.00↓	0.50	0.00↓	0.50	0.00↓	0.00↓	0.00↓	0.00↓	—	3.00↑
15 Description and Workflow Correction	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.67↑	0.56↑	0.00	0.00	0.56↑	0.00	0.00	0.00	—

C.8.2 CATEGORY-LEVEL TRAINING

Category models train and validate on the combined member-task partitions and are tested on all tasks. Categories A–F contain tasks 01–02, 03–05, 06–10, 11–13, 14, and 15. Categories E and F reuse their per-task models. Table 14 separates accuracy within and outside the training category. Tables 15, 16, and 17 give the corresponding task-level results.

Table 14: Category training: within and outside the trained category (accuracy, %). Each cell reports within-category / outside-category accuracy, pooling test questions separately for each group. Base uses the per-task baselines in Table 2 on the same questions. SFT reporting follows Section C.4. Each arrow compares a value with its corresponding within- or outside-category Base. Equal values have no marker.

Training category	Tasks	Base within / outside	SFT within / outside	RL within / outside	SFT + RL within / outside
A	01 Phase Recognition 02 Phase Consistency	6.42 / 1.66	45.19↑ / 2.89↑	31.28↑ / 2.71↑	50.27↑ / 3.15↑
B	03 Action Description Matching 04 Shared CVS Correction 05 Spatial Repair	0.23 / 2.74	47.37↑ / 12.55↑	15.56↑ / 3.10↑	50.57↑ / 11.65↑
C	06 Reference Alignment 07 Temporal Order Repair 08 CVS Milestone Tracking 09 State Change Tracking 10 Spatial Relation Tracking	2.25 / 2.37	29.37↑ / 9.21↑	22.83↑ / 2.78↑	30.44↑ / 9.26↑
D	11 Workflow Ordering 12 Workflow Path Recovery 13 Workflow Repair	2.80 / 2.21	57.46↑ / 7.63↑	7.46↑ / 7.53↑	56.53↑ / 8.90↑
E	14 Description Correction	0.50 / 2.48	61.50↑ / 7.24↑	2.00↑ / 3.66↑	62.50↑ / 6.87↑
F	15 Description and Workflow Correction	0.00 / 2.50	38.89↑ / 8.75↑	31.67↑ / 3.47↑	44.44↑ / 8.47↑

The learning histories in Figure 12 vary by category. In Description Correction, direct RL produces longer responses while validation accuracy remains low. Category validation curves describe learning on the supervised tasks. The outside-category test results measure transfer.

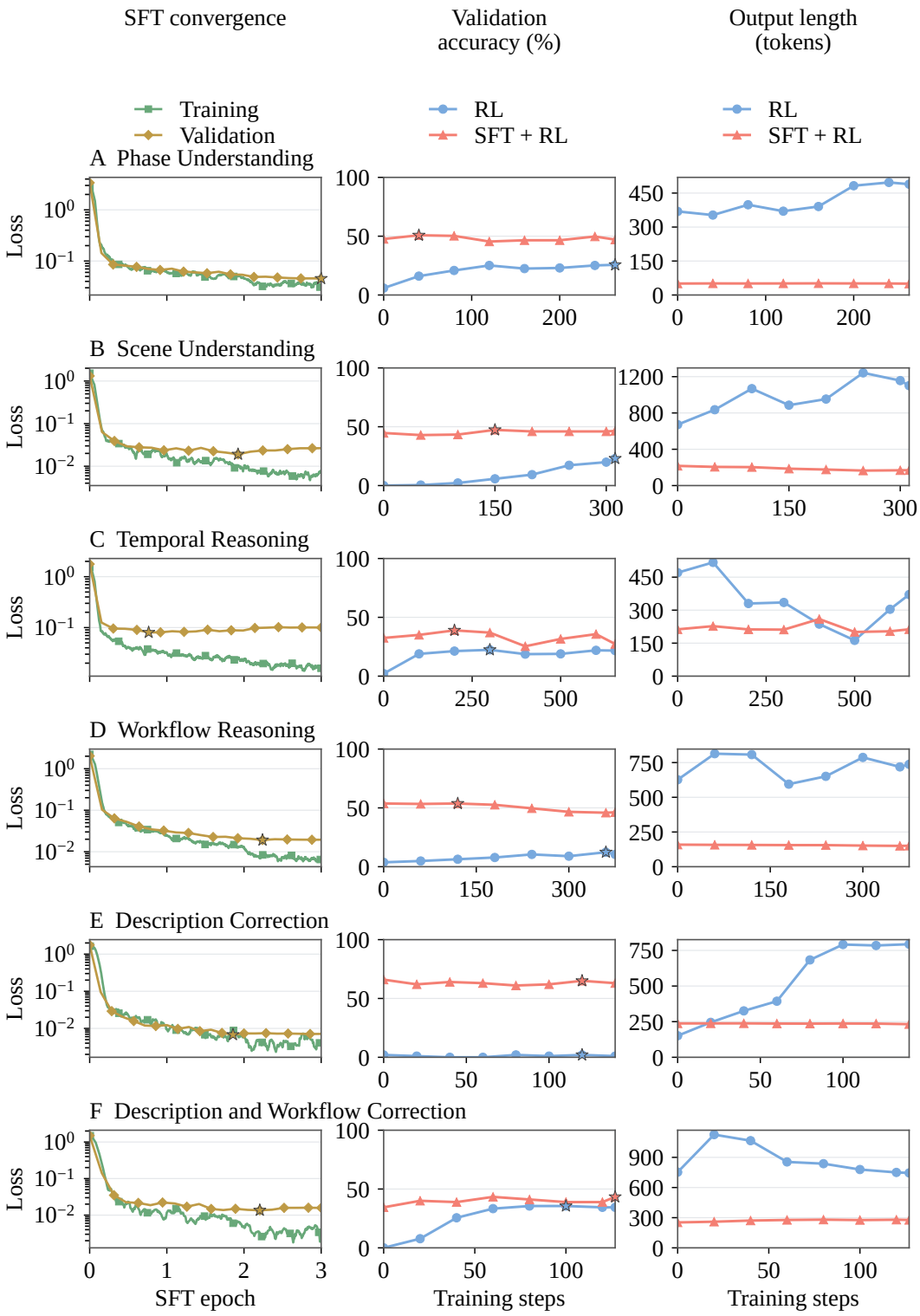


Figure 12: Category-level learning dynamics in reasoning mode. Rows are categories A–F. Columns show SFT loss, RL validation accuracy, and mean validation output length. Squares and diamonds denote training and validation loss. Circles and triangles denote RL and SFT + RL. Accuracy and length pool the category's fixed validation questions. Stars mark selected checkpoints. Categories E and F reuse the Description Correction and Description and Workflow Correction runs, respectively. They are not additional training repetitions.

Table 15: Category transfer after SFT (accuracy, %). Rows group test tasks by their category. Columns A–F identify training categories. Results include both within-category and outside-category tests. Base follows Table 2. SFT reporting follows Section C.4. Arrows indicate higher (↑) or lower (↓) accuracy than Base for the same test task. Equal values have no marker.

Category	Test task	Base	Training category					
			A	B	C	D	E	F
A	01 Phase Recognition	11.58	52.11↑	31.05↑	33.16↑	44.21↑	31.58↑	32.63↑
	02 Phase Consistency	1.09	38.04↑	23.91↑	28.26↑	19.02↑	19.02↑	19.57↑
B	03 Action Description Matching	0.00	0.00	20.95↑	0.00	0.00	0.68↑	0.00
	04 Shared CVS Correction	0.00	2.31↑	18.46↑	0.00	3.85↑	0.77↑	7.69↑
	05 Spatial Repair	0.63	0.63	95.60↑	0.63	0.63	0.00↓	0.63
C	06 Reference Alignment	9.47	6.32↓	16.32↑	35.26↑	8.42↓	15.26↑	13.16↑
	07 Temporal Order Repair	0.00	5.11↑	8.52↑	11.93↑	6.82↑	6.25↑	7.39↑
	08 CVS Milestone Tracking	0.00	1.08↑	30.27↑	64.86↑	0.54↑	1.62↑	1.62↑
	09 State Change Tracking	1.09	0.00↓	1.09	3.83↑	1.09	1.09	0.00↓
	10 Spatial Relation Tracking	0.50	0.50	9.55↑	29.65↑	3.02↑	0.00↓	4.02↑
D	11 Workflow Ordering	8.15	15.22↑	19.57↑	18.48↑	59.24↑	15.76↑	15.22↑
	12 Workflow Path Recovery	0.00	4.27↑	9.15↑	3.05↑	68.29↑	4.27↑	9.76↑
	13 Workflow Repair	0.00	0.53↑	0.00	1.06↑	46.28↑	0.00	2.66↑
E	14 Description Correction	0.50	1.00↑	1.00↑	1.00↑	0.00↓	61.50↑	5.00↑
F	15 Description and Workflow Correction	0.00	0.00	0.00	0.00	0.00	0.00	38.89↑

Table 16: Category transfer after RL (accuracy, %). Rows group test tasks by their category. Columns A–F identify training categories. Results include both within-category and outside-category tests. Base follows Table 2. Arrows indicate higher (↑) or lower (↓) accuracy than Base for the same test task. Equal values have no marker.

Category	Test task	Base	Training category					
			A	B	C	D	E	F
A	01 Phase Recognition	11.58	36.84↑	12.11↑	1.05↓	31.58↑	11.05↓	1.58↓
	02 Phase Consistency	1.09	25.54↑	9.78↑	1.09	13.59↑	6.52↑	2.17↑
B	03 Action Description Matching	0.00	0.68↑	0.00	0.00	0.00	0.00	0.68↑
	04 Shared CVS Correction	0.00	3.85↑	18.46↑	0.00	19.23↑	3.08↑	1.54↑
	05 Spatial Repair	0.63	3.14↑	27.67↑	0.63	1.26↑	3.14↑	0.63
C	06 Reference Alignment	9.47	10.00↑	6.32↓	16.84↑	8.95↓	12.63↑	15.79↑
	07 Temporal Order Repair	0.00	0.57↑	0.00	8.52↑	2.27↑	3.41↑	6.82↑
	08 CVS Milestone Tracking	0.00	0.00	0.54↑	64.86↑	1.08↑	0.54↑	1.08↑
	09 State Change Tracking	1.09	1.09	3.28↑	1.09	1.09	1.09	0.55↓
	10 Spatial Relation Tracking	0.50	3.02↑	3.02↑	22.11↑	10.05↑	0.50	0.00↓
D	11 Workflow Ordering	8.15	9.24↑	0.54↓	15.22↑	8.15	6.52↓	15.76↑
	12 Workflow Path Recovery	0.00	0.00	0.00	1.22↑	6.10↑	0.61↑	0.61↑
	13 Workflow Repair	0.00	1.60↑	0.00	4.26↑	7.98↑	0.53↑	0.00
E	14 Description Correction	0.50	1.50↑	1.00↑	2.50↑	1.50↑	2.00↑	0.00↓
F	15 Description and Workflow Correction	0.00	0.00	0.00	0.00	0.00	0.00	31.67↑

Table 17: Category transfer after SFT + RL (accuracy, %). Rows group test tasks by their category. Columns A–F identify training categories. Results include both within-category and outside-category tests. Base follows Table 2. Arrows indicate higher (↑) or lower (↓) accuracy than Base for the same test task. Equal values have no marker.

Category	Test task	Base	Training category					
			A	B	C	D	E	F
A	01 Phase Recognition	11.58	57.37↑	37.37↑	30.00↑	47.89↑	28.95↑	33.16↑
	02 Phase Consistency	1.09	42.93↑	28.26↑	30.98↑	22.28↑	19.57↑	18.48↑
B	03 Action Description Matching	0.00	0.00	27.70↑	1.35↑	0.68↑	2.03↑	0.68↑
	04 Shared CVS Correction	0.00	3.08↑	41.54↑	0.00	9.23↑	1.54↑	6.15↑
	05 Spatial Repair	0.63	1.26↑	79.25↑	0.00↓	0.00↓	0.00↓	0.63
C	06 Reference Alignment	9.47	7.37↓	13.16↑	34.21↑	8.95↓	13.16↑	15.26↑
	07 Temporal Order Repair	0.00	6.25↑	5.11↑	12.50↑	9.09↑	3.98↑	7.39↑
	08 CVS Milestone Tracking	0.00	0.54↑	9.73↑	64.86↑	1.62↑	1.08↑	4.32↑
	09 State Change Tracking	1.09	0.55↓	1.64↑	8.20↑	1.09	0.55↓	0.55↓
	10 Spatial Relation Tracking	0.50	0.50	18.59↑	31.16↑	3.02↑	0.50	3.02↑
D	11 Workflow Ordering	8.15	14.13↑	20.65↑	19.57↑	58.70↑	17.39↑	16.30↑
	12 Workflow Path Recovery	0.00	4.27↑	3.66↑	2.44↑	70.73↑	3.05↑	3.66↑
	13 Workflow Repair	0.00	0.53↑	0.00	1.06↑	42.02↑	0.00	2.13↑
E	14 Description Correction	0.50	2.00↑	0.00↓	1.00↑	0.00↓	62.50↑	3.00↑
F	15 Description and Workflow Correction	0.00	0.00	0.00	0.00	0.00	0.00	44.44↑

C.8.3 JOINT TRAINING

Joint models train on all task partitions and validate on their combined validation questions. Table 18 reports each task and the pooled test accuracy. Figure 13 shows learning curves and comparisons with per-task models. In its lower panels, positive bars favor category or joint training. Negative bars favor the per-task model under the same method.

Table 18: Joint training on all tasks (accuracy, %). Each method trains one model on the combined training sets and evaluates its validation-selected best checkpoint on each task's test set. Base follows Table 2. SFT reporting follows Section C.4. Overall pools all test questions. Arrows indicate higher (↑) or lower (↓) accuracy than the corresponding Base. Equal values have no marker.

Category	Test task	Base	Joint SFT	Joint RL	Joint SFT + RL
A	01 Phase Recognition	11.58	66.84↑	21.05↑	70.00↑
	02 Phase Consistency	1.09	51.63↑	32.61↑	64.13↑
B	03 Action Description Matching	0.00	19.59↑	0.00	21.62↑
	04 Shared CVS Correction	0.00	18.46↑	2.31↑	9.23↑
	05 Spatial Repair	0.63	80.50↑	0.00↓	87.42↑
C	06 Reference Alignment	9.47	46.32↑	16.84↑	47.37↑
	07 Temporal Order Repair	0.00	11.36↑	9.09↑	15.91↑
	08 CVS Milestone Tracking	0.00	64.86↑	64.86↑	64.32↑
	09 State Change Tracking	1.09	8.20↑	2.73↑	10.38↑
	10 Spatial Relation Tracking	0.50	45.23↑	44.72↑	56.78↑
D	11 Workflow Ordering	8.15	55.98↑	17.93↑	61.41↑
	12 Workflow Path Recovery	0.00	63.41↑	0.61↑	59.76↑
	13 Workflow Repair	0.00	36.70↑	13.30↑	29.79↑
E	14 Description Correction	0.50	61.00↑	0.00↓	63.00↑
F	15 Description and Workflow Correction	0.00	46.11↑	3.33↑	48.33↑
	Overall	2.33	45.75↑	16.17↑	48.23↑

The task breakdown includes regressions, particularly for Shared CVS Correction. The best validation checkpoint for joint SFT + RL occurs before the end of RL, followed by declining validation accuracy. Broader training changes data volume and possible cross-task image exposure, so these results describe the tested training scopes rather than an isolated effect of combining tasks.

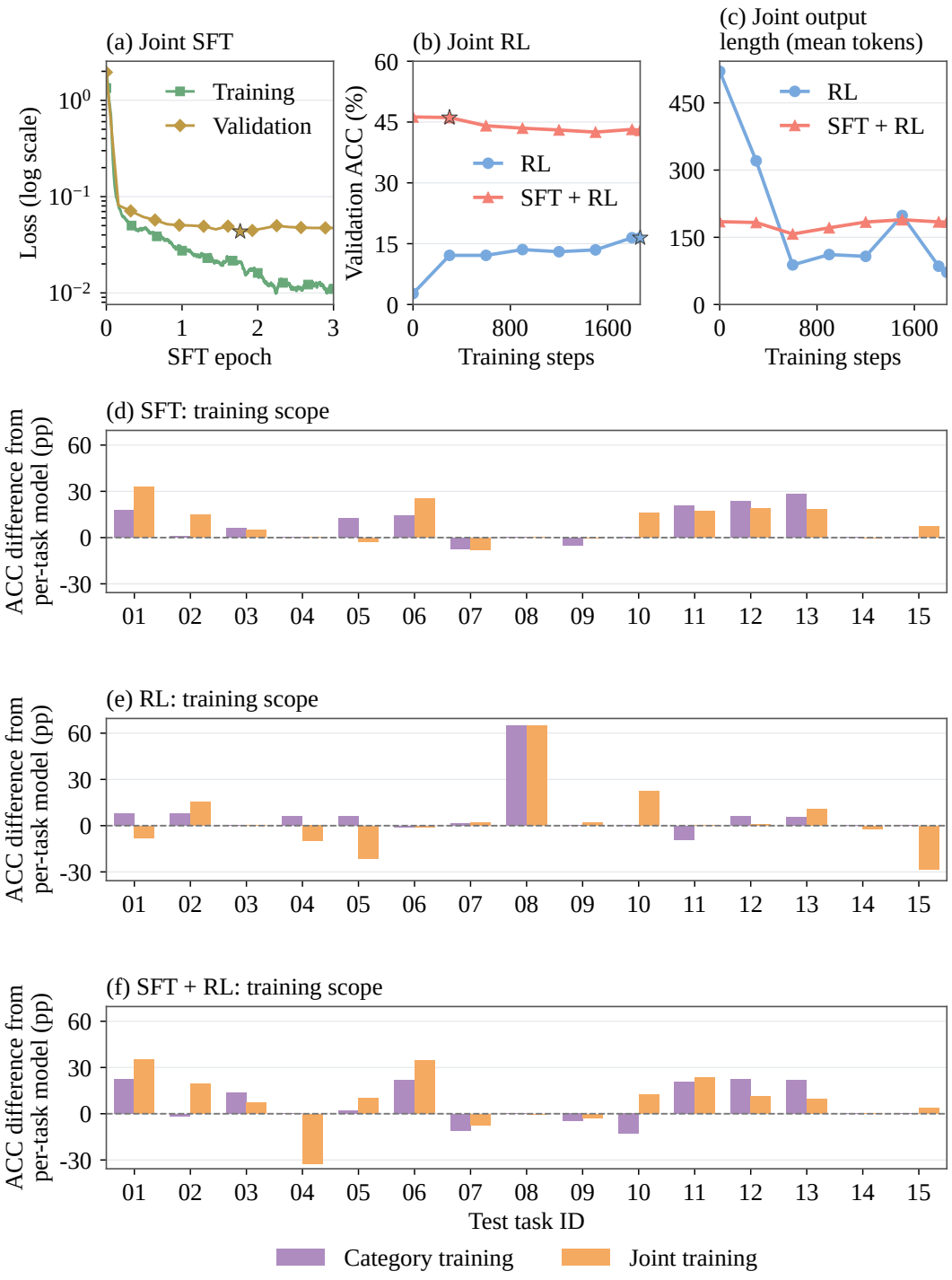


Figure 13: Joint learning dynamics and the effect of training scope. Top: joint SFT loss, RL validation accuracy, and validation output length. Validation metrics pool all validation questions and stars mark selected checkpoints. Panels (d)–(f) compare SFT, RL, and SFT + RL models, respectively. Each bar subtracts the per-task-trained model's accuracy from the category-trained model's accuracy (purple) or jointly trained model's accuracy (orange), using the same method and test questions. Values are percentage-point differences between trained models, not before-and-after training gains. Zero denotes the per-task-trained baseline. Task identifiers (IDs) follow Figure 11 and Table 18. Training scopes also differ in data volume and cross-task image overlap, so these comparisons do not isolate a causal benefit of joint training.

Table 19: Accuracy (%) aggregated by GATE capability group.

Capability	Random	GPT-6 Astra	GPT-6 Luna	GPT-5.6 Sol	GPT-5.6 Luna	Claude Opus 5.5	Claude Sonnet 5	Kimi K3	GLM-5.3 Flash	Intern VL3.5	Nemotron-3 Omni	Gemma 4 31B IT
Phase Understanding	21.69	62.82	35.52	50.69	40.38	57.53	34.72	41.35	47.81	34.29	27.88	45.03
Scene Understanding	8.26	47.79	24.13	40.57	29.00	38.23	17.17	33.46	29.08	20.71	4.26	25.24
Temporal Reasoning	6.88	54.26	26.77	35.61	24.42	45.17	19.68	25.04	23.84	10.91	6.75	23.73
Workflow Reasoning	8.26	43.76	14.24	24.85	14.24	35.69	11.81	17.38	18.54	9.83	7.88	15.70
Description Correction	0.50	29.36	23.35	23.25	22.14	37.27	23.75	23.25	20.84	13.03	0.50	24.75
Description and Workflow Correction	1.44	33.96	11.43	17.76	21.98	45.17	7.55	27.30	16.87	18.31	10.10	14.98

Table 20: Accuracy and output diagnostics (%). Strict format requires the specified opening and closing markers and an answer structure that follows the task's rules.

Metric	GPT-6 Astra	GPT-6 Luna	GPT-5.6 Sol	GPT-5.6 Luna	Claude Opus 5.5	Claude Sonnet 5	Kimi K3	GLM-5.3 Flash	Intern VL3.5	Nemotron-3 Omni	Gemma 4 31B IT
Accuracy	49.00	23.76	34.44	25.15	43.11	19.22	27.39	26.39	16.39	9.15	24.86
Task-valid answer	99.44	97.59	99.53	97.18	99.00	87.98	99.42	87.25	93.34	61.87	89.54
Strict format	99.44	97.59	99.53	97.18	0.00	0.00	99.38	27.92	88.10	32.64	0.00
Valid, other format	0.00	0.00	0.00	0.00	99.00	87.98	0.04	59.34	5.24	29.24	89.54
No valid answer	0.56	2.41	0.47	2.82	1.00	12.02	0.58	12.75	6.66	38.13	10.46

D Additional Benchmark Results

D.1 Results by Capability and Anatomical Site

The 11 models answer the same 13,715 questions. Table 19 pools accuracy by capability, and Table 20 summarizes output status.

Strong local judgments do not ensure accurate temporal comparisons. GPT-6 Astra reaches 49.00% overall, but its task scores range from 88.52% on Spatial Repair and 72.84% on CVS Milestone Tracking to 9.96% on State Change Tracking and 33.96% on Description and Workflow Correction. GPT-5.6 Sol similarly reaches 89.66% on Spatial Repair but 6.56% on State Change Tracking. These gaps identify different capability requirements, although tasks also differ in answer structure and source coverage.

Answer validity and correctness reveal different weaknesses. Gemma 4 31B IT obtains 24.86% accuracy despite consistently using a response format that differs from the requested one. InternVL3.5 reaches 16.39% overall, with 31.91% on Shared CVS Correction and 4.92% on State Change Tracking. Producing a usable response therefore does not ensure a correct surgical judgment.

Table 21 reports results by anatomical site. Gallbladder combines Cholec80, CholecT50, and Endoscapes2023. The other rows represent the remaining sites. Different sites contribute different tasks and answer structures, so these scores describe the benchmark composition rather than controlled differences in organ difficulty.

D.2 Text-Only Evaluation

We sample 1,370 questions at approximately 10% within each task using seed 20260920. Text-only prompts retain the question, options, claims, rules, and reasoning request, but remove images and the text marking where they would appear. Table 22 reports all 11 models. The matched visual comparison uses exactly the same questions and scoring rules. Full-benchmark visual scores have a different denominator.

Across the 15,070 model–question pairs, 4,112 answers are correct with images and 2,162 without them (Table 23).

Joint auditing remains vulnerable even when images are available. Shared CVS Correction, CVS Milestone Tracking, and Description and Workflow Correction are the task-level exceptions to improved accuracy with images. Description and Workflow Correction is the only category with lower visual accuracy: 224 of 990 visual responses are correct, compared with 289 text-only responses. Textual constraints and answer preferences can support correct answers without images,

Table 21: VLM performance across six anatomical sites. Values are sample-level accuracies (%).

Anatomical site	N	GPT-6 Astra	GPT-6 Luna	GPT-5.6 Sol	GPT-5.6 Luna	Claude Opus 5.5	Claude Sonnet 5	Kimi K3	GLM-5.3 Flash	Intern VL3.5	Nemotron-3 Omni	Gemma 4 31B IT
Gallbladder	6,508	**45.37**	17.92	29.98	17.89	39.64	18.45	20.48	23.37	16.49	7.76	21.60
Eye	2,221	**64.84**	29.81	43.09	36.33	50.38	18.51	37.55	28.01	16.34	9.59	27.24
Prostate	2,081	**49.83**	29.99	35.22	29.26	44.40	20.57	29.17	28.69	15.62	10.33	25.08
Gynecologic/Pelvic	996	53.01	37.15	48.39	42.27	**56.53**	22.69	48.29	42.07	11.24	6.22	39.96
Lung	630	**49.52**	25.40	36.35	28.89	47.78	18.25	26.19	27.78	18.89	16.19	25.71
Liver	1,279	**35.26**	21.66	29.01	20.88	33.31	19.94	26.27	22.36	20.02	12.35	24.71
Overall	**13,715**	**49.00**	23.76	34.44	25.15	43.11	19.22	27.39	26.39	16.39	9.15	24.86

Note: Bold values indicate the best-performing model in each row. Model names follow Table 1. Gallbladder aggregates Cholec80, CholecT50, and Endoscapes2023. GynSurg covers the Gynecologic/Pelvic site.

Table 22: Text-only VLM accuracy on 1,370 fixed questions, sampled at approximately 10% within each of the 15 GATE tasks and grouped as in Table 1. Values are accuracy (%). Model average is the mean across 11 models.

Model	**Phase Understanding**		**Scene Understanding**			**Temporal Reasoning**					**Workflow Reasoning**			**Description Correction**	**Description and Workflow Correction**
	Phase Recognition	Phase Consistency	Action Matching	CVS Correction	Spatial Repair	Reference Alignment	Temporal Repair	CVS Tracking	State Tracking	Spatial Relation	Workflow Ordering	Workflow Recovery	Workflow Repair	Description Correction	Description and Workflow Correction
Proprietary Models															
GPT-6 Astra	28.42	17.39	0.00	7.14	0.00	22.11	7.95	67.39	0.00	0.00	17.39	23.17	0.00	2.00	44.44
GPT-6 Luna	29.47	20.65	0.00	26.53	24.05	18.95	7.95	66.30	0.00	0.00	17.39	20.73	2.13	6.00	41.11
GPT-5.6 Sol	26.32	18.48	0.00	30.61	32.91	15.79	7.95	65.22	0.00	1.01	18.48	12.20	23.40	3.00	50.00
GPT-5.6 Luna	20.00	19.57	0.00	29.59	21.52	22.11	10.23	64.13	1.10	4.04	15.22	2.44	25.53	5.00	34.44
Claude Opus 5.5	29.47	18.48	0.00	13.27	24.05	21.05	10.23	5.43	0.00	7.07	17.39	3.66	28.72	11.00	45.56
Claude Sonnet 5	25.26	13.04	0.00	26.53	0.00	11.58	7.95	14.13	0.00	1.01	18.48	1.22	9.57	8.00	42.22
Open-Weight models															
Kimi K3	24.21	21.74	0.00	13.27	13.92	16.84	13.64	2.17	0.00	2.02	14.13	2.44	18.09	4.00	22.22
GLM-5.3-Flash	22.11	22.83	0.00	22.45	22.78	12.63	7.95	66.30	0.00	0.00	8.70	1.22	7.45	5.00	14.44
InternVL3.5	30.53	17.39	0.00	30.61	20.25	8.42	9.09	9.78	0.00	4.04	16.30	1.22	8.51	6.00	3.33
Nemotron-3 Omni	22.11	17.39	0.00	13.27	8.86	18.95	11.36	25.00	0.00	0.00	19.57	1.22	2.13	0.00	11.11
Gemma 4 31B IT	25.26	17.39	0.00	11.22	20.25	8.42	3.41	2.17	0.00	6.06	20.65	0.00	6.38	10.00	12.22
Model average	25.74	18.58	0.00	20.41	17.15	16.08	8.88	35.28	0.10	2.30	16.70	6.32	11.99	5.45	29.19

Note: Images and the text marking where they would appear were removed. Per-task sample sizes are 79–100. Full-pool visual scores in Table 1 use different denominators and are not matched ablation comparisons.

while visual access introduces clips that the model must interpret and integrate. This aggregate comparison does not identify the cause of individual errors.

E Reasoning and Error Analysis

We examine final-answer outcomes, visible reasoning, and changes after training. These analyses describe recorded responses. They do not independently verify the models' visual statements or reveal their internal reasoning process.

E.1 Benchmark Evaluation

We analyze 150,865 responses from 11 models on 13,715 image-based questions. Figure 4 partitions them into correct answers (40,990, 27.2%), scorable wrong answers (97,826, 64.8%), nonempty unscorable outputs (10,854, 7.2%), and empty outputs (1,195, 0.8%). The absence of final text

Table 23: Visual and text-only accuracy on the same 1,370 questions for all 11 models. Each count is a model–question pair. Category accuracies pool correct answers across models and tasks. Difference is visual minus text-only accuracy in percentage points.

Task category	Pairs	Visual (%)	Text only (%)	Difference
Phase Understanding	2,057	42.1	22.2	+19.9
Scene Understanding	2,860	26.9	12.9	+14.0
Temporal Reasoning	5,115	26.5	12.5	+14.1
Workflow Reasoning	2,948	21.6	11.9	+9.7
Description Correction	1,100	23.5	5.5	+18.1
Description and Workflow Correction	990	22.6	29.2	−6.6
Overall	15,070	27.29	14.35	+12.94

Table 24: Visible-reasoning support among sampled correct answers. Counts describe the reviewed sample, not all correct answers. "Clear mismatch" includes a non sequitur or a direct contradiction. Only full-support and clear-mismatch counts are shown. Reviewed totals also include partial-support and indeterminate cases.

Model	Reviewed	Full support	Clear mismatch
Claude Opus 5.5	300	297	0
Claude Sonnet 5	300	276	1
Gemma 4 31B IT	300	274	2
GLM-5.3 Flash	300	270	2
GPT-5.6 Luna	300	269	3
GPT-5.6 Sol	300	262	0
GPT-6 Astra	300	289	1
GPT-6 Luna	297	243	5
InternVL3.5	300	227	10
Kimi K3	300	274	1
Nemotron-3 Omni	256	146	28
Total	3,253	2,827	53

takes precedence when assigning groups. Pooled rates weight responses equally. Model-mean rates weight models equally.

E.1.1 Correct Answers

We sample up to 20 correct responses from each of 165 model–task cells, reviewing smaller cells in full. Table 24 reports the resulting 3,253 responses.

A correct answer does not always follow from its visible reasoning. The explanation fully supports the answer in 2,827 responses (86.9%), partly supports it in 341 (10.5%), and clearly conflicts with or cannot yield it in 53 (1.6%). The remaining 32 are indeterminate. Full support ranges from 57.0% for Nemotron-3 Omni to 99.0% for Claude Opus 5.5. These differences concern the visible text, whose length and level of detail also vary across models.

The sample is stratified by model and task. Weighting each cell by its full correct-answer count gives 91.38% full support, compared with 86.9% in the unweighted reviewed sample. An independent reviewer assessed 90 randomly selected responses, with 78 initial exact agreements before adjudication. Neither measure independently checks the visual evidence.

E.1.2 Scorable Wrong Answers

We examine 440 wrong responses from eight temporal and workflow tasks: five responses from each of 88 model–task cells, selected in a fixed, reproducible order. These tasks contain 55,396 scorable wrong responses in total. Review follows the visible path from task interpretation through single-clip judgments, cross-clip comparisons, rule application, and the final answer. Full text was read for 146 responses. Opening and decisive passages were inspected for the remainder. Source frames were not re-examined.

Table 25: Earliest divergence supported by the examined text in the 440 sampled scorable wrong answers. Categories describe visible evidence, not mutually exclusive internal failure mechanisms.

Observed divergence	Responses
Task condition or criterion	23
Single-clip or attribute judgment	283
Cross-clip relation, pairing, or order	116
Rule application or final selection	9
Written decision versus submitted answer	2
Not localizable	6
Reference ambiguity	1
Total	440

Table 26: Phase selection in the scorable PmLR50 Phase Recognition responses. Scorable gives the number of scorable responses. The Resecting and Postprocessing columns give reference count / predicted count on those responses.

Model	Scorable	Resecting	Postprocessing
Claude Opus 5.5	165	33 / 59	33 / 23
Claude Sonnet 5	165	33 / 99	33 / 0
Gemma 4 31B IT	164	33 / 59	33 / 15
GLM-5.3 Flash	165	33 / 73	33 / 3
GPT-5.6 Luna	165	33 / 91	33 / 0
GPT-5.6 Sol	165	33 / 71	33 / 7
GPT-6 Astra	165	33 / 54	33 / 20
GPT-6 Luna	165	33 / 107	33 / 0
InternVL3.5	165	33 / 105	33 / 0
Kimi K3	165	33 / 87	33 / 1
Nemotron-3 Omni	162	33 / 45	33 / 0
Total	1,811	363 / 850	363 / 69

Many wrong answers are consistent with an earlier incorrect claim. Table 25 places the earliest supported divergence in a single-clip or attribute judgment for 283 responses (64.3%), in cross-clip relationships for 116 (26.4%), and in the task criterion for 23 (5.2%). Nine diverge during rule application or selection, and two between the written decision and submitted answer. Six cannot be localized and one has an ambiguous reference. The last written decision fully supports the submitted answer in 408 responses, suggesting that many errors occur before final answer expression. These are divergences in visible statements, not verified perceptual failures.

Answer choices show systematic preferences for phases and positions. In the PmLR50 Phase Recognition subset, models select Resecting 850 times against 363 reference occurrences, but Postprocessing only 69 times against 363 occurrences (Table 26). All 11 models show this direction. Both options appear with similar frequency in scorable responses (1,427 and 1,438), so option availability alone does not explain the gap.

Position preferences also appear in temporal tasks. Among wrong swap answers in Temporal Order Repair, 2,282 of 3,499 (65.2%) select adjacent positions, compared with 40.8% under uniform selection among the incorrect pairs. In Reference Alignment, we restrict the comparison to scorable wrong responses whose reference is not the displayed order A, B, C. Of these, 1,965 of 5,381 (36.5%) return that displayed order, compared with 20% under uniform selection among incorrect permutations. These patterns concern submitted choices rather than which images the model actually inspected.

Models often select a finding or repair when the reference specifies none. We compare six tasks with an explicit null answer: CVS Milestone Tracking, Description and Workflow Correction, Phase Consistency, Shared CVS Correction, Temporal Order Repair, and Workflow Path Recovery. Here, null means no attained CVS milestone, no required correction or outlier, or no qualifying middle clip, depending on the task. The rates reported in Section 3.4 use only scorable responses under the corresponding null or non-null reference condition. On matched questions scorable with and without images, the model-mean share of null answers is 11.7–66.9 percentage points higher without images. This shift is not uniform across model–task combinations.

Table 27: Shared CVS Correction by model. False positives propose a common operation when none is required. False negatives omit one when a common operation is required. Each rate uses only scorable responses under the corresponding reference condition.

Model	Scorable	False positive	False negative
Claude Opus 5.5	981	80.4%	11.2%
Claude Sonnet 5	966	65.6%	34.8%
Gemma 4 31B IT	980	99.4%	2.0%
GLM-5.3 Flash	969	75.2%	5.9%
GPT-5.6 Luna	981	94.0%	3.7%
GPT-5.6 Sol	981	58.0%	21.9%
GPT-6 Astra	981	51.2%	10.2%
GPT-6 Luna	981	95.5%	2.0%
InternVL3.5	981	7.4%	95.2%
Kimi K3	981	93.2%	2.6%
Nemotron-3 Omni	149	89.4%	14.7%

For Shared CVS Correction, 2,453 of 3,396 scorable responses with no required shared operation (72.2%) propose one. Conversely, 1,234 of 6,535 responses requiring a shared operation (18.9%) instead indicate that no shared operation is needed. Table 27 shows that ten models have a higher false-positive than false-negative rate. InternVL3.5 reverses the pattern, with rates of 7.4% and 95.2%. The no-change answer means that no operation is required by every clip, not that every clip has complete CVS. The observed errors expose a difficulty in the joint requirement of this task: recognizing a missing criterion in one clip is insufficient to justify a correction shared by all clips.

Related preferences appear in Workflow Repair: swapping is selected in 47.0% of scorable responses but required in 19.3%, while replacement is selected in 3.7% but required in 17.4%. Across CVS tasks, the same four-frame clip can receive incompatible answers. We match clips that reference the same image files and retain pairs where the milestone answer is correct and the reference marks the shared clip as having complete CVS. In 1,783 of these 2,775 model–question pairs (64.3%), the shared-correction answer nevertheless selects a nonzero operation, implying that the complete-CVS clip also needs that correction. The paired tasks differ in context and aggregation, and reuse clips and questions, so these counts measure output inconsistency rather than independent clinical errors.

E.1.3 Nonempty Unscorable Outputs

We inspect all 10,854 nonempty unscorable responses. Table 28 reports model-level counts and representative failure modes. Its percentages use each model's full 13,715-question evaluation set. The following percentages instead use the pooled unscorable responses.

Incomplete answers and invalid answer structures are distinct failures. Of these responses, 6,506 (59.9%) never complete an answer, 3,561 (32.8%) produce a well-formed answer structure that violates task constraints, 784 (7.2%) mark the beginning and end of an answer whose contents cannot be interpreted, and three are unusual cases where answer extraction fails. Missing required information or an operation that the task does not allow makes an answer invalid. A task-valid structure with the wrong decision counts as an incorrect answer.

A well-formed answer can still violate task rules. Among InternVL3.5's 913 unscorable responses, 745 have well-formed answer structures. Of its 332 constraint-violating Spatial Repair answers, 323 alter a relation that the prompt disallows changing. Nemotron-3 Omni gives a single operation number instead of a list in 599 Shared CVS Correction responses. Conversely, at least 98 outputs missing the closing marker for the answer already contain a task-valid response, and at least 124 present an explicit choice in a different surrounding format. These sets can overlap. These failures include both invalid task structures and incomplete or nonstandard answer submissions.

E.1.4 Empty Outputs

All 1,195 responses without final text come from Claude Sonnet 5 and are scored as incorrect. Each contains recorded reasoning but no final answer.

A response can contain reasoning without a final answer. Empty outputs account for 8.7% of Claude Sonnet 5's responses and are concentrated in Description and Workflow Correction, Workflow Repair, Temporal Order Repair, and Spatial Repair. Together these tasks account for 1,035 of

Table 28: Nonempty responses without a scorable answer in the 11 full evaluation runs. Each model answered 13,715 questions. Percentages use that denominator. The listed failure modes are observable examples within each row and need not exhaust or partition the row total.

Model	n (%)	Observed failure modes
Nemotron-3 Omni	5,229 (38.1)	3,646 lack a complete marked answer section. 1,124 have well-formed answers that violate task rules in the last complete answer section.
GLM-5.3 Flash	1,748 (12.7)	1,017 have no usable final answer. 403 last complete answer sections contain text outside the required structured format.
Gemma 4 31B IT	1,434 (10.5)	1,292 lack a complete marked answer section. 142 have well-formed answers that violate task rules.
InternVL3.5	913 (6.7)	745 have well-formed answer structures that violate task constraints, often in Spatial Repair.
Claude Sonnet 5	455 (3.3)	162 reach the output-token limit. 293 have well-formed answers that violate task rules.
GPT-5.6 Luna	387 (2.8)	All 387 have well-formed answers that violate task rules. 307 involve workflow repair.
GPT-6 Luna	331 (2.4)	All 331 have well-formed answers that violate task rules. 281 involve workflow repair.
Claude Opus 5.5	137 (1.0)	All 137 have well-formed answers that violate task rules. 113 include an unchanged spatial relation in a repair.
Kimi K3	79 (0.6)	65 have well-formed answers that violate task rules. 14 lack a complete final answer.
GPT-6 Astra	77 (0.6)	37 only lack the answer's closing marker. 3 have malformed answer structures. 37 have well-formed answers that violate task rules.
GPT-5.6 Sol	64 (0.5)	All 64 have well-formed answers that violate workflow-repair rules.

1,195 empty outputs (86.6%). Description and Workflow Correction alone has 543 empty outputs among 901 questions (60.3%). Claude Opus 5.5 returns final text on all 1,195 matched questions. Explicit rechecking language is common in both empty outputs (616/1,195, 51.5%) and length-matched outputs with final answers (38/67, 56.7%). These counts do not establish repeated image inspection as the cause of missing final answers.

E.2 Errors Before and After Training

We pair selected-checkpoint test responses to the same question within each task and dataset. For both RL and SFT, accuracy measures whether each answer is correct; it is not calculated from the RL reward. These analyses use each evaluation's original responses, whereas result tables use the shared reference scores in Section C.4.

The comparison includes 1,035 source-model/target-task entries: 174 with a separate pre-test, 570 with an existing Base reference, 285 with the source SFT model's post-test as the SFT-to-RL reference, and six reusing a single-task pre-test. Thus, not every comparison is an independent before-and-after test of the same run. Available evidence includes test outputs, answers, configurations, and checkpoint identities, but not the source frames or training-time generations. We use it to measure answer changes without assigning an unobserved visual or optimization cause.

E.2.1 Per-Task Reasoning Models

SFT improves answer delivery as well as correctness. Among the 2,660 recorded Base SFT-pre responses, all fail the strict reasoning-format check and 2,110 are invalid. After SFT, only 39 are invalid, while correct answers increase from 68 to 958. The paired changes include 938 gains and 48 losses. Learning the response format therefore contributes to the observed improvement, alongside changes in answer content.

A higher pooled score can conceal many newly incorrect answers. Sequential RL raises pooled accuracy from 36.02% to 38.91%, with 272 wrong-to-correct changes and 195 correct-to-wrong changes. Four tasks decline and one is unchanged. Reference Alignment falls from 40/190 (21.1%) to 24/190 (12.6%): 18 gains are outweighed by 34 losses. All 166 wrong post-RL answers on that task still satisfy the answer format requirements. The model still produces the required structure, but chooses different correspondences. Answer-format acquisition and reliable cross-case matching are therefore separable: preserving the former does not prevent regression in the latter.

E.3 Answer-Only Errors and Repeated Answers

The answer-only SFT pre-test contains 1,083 invalid outputs and 169 correct answers. Its post-test contains no invalid outputs and 865 correct answers. The remaining errors therefore persist after the model learns to deliver valid answers.

A model can learn a legal response without distinguishing questions. The answer-only Workflow Ordering model always lists clips in the displayed order A, B, C on all 184 test questions, matching only 30 references (16.3%). Accuracy declines from 37/184 (20.1%), with 19 gains and 26 losses. Reference Alignment similarly matches the reference clips to A, C, B, respectively, on all 190 questions, matching 39 (20.5%). Workflow Repair always selects deletion of the third clip, matching 26 of 188 targets (13.8%), and Temporal Order Repair proposes swapping clips A and C on 175 of 176 questions.

A frequent reference answer can make a constant prediction look strong. Both reasoning-mode and answer-only SFT give the same answer—neither CVS milestone is reached—on every CVS Milestone Tracking test question. That answer is correct for 120 of 185 questions, yielding 64.9% accuracy without any variation across inputs. By comparison, answer-only Spatial Repair produces 65 distinct answers across 159 questions and reaches 89.3%. This exposes a failure of input-dependent discrimination: the CVS models can match the most frequent reference without distinguishing complete from incomplete CVS in individual clips. Accuracy alone hides this behavior. Prediction diversity and reference-answer frequency reveal it.

E.4 Errors across Training Scopes

Benefits remain concentrated on the supervised tasks. Across 210 off-diagonal SFT comparisons, accuracy is 5.93% against 2.56% in the Base references, well below the 36.02% obtained on each model's own task. Category SFT raises within-category accuracy from 2.37% to 43.35%, compared with 2.56% to 7.98% outside the category. Sequential RL changes the corresponding recorded SFT references from 43.53% to 45.15% within the category and 7.98% to 7.96% outside it. These comparisons reuse target questions across source models and allow cross-task image overlap.

Joint training also improves output validity while retaining local regressions. Joint SFT reduces invalid responses from 2,103 to 32 and increases correct answers from 64 to 1,223, giving 45.98% on its recorded post-test. Sequential RL improves its separately recorded SFT reference from 46.13% to 48.23%, with 238 gains and 182 losses. Yet Shared CVS Correction falls from 24/130 to 12/130 (18.5% to 9.2%), and Workflow Repair from 71/188 to 56/188 (37.8% to 29.8%). Their paired changes are 12 gains versus 24 losses, and seven gains versus 22 losses, respectively.

The changed answers can remain complete and interpretable. In one Workflow Repair example, the correct SFT deletion of clip B becomes an insertion of Hydrodissection between clips B and C after RL. In a Shared CVS Correction example, the post-RL answer drops operation 2 from the correct set of operations 1, 2, and 3. These cases locate the regression in the submitted repair, while identifying its visual or training-time cause would require additional evidence.

F Action-Conditioned Image and Video Generation

F.1 State Transitions and Rollouts

Control theory studies how inputs influence a dynamical system and how to choose those inputs to achieve a desired behavior (Åström and Murray, 2021). In a discrete-time state-space description, the current state s_t and control input a_t determine the next state through a transition function f. A learned world model approximates these dynamics and predicts future states under supplied actions (Ha and Schmidhuber, 2018). In the deterministic form, the true and predicted transitions are

$$s_{t+1} = f(s_t, a_t), \qquad \hat{s}_{t+1} = f_\theta(\hat{s}_t, a_t), \qquad \hat{s}_0 = s_0. \tag{4}$$

The predicted state becomes the input to the next step: $\hat{s}_{t+2} = f_\theta(\hat{s}_{t+1}, a_{t+1})$. Repeating this operation produces a *rollout*, $\hat{s}_{1:T} = (\hat{s}_1, \ldots, \hat{s}_T)$: a sequence of predicted future states from the initial state under the supplied actions.

F.2 Task and Conditioning

We use this perspective to test whether a generative model can continue a surgical scene while following a specified instrument–action–target relation. Here, s_t denotes the visible scene represented by a frame, rather than a fully observed physical state. Each example supplies an initial image s_0 and an annotated action description A over a six-second horizon. The description specifies a semantic action over that interval, not a sequence of measured control inputs a_t. Video generation is therefore a conditional prediction $\hat{s}_{1:T} \sim p_\theta(\cdot \mid s_0, A)$, with T indexing the end of the horizon. The visual rollout is generated in a single request. This formulation does not assume that the generator explicitly implements the recurrence in Equation 4.

Gemini 3.1 Flash Image predicts only the terminal image $\hat{s}_T$. Veo 3.1 predicts the sequence $\hat{s}_{1:T}$, including the endpoint and intervening changes. Both receive the same starting image and action, and neither sees the recorded continuation. The terminal image is generated independently of the video.

We compare the predicted endpoint or trajectory with the recorded continuation $s_{1:T}$:

$$e_{\text{image}} = D_{\text{end}}(\hat{s}_T, s_T), \qquad e_{\text{video}} = D_{\text{traj}}(\hat{s}_{1:T}, s_{1:T}). \tag{5}$$

Here, D_{end} and D_{traj} denote the categorical human judgments of action and scene consistency defined in Section F.3, rather than numerical pixel distances. Image review assesses the resulting visible state. Video review additionally assesses whether the intervening transition remains consistent with the action and initial scene.

We compare action-only prompts with prompts that additionally name the instruments visible at the initial frame. These current-state labels provide no coordinates, future trajectory, or outcome. All prompts ask the model to preserve the initial anatomy, instruments, viewpoint, and scene geometry while showing the requested action. They permit changes caused by that action, but disallow unrelated events, additional tools, overlays, and camera transitions. The action may remain in progress at the end of the horizon.

F.3 Generation and Evaluation Protocol

Gemini 3.1 Flash Image (Google, 2026a) generates one 16:9 terminal-state image. Veo 3.1 uses `veo-3.1-generate-preview` (Google, 2026b), as listed in Table 7, to generate one continuous six-second, 16:9 video.

The paired set contains 100 starting points from five Cholec80 videos, contributing 29, 25, 18, 14, and 14 examples. Frame-aligned CholecT50 triplets specify the action, and Cholec80 tool-presence annotations supply the optional state labels. Every point has all four generated conditions and their exact prompts. A ten-second, 25-fps source-video slice is available to the reviewer for context. Filtered or incomplete requests are excluded before forming this paired set.

A reviewer compares the initial frame, action, recorded continuation, and generated output, assigning one primary outcome to each of the 400 outputs:

- *Consistent with reference*: the action and scene continuation are plausible and compatible with the recording. Exact pixel or motion matching is not required.
- *Action-triplet error*: the requested instrument, action, or target is incorrect or absent.
- *Instrument–tissue interaction error*: contact or tissue response is inconsistent despite recognizable entities.
- *Initial-scene deviation*: the output changes the starting anatomy, instruments, arrangement, or viewpoint without support from the requested action.
- *Unrequested surgical event*: an additional step or effect falls outside the supplied command.
- *Visual/geometry generation error*: objects or motion exhibit distortion, disappearance, duplication, or discontinuity.

Only the most salient outcome is recorded, so error counts do not capture all co-occurring faults. The review has no independent second rating or written case notes. Source action and tool-label alignment has not received a separate completed human audit. Each condition has one sampled

Table 29: Primary human outcome for action-conditioned generation. Each column has 100 outputs, so each count is also a percentage. The six labels are mutually exclusive primary outcomes.

	Image		Video	
Primary outcome	Action only	Action + labels	Action only	Action + labels
Action-triplet error	9	7	27	21
Instrument–tissue interaction	2	4	2	4
Initial-scene deviation	2	0	19	16
Unrequested surgical event	3	9	25	25
Visual/geometry generation	2	0	5	6
Consistent with reference	**82**	**80**	**22**	**28**

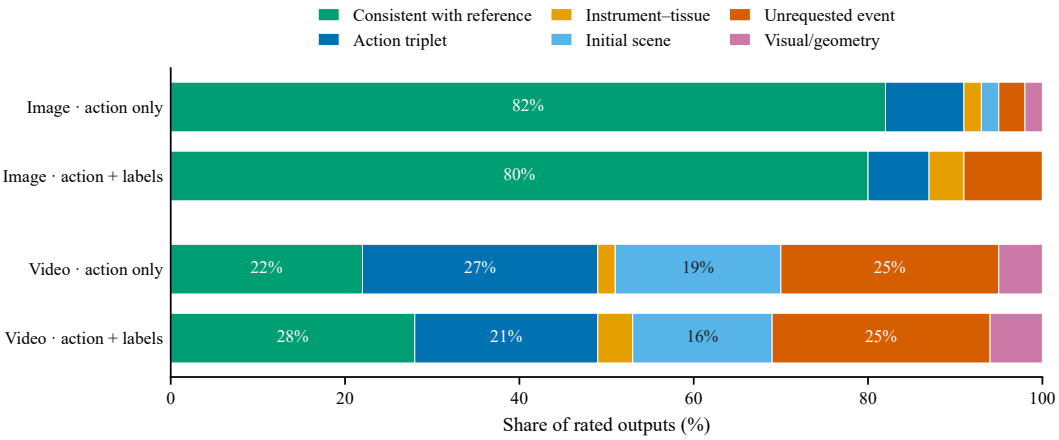


Figure 14: Distribution of the six primary outcomes in each generation condition ($n = 100$ per condition). The green segment denotes a continuation judged consistent with the requested action and source recording. The other segments denote the five error types. The two modalities use different generators and different output requirements.

generation per starting point, and examples cluster within five videos. The results do not estimate variation across generations, reviewers, or surgical cases.

F.4 Results

Table 29 reports outcome counts, and Figure 14 shows their distribution. Each condition contains 100 outputs, so counts also give percentages.

Plausible endpoints are much more common than consistent video continuations. A favorable endpoint does not establish a valid intervening sequence. The conditions use different generators and assessment requirements, so the gap is not an isolated effect of output modality.

Video errors mainly concern the commanded action and allowed events. Among 150 error-labeled videos, 50 contain an unrequested surgical event, 48 have an action-triplet error, and 35 deviate from the initial scene. These account for 133/150 primary errors. Interaction and visual/geometry errors contribute six and 11 ratings. Because only one label is recorded, their smaller counts do not imply that contact and rendering are otherwise reliable. Images also retain semantic errors: among 38 non-positive ratings, 16 concern the action triplet and 12 an unrequested event.

Current-state labels yield mixed paired outcomes. With labels, nine image starting points receive a positive rating where the action-only output received an error, while 11 show the reverse, yielding a net change of -2 percentage points. For videos, 16 improve and ten worsen, yielding $+6$ points. With one sampled output per condition, these differences cannot isolate a causal effect of the added labels from generation variability. The labels specify instrument identity but leave contact, motion, and tissue response unspecified. This distinction offers an interpretation of the mixed results: nam-

ing the entities does not specify the transition needed to execute the requested relation. The ratings alone cannot determine which part of that transition each model represents.

F.5 Action-Level Differences

Four triplets account for 76 of the 100 starting points. For gallbladder retraction with a grasper, only 4/44 videos across both prompt variants receive a positive rating, while 33/44 are labeled as unrequested events. For grasping the specimen bag with a grasper, 17/24 videos are judged consistent. This identifies retraction examples as a useful source of observed failures, but the actions differ in context and sample size. The primary labels do not identify the precise extra event or its onset, so they support an action-specific description of failures rather than a causal account of surgical difficulty.

G Prompt Templates and Task Definitions

G.1 Complete Prompt Examples

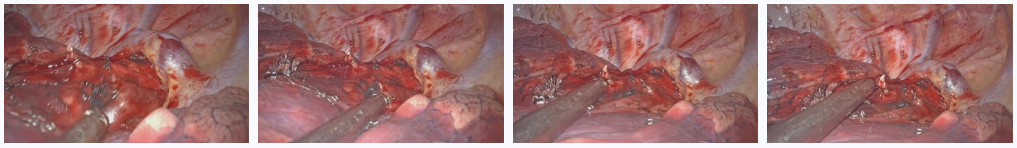
Figures 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, and 29 provide one reviewed example per task, including visual evidence, rules, and the required answer. Continued pages belong to the same example. JavaScript Object Notation (JSON) specifies the answer format shown in the cards.

Phase Recognition

[Task]

You are solving a cataract surgery video phase recognition task. You must decide which surgical phase is shown in Clip A. The visual evidence for this question is provided below.

Clip A, frame 1 | Clip A, frame 2 | Clip A, frame 3 | Clip A, frame 4



[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.

[Question]

What surgical phase is shown in this clip?
Options:
A. Capsulorhexis
B. Viscoelastic
C. Anterior chamber flushing
D. Irrigation/ aspiration

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": "{ans}"}
and nothing else. Replace {think} with your reasoning. Replace {ans} with the single legal answer for this task, chosen from "A", "B", "C", and "D".

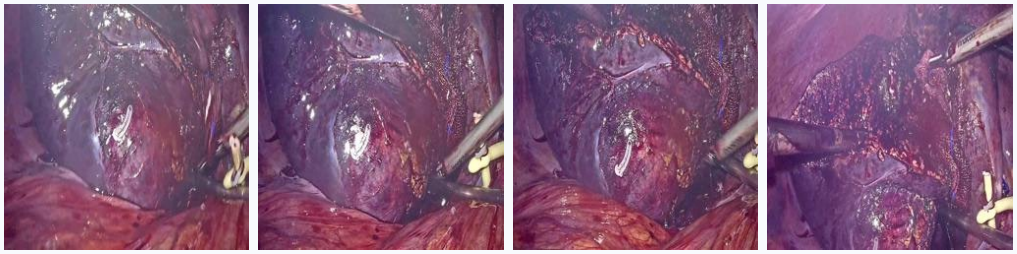
Figure 15: Example of the Phase Recognition task under the Phase Understanding category.

Phase Consistency

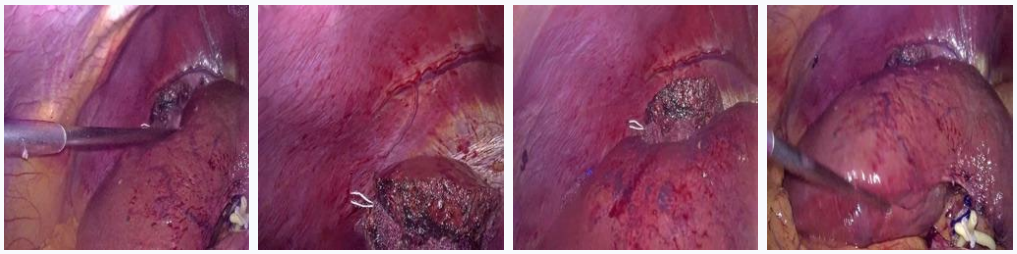
[Task]

You are solving a cataract surgery video phase consistency task over four clips. You must decide whether one clip shows a different surgical phase or all four clips show the same surgical phase. The visual evidence for this question is provided below.

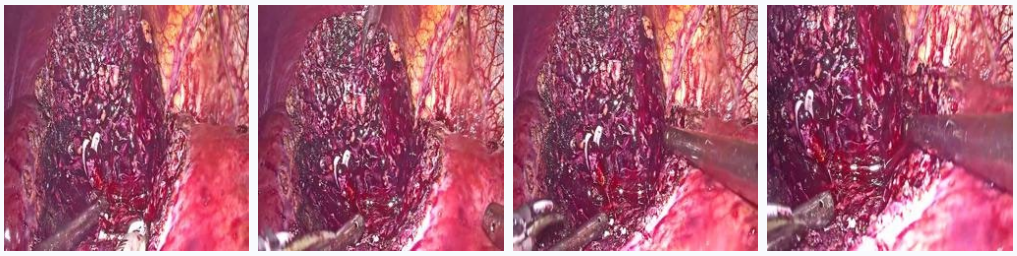
Clip A, frame 1 | Clip A, frame 2 | Clip A, frame 3 | Clip A, frame 4



Clip B, frame 1 | Clip B, frame 2 | Clip B, frame 3 | Clip B, frame 4



Clip C, frame 1 | Clip C, frame 2 | Clip C, frame 3 | Clip C, frame 4



Clip D, frame 1 | Clip D, frame 2 | Clip D, frame 3 | Clip D, frame 4

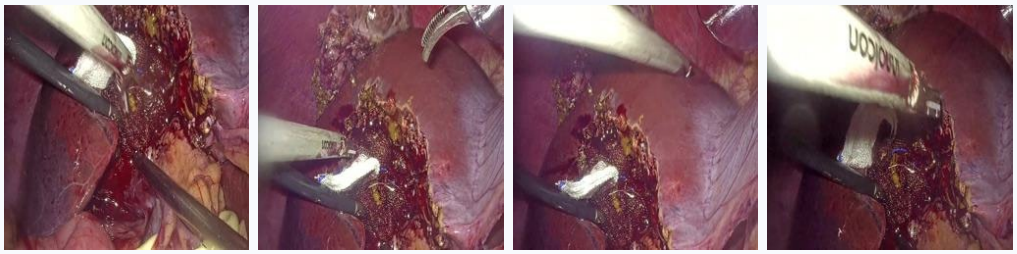


Figure 16: Example of the Phase Consistency task under the Phase Understanding category.

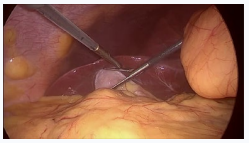
[Rules]	[Question]	[Answer Format]
1. Use only the images and text provided in this prompt. 2. If answer options are provided, choose only from the provided options. 3. Follow the required answer format exactly.	Which clip, if any, belongs to a different surgical phase from the other three? Options: A. Clip A. B. Clip B. C. Clip C. D. Clip D. E. None of the above; all four clips belong to the same surgical phase.	Output exactly one JSON object: {"think": "{think}", "answer": "{ans}"} and nothing else. Replace {think} with your reasoning. Replace {ans} with the single legal answer for this task, chosen from "A", "B", "C", "D", and "E".

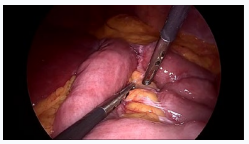
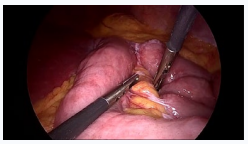
Figure 16 (continued): Example of the Phase Consistency task under the Phase Understanding category.

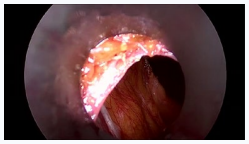
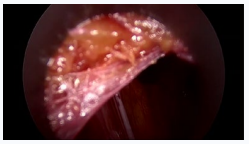
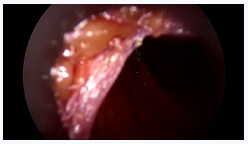
Action Description Matching

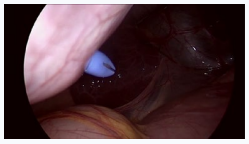
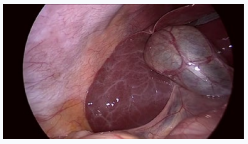
[Task]

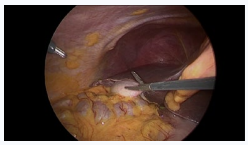
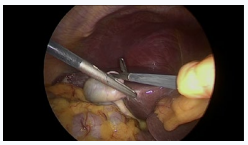
You are solving an instrument-action-target triplet matching task using laparoscopic cholecystectomy clips. You must identify the instrument, action, and target shown in each clip and match all three to the correct triplet description. The visual evidence for this question is provided below.

Clip A, frame 1 
Clip A, frame 2 
Clip A, frame 3 
Clip A, frame 4 

Clip B, frame 1 
Clip B, frame 2 
Clip B, frame 3 
Clip B, frame 4 

Clip C, frame 1 
Clip C, frame 2 
Clip C, frame 3 
Clip C, frame 4 

Clip D, frame 1 
Clip D, frame 2 
Clip D, frame 3 
Clip D, frame 4 

Clip E, frame 1 
Clip E, frame 2 
Clip E, frame 3 
Clip E, frame 4 

[Rules]

1. Use only the images and text provided in this prompt.
2. Follow the required answer format exactly.
3. Match each Clip to exactly one of the provided description IDs.
4. The same description ID may be selected for multiple Clips.

[Question]

Match each Clip to its correct instrument-action-target triplet description.
Descriptions:
1. grasper — grasp — gallbladder
2. grasper — retract — liver
3. grasper — retract — gut
4. scissors — cut — adhesion

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Replace {ans} with a JSON array of five integer description IDs (1-4), corresponding to Clips A, B, C, D, and E in that order.

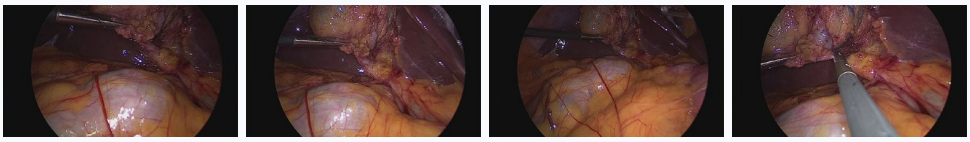
Figure 17: Example of the Action Description Matching task under the Scene Understanding category.

Shared CVS Correction

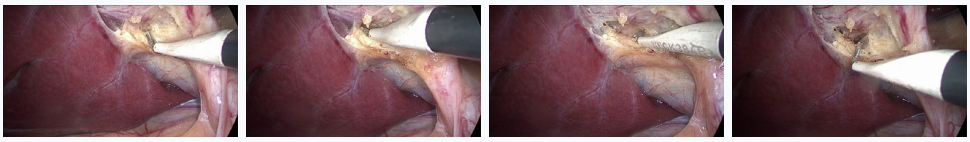
[Task]

You are solving a shared Critical View of Safety corrective-action selection task for laparoscopic cholecystectomy. Compare the four Clips and identify every corrective operation required for all four. The Clips come from different surgeries, and their CVS states may be the same or different. Each Clip contains four chronological frames with the same three-criterion CVS state. Judge the whole Clip. The visual evidence for this question is provided below.

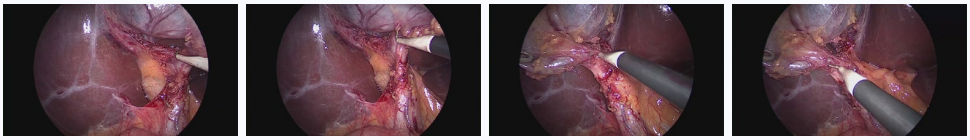
Clip A, frame 1 | Clip A, frame 2 | Clip A, frame 3 | Clip A, frame 4



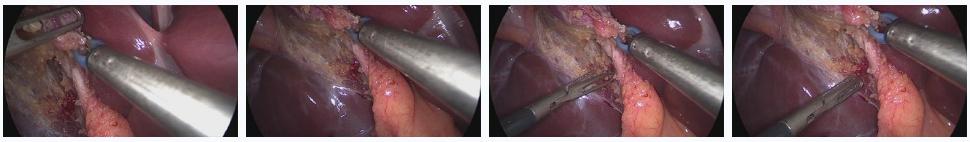
Clip B, frame 1 | Clip B, frame 2 | Clip B, frame 3 | Clip B, frame 4



Clip C, frame 1 | Clip C, frame 2 | Clip C, frame 3 | Clip C, frame 4



Clip D, frame 1 | Clip D, frame 2 | Clip D, frame 3 | Clip D, frame 4



[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.
4. Select an operation only if it is required for every one of Clips A, B, C, and D. Select all such operations.
5. Select 0 alone if no operation is required for all four Clips. This includes both complete CVS in every Clip and any other combination with no shared corrective operation.

[Question]

Based on the current CVS status, which corrective operations are required for all four Clips?
Options:
0. None of the corrective operations below is required for all four Clips.
1. Continue clearing the hepatocystic triangle.
2. Continue separating the lower gallbladder to expose the cystic plate.
3. Continue identification and separation until only two structures enter the gallbladder.

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Replace {ans} with a JSON array containing all shared corrective-operation IDs as integers in ascending order, without duplicates.

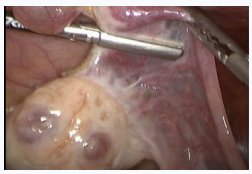
Figure 18: Shared CVS Correction

Spatial Repair

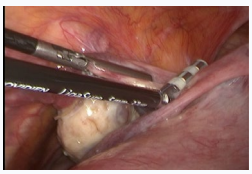
[Task]

You are solving an instrument–anatomy spatial relation audit and repair task for cataract surgery. You must decide whether each Claim's one or two spatial relations agree with the visible geometry in every Frame of Clip A. The visual evidence for this question is provided below.

Clip A, frame 1



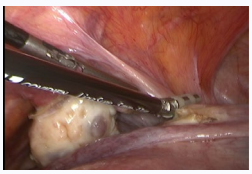
Clip A, frame 2



Clip A, frame 3



Clip A, frame 4



[Rules]

1. Use only the images and text provided in this prompt.
2. The instrument and anatomy names in every Claim are correct; only the spatial relations may be corrupted.
3. Follow the required answer format exactly.
4. Every Claim applies to all four Frames shown in Clip A and is correct only when all stated relations hold in every Frame.

[Question]

Audit each spatial Claim against all four Frames in Clip A.
Claims:
Claim 1: The Spatula is below the Pupil.
Claim 2: The Irrigation Aspiration is disjoint from and above the Pupil.
Claim 3: The Spatula is below the Cornea.

[Answer Format]

Output exactly one JSON object:
{"think": "{think}",
"answer": {ans}}
and nothing else. Replace {think} with your reasoning. Replace {ans} with a JSON object mapping incorrect Claim IDs to arrays of corrected relations. Include only the relations that need correction, in the order they appear in the Claim. Only switch between "left_of" and "right_of", between "above" and "below", or between "overlapping" and "disjoint".

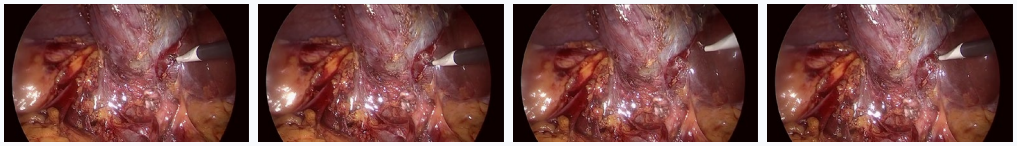
Figure 19: Example of the Spatial Repair task under the Scene Understanding category.

Reference Alignment

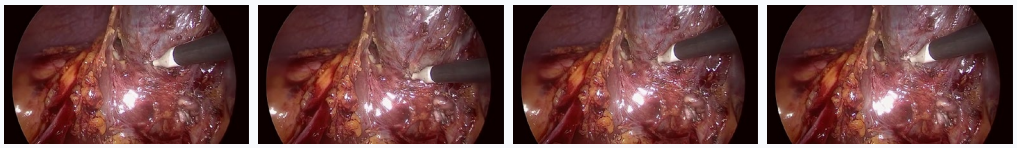
[Task]

You are solving a surgical action-ordering task in laparoscopic cholecystectomy. You must use the ordered reference action sequence to recover the order of the query Clips. The visual evidence for this question is provided below.

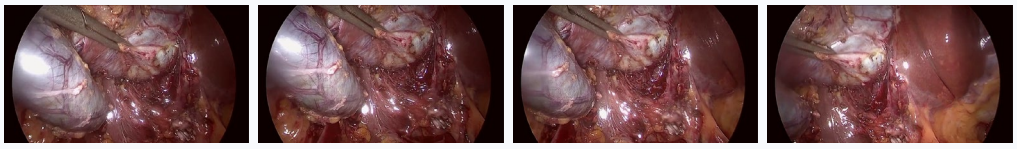
Clip R1, frame 1 Clip R1, frame 2 Clip R1, frame 3 Clip R1, frame 4



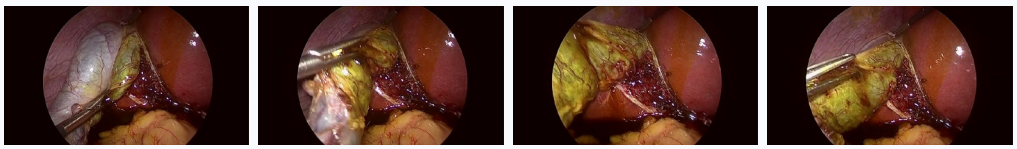
Clip R2, frame 1 Clip R2, frame 2 Clip R2, frame 3 Clip R2, frame 4



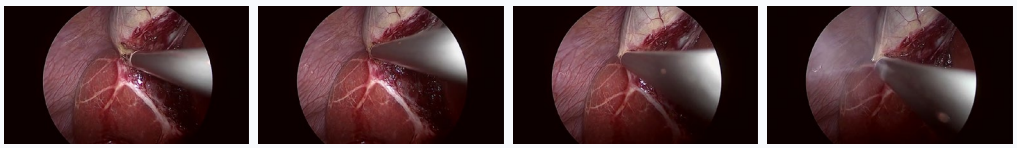
Clip R3, frame 1 Clip R3, frame 2 Clip R3, frame 3 Clip R3, frame 4



Clip A, frame 1 Clip A, frame 2 Clip A, frame 3 Clip A, frame 4



Clip B, frame 1 Clip B, frame 2 Clip B, frame 3 Clip B, frame 4



Clip C, frame 1 Clip C, frame 2 Clip C, frame 3 Clip C, frame 4

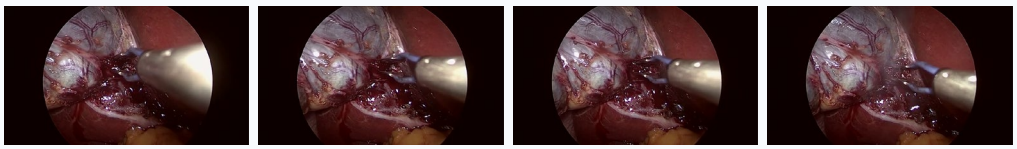


Figure 20: Example of the Reference Alignment task under the Temporal Reasoning category.

[Rules]	[Question]	[Answer Format]
1. Use only the images and text provided in this prompt. 2. If answer options are provided, choose only from the provided options. 3. Follow the required answer format exactly. 4. Use each of "A", "B", and "C" exactly once.	Reference Clips R1, R2, and R3 show three surgical actions in reference order. Each query Clip shows one of the same three actions. Arrange query Clips A, B, and C to match the reference action order.	Output exactly one JSON object: {"think": "{think}", "answer": {ans}} and nothing else. Replace {think} with your reasoning. Replace {ans} with the actual answer: a JSON array of the query Clip IDs in the corresponding reference order.

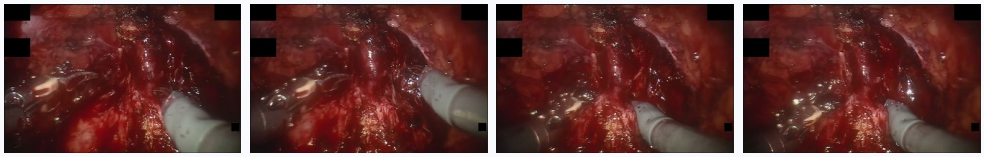
Figure 20 (continued): Example of the Reference Alignment task under the Temporal Reasoning category.

Temporal Order Repair

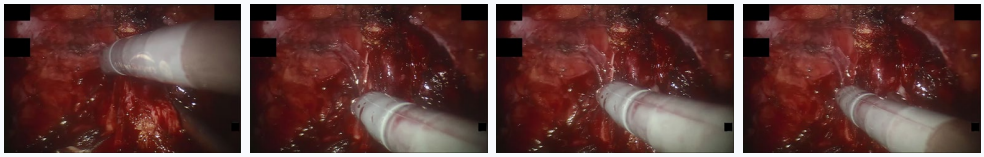
[Task]

You are solving a within-phase sequence repair task over five clips from one cataract surgery case. You must decide whether the five displayed clips preserve their true within-phase chronological order. The visual evidence for this question is provided below.

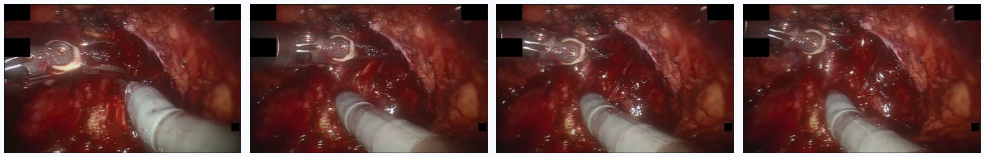
Clip A, frame 1 | Clip A, frame 2 | Clip A, frame 3 | Clip A, frame 4



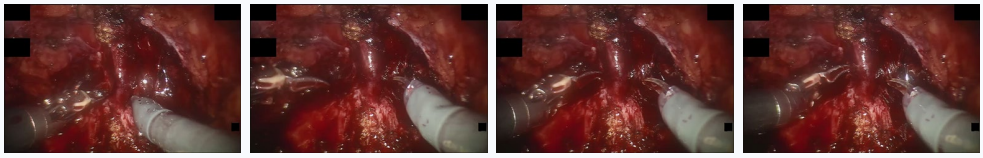
Clip B, frame 1 | Clip B, frame 2 | Clip B, frame 3 | Clip B, frame 4



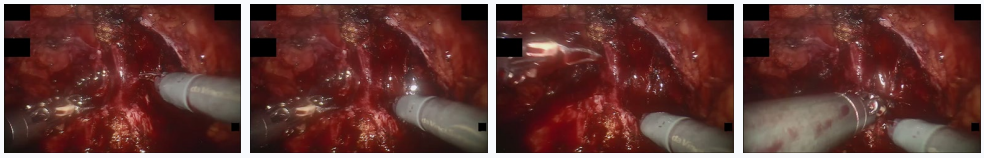
Clip C, frame 1 | Clip C, frame 2 | Clip C, frame 3 | Clip C, frame 4



Clip D, frame 1 | Clip D, frame 2 | Clip D, frame 3 | Clip D, frame 4



Clip E, frame 1 | Clip E, frame 2 | Clip E, frame 3 | Clip E, frame 4



[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.

[Question]

Which two Clips, if any, must be swapped to restore the chronological order?

[Answer Format]

Output exactly one JSON object: {"think": "{think}", "answer": {ans}} and nothing else. Replace {think} with your reasoning. Replace {ans} with the actual answer: a JSON array containing the two distinct Clip IDs that must be swapped, listed in alphabetical order. Use [] when the displayed Clips are already in chronological order.

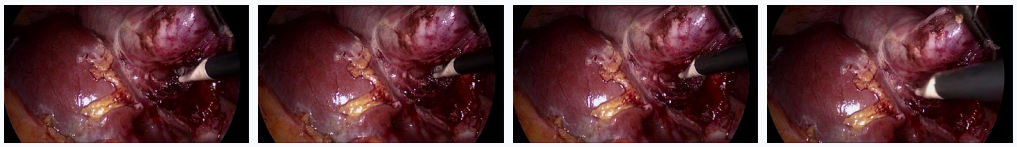
Figure 21: Example of the Temporal Order Repair task under the Temporal Reasoning category.

CVS Milestone Tracking

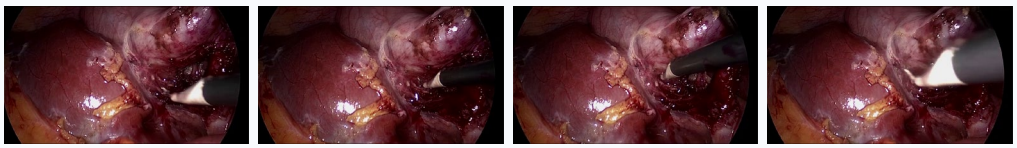
[Task]

You are solving a temporal Critical View of Safety tracking task in laparoscopic cholecystectomy. You must track complete CVS across a chronologically ordered sequence of clips. Each Clip contains four chronological frames with the same three-criterion CVS state. Judge the whole Clip. The visual evidence for this question is provided below.

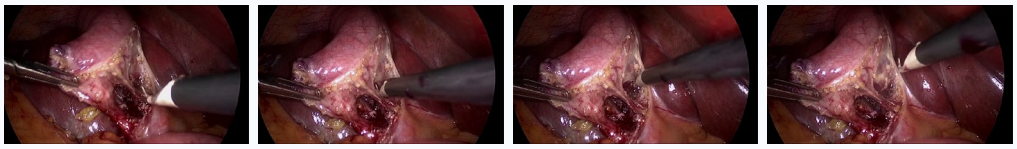
Clip A, frame 1 Clip A, frame 2 Clip A, frame 3 Clip A, frame 4



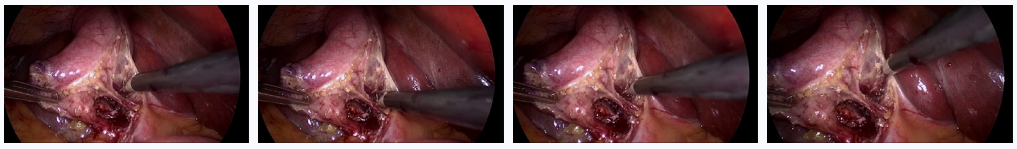
Clip B, frame 1 Clip B, frame 2 Clip B, frame 3 Clip B, frame 4



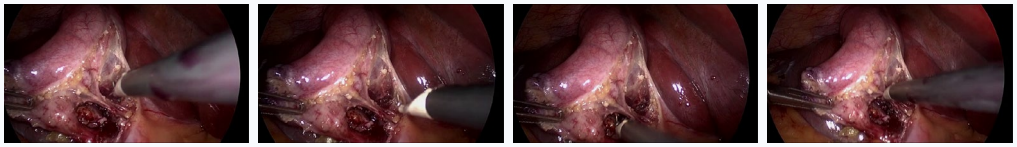
Clip C, frame 1 Clip C, frame 2 Clip C, frame 3 Clip C, frame 4



Clip D, frame 1 Clip D, frame 2 Clip D, frame 3 Clip D, frame 4



Clip E, frame 1 Clip E, frame 2 Clip E, frame 3 Clip E, frame 4



Clip F, frame 1 Clip F, frame 2 Clip F, frame 3 Clip F, frame 4

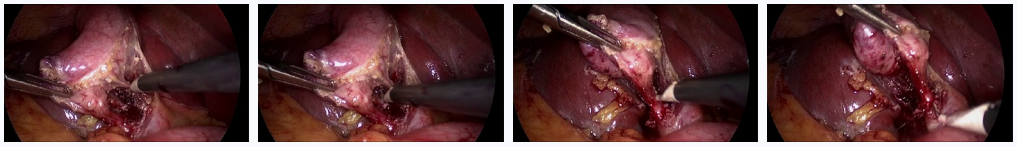


Figure 22: Example of the CVS Milestone Tracking task under the Temporal Reasoning category.

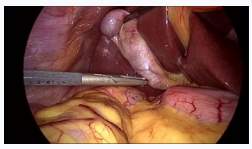
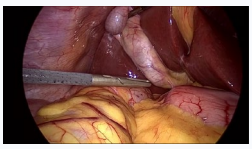
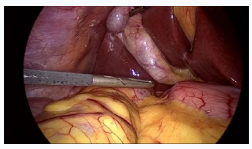
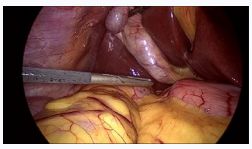
[Rules]	[Question]	[Answer Format]
1. Use only the images and text provided in this prompt. 2. If answer options are provided, choose only from the provided options. 3. Follow the required answer format exactly.	Which displayed Clip first shows complete CVS, and which is the first Clip where both it and the previous Clip show complete CVS? Complete CVS means that all three criteria are satisfied. Sustained complete CVS here means at least two consecutive displayed Clips are complete.	Output exactly one JSON object: {"think": "{think}", "answer": {ans}} and nothing else. Replace {think} with your reasoning. Replace {ans} with a JSON array of two values: the first complete-CVS Clip ID, followed by the first Clip ID where both that Clip and the previous Clip show complete CVS. Use "NONE" if the corresponding result does not exist.

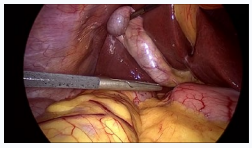
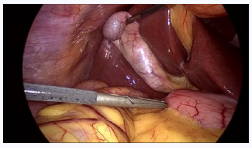
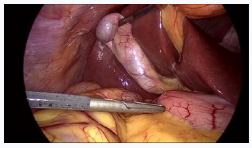
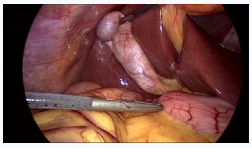
Figure 22 (continued): Example of the CVS Milestone Tracking task under the Temporal Reasoning category.

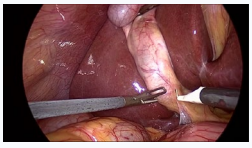
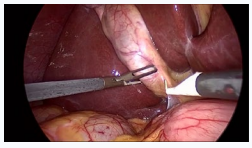
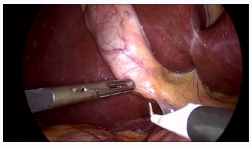
State Change Tracking

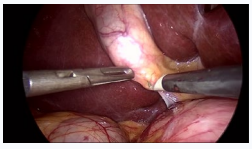
[Task]

You are solving a multi-field surgical-state change comparison task in laparoscopic cholecystectomy. You must identify where the specified fields change across Clips A-D. Each Clip contains four chronological frames with consistent values for all compared fields. The visual evidence for this question is provided below.

Clip A, frame 1 
Clip A, frame 2 
Clip A, frame 3 
Clip A, frame 4 

Clip B, frame 1 
Clip B, frame 2 
Clip B, frame 3 
Clip B, frame 4 

Clip C, frame 1 
Clip C, frame 2 
Clip C, frame 3 
Clip C, frame 4 

Clip D, frame 1 
Clip D, frame 2 
Clip D, frame 3 
Clip D, frame 4 

[Rules]

1. Use only the images and text provided in this prompt.
2. Follow the required answer format exactly.
3. Compare each Clip with the immediately preceding Clip.
4. When a field changes, report the later Clip ID: "B" for A→B, "C" for B→C, and "D" for C→D.

[Question]

Clips A, B, C, and D come from the same surgical case and are shown in chronological order. Compare only the following fields:
- phase: the surgical phase
- instrument: the visible instrument set
For each listed field, in which Clips does its value change from the immediately preceding Clip?
For a set-valued field, any addition or removal of a category counts as a change. Do not report changes in other attributes.

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Replace {ans} with a JSON object. Use each field listed in [Question] as a key. Its value must be an array of Clip IDs where that field changes compared with the immediately preceding Clip. List Clip IDs in chronological order; use [] if there is no change.

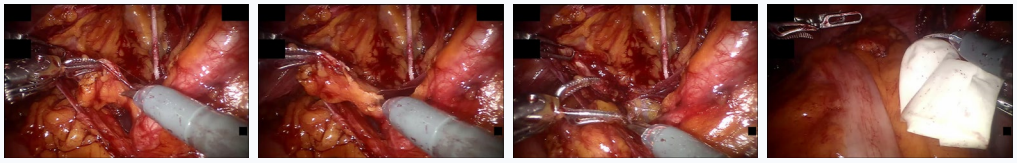
Figure 23: Example of the State Change Tracking task under the Temporal Reasoning category.

Spatial Relation Tracking

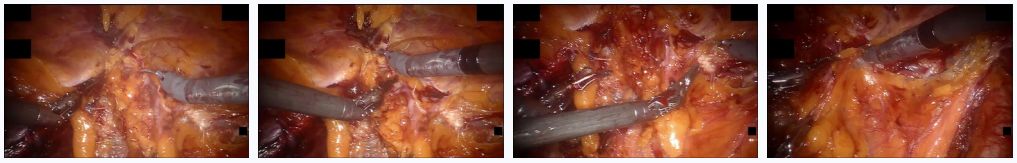
[Task]

You are solving a spatial relation tracking task in radical prostatectomy. You must identify where the specified spatial relations between the Subject and Object change across Clips A-D. The visual evidence for this question is provided below.

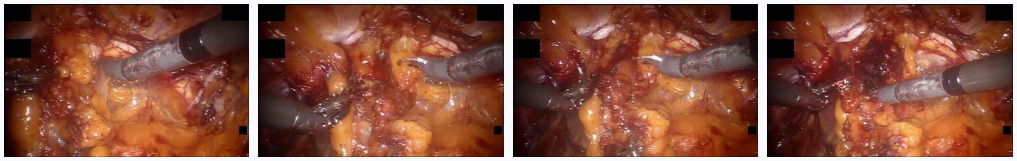
Clip A, frame 1 Clip A, frame 2 Clip A, frame 3 Clip A, frame 4



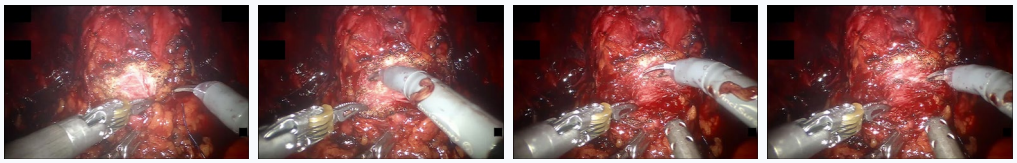
Clip B, frame 1 Clip B, frame 2 Clip B, frame 3 Clip B, frame 4



Clip C, frame 1 Clip C, frame 2 Clip C, frame 3 Clip C, frame 4



Clip D, frame 1 Clip D, frame 2 Clip D, frame 3 Clip D, frame 4



[Rules]

1. Use only the images and text provided in this prompt.
2. Follow the required answer format exactly.
3. Compare each Clip with the immediately preceding Clip.
4. When a relation changes, report the later Clip ID: "B" for A→B, "C" for B→C, and "D" for C→D.

[Question]

Subject: Bipolar Forceps
Object: Monopolar Curved Scissors
Relations to compare: horizontal, vertical
For each listed relation, in which Clips does the Subject's relation to the Object change from the immediately preceding Clip?

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Replace {ans} with a JSON object. Use each relation type listed in [Question] as a key. Its value must be an array of Clip IDs where that relation changes compared with the immediately preceding Clip. List Clip IDs in chronological order; use [] if there is no change.

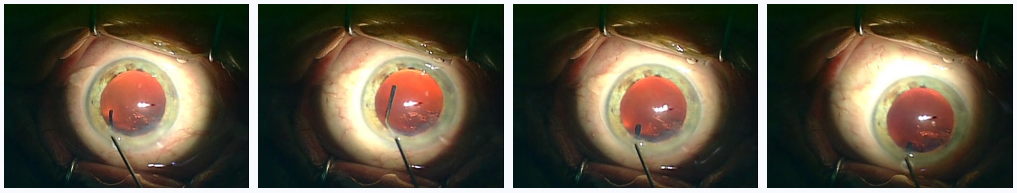
Figure 24: Example of the Spatial Relation Tracking task under the Temporal Reasoning category.

Workflow Ordering

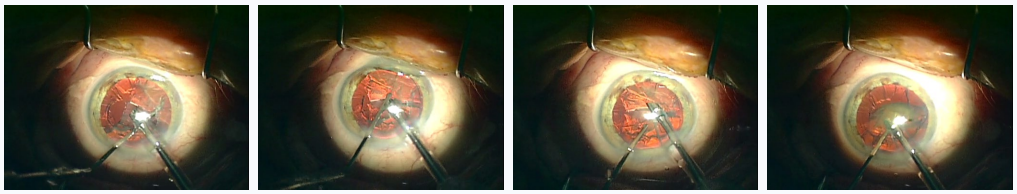
[Task]

You are solving a cataract surgery video procedural ordering task. You must decide the relative order of the three displayed Clips, which come from the same surgery. The visual evidence for this question is provided below.

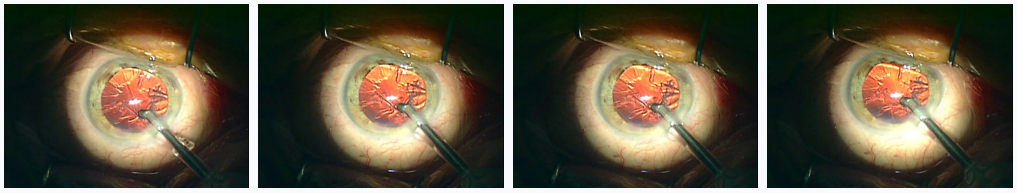
Clip A, frame 1 | Clip A, frame 2 | Clip A, frame 3 | Clip A, frame 4



Clip B, frame 1 | Clip B, frame 2 | Clip B, frame 3 | Clip B, frame 4



Clip C, frame 1 | Clip C, frame 2 | Clip C, frame 3 | Clip C, frame 4



[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.
4. Use each displayed Clip ID exactly once.

[Question]

Arrange the three displayed Clips in procedural order from earliest to latest.

[Answer Format]

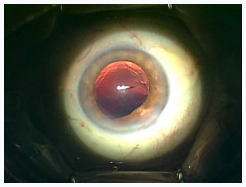
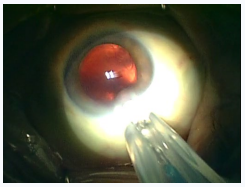
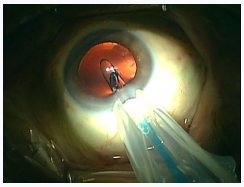
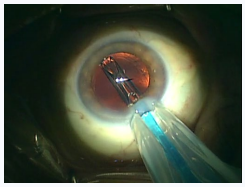
Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Replace {ans} with the actual answer: a JSON array containing the three Clip IDs from earliest to latest.

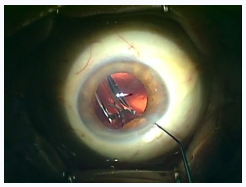
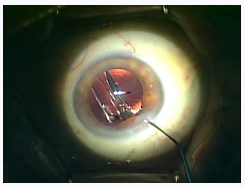
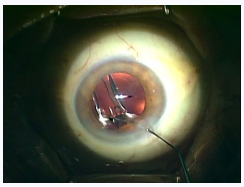
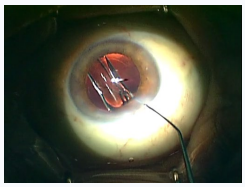
Figure 25: Example of the Workflow Ordering task under the Workflow Reasoning category.

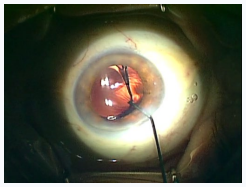
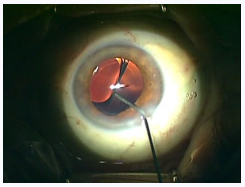
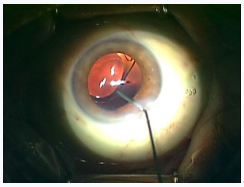
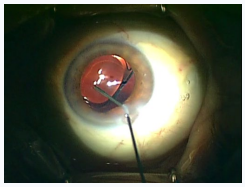
Workflow Path Recovery

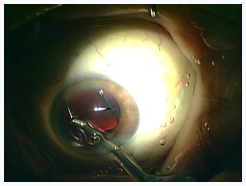
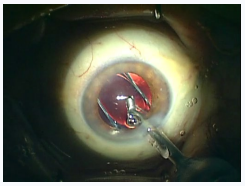
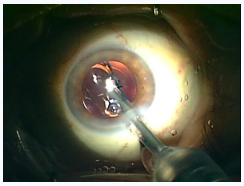
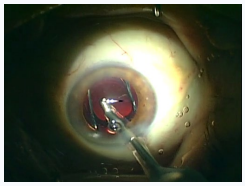
[Task]

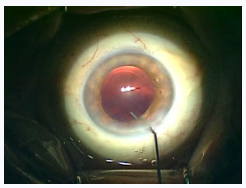
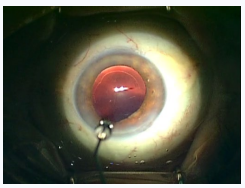
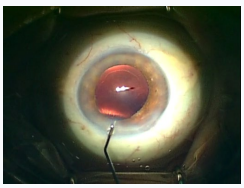
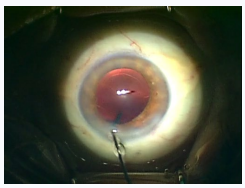
You are solving a cataract surgery video workflow path recovery task. You must recover the missing middle stages between fixed start and end Clips. The visual evidence for this question is provided below.

Clip A, frame 1 
Clip A, frame 2 
Clip A, frame 3 
Clip A, frame 4 

Clip B, frame 1 
Clip B, frame 2 
Clip B, frame 3 
Clip B, frame 4 

Clip C, frame 1 
Clip C, frame 2 
Clip C, frame 3 
Clip C, frame 4 

Clip D, frame 1 
Clip D, frame 2 
Clip D, frame 3 
Clip D, frame 4 

Clip E, frame 1 
Clip E, frame 2 
Clip E, frame 3 
Clip E, frame 4 

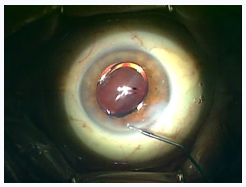
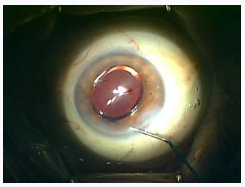
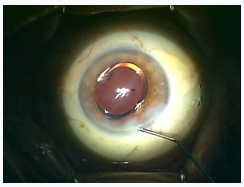
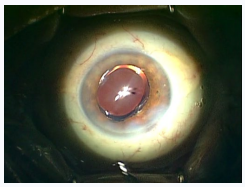
Clip F, frame 1 
Clip F, frame 2 
Clip F, frame 3 
Clip F, frame 4 

Figure 26: Example of the Workflow Path Recovery task under the Workflow Reasoning category.

[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.

[Question]

Clip A is the fixed start and Clip F is the fixed end.
Which of Clips B, C, D, and E belong between them in the procedural workflow?

[Answer Format]

Output exactly one JSON object: {"think": "{think}", "answer": {ans}} and nothing else. Replace {think} with your reasoning. Replace {ans} with the actual answer: a JSON array containing every candidate Clip ID that belongs between Clip A and Clip F, in procedural order. The answer may contain zero to four candidate Clip IDs. Use [] when none of the candidate Clips belongs between them.

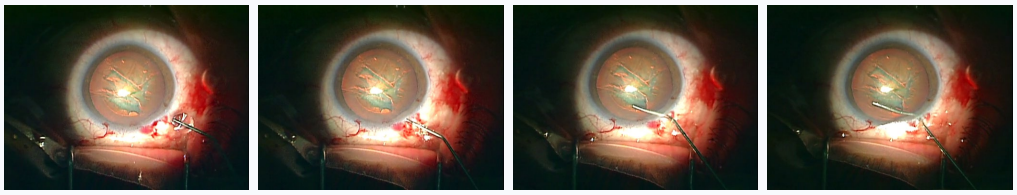
Figure 26 (continued): Example of the Workflow Path Recovery task under the Workflow Reasoning category.

Workflow Repair

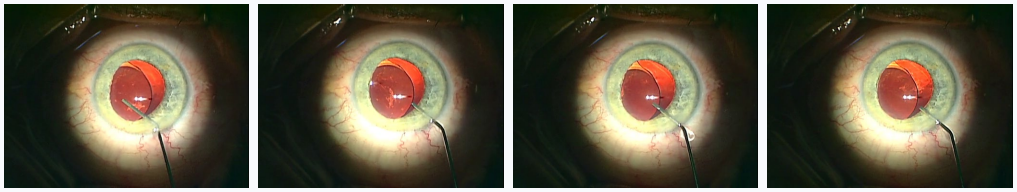
[Task]

You are solving a cataract surgery video workflow repair task over four displayed Clips. You must locate exactly one workflow error and provide its minimum repair. The expected surgical workflow is: 1. Incision 2. Viscoelastic 3. Capsulorhexis 4. Hydrodissection 5. Phacoemulsification 6. Irrigation/aspiration 7. Capsule polishing 8. Lens implantation 9. Lens positioning 10. Viscoelastic suction 11. Anterior chamber flushing 12. Tonifying/antibiotics The visual evidence for this question is provided below.

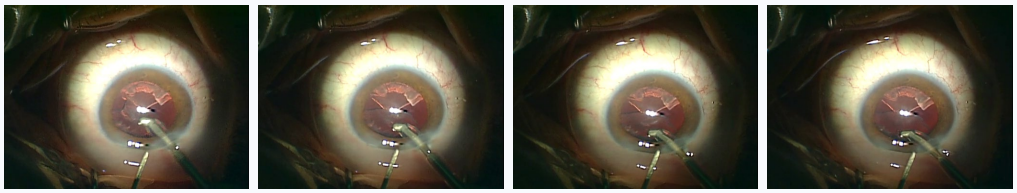
Clip A, frame 1 Clip A, frame 2 Clip A, frame 3 Clip A, frame 4



Clip B, frame 1 Clip B, frame 2 Clip B, frame 3 Clip B, frame 4



Clip C, frame 1 Clip C, frame 2 Clip C, frame 3 Clip C, frame 4



Clip D, frame 1 Clip D, frame 2 Clip D, frame 3 Clip D, frame 4

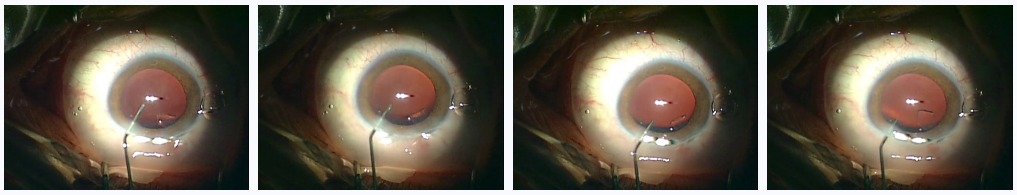


Figure 27: Example of the Workflow Repair task under the Workflow Reasoning category.

[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.

[Question]

Locate the workflow error in the displayed sequence and provide the minimum repair.

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Positions are one-based: Clip A = 1, Clip B = 2, Clip C = 3, and Clip D = 4. Replace {ans} with the actual answer: a JSON object containing "operation", "position", and "parameter". "operation" must be "insert", "swap", "delete", or "replace". Use "insert" when one or more required workflow stages are missing. Use "swap" when two adjacent Clips appear in reversed procedural order. Use "delete" when a workflow stage is duplicated or an unrelated Clip has been inserted. When two adjacent Clips show the same duplicated workflow stage, treat the later occurrence as the extra Clip. Use "replace" when an incorrect Clip occupies the position of a required workflow stage. "position" must be a JSON array. For "insert", list the two displayed positions surrounding the gap. For "swap", list the two positions to swap, in ascending order. For "delete" or "replace", list the single position to modify. "parameter" must be a JSON array. For "insert", list the missing stages in procedural order. For "replace", list the expected stage.
For "swap" or "delete", use [].

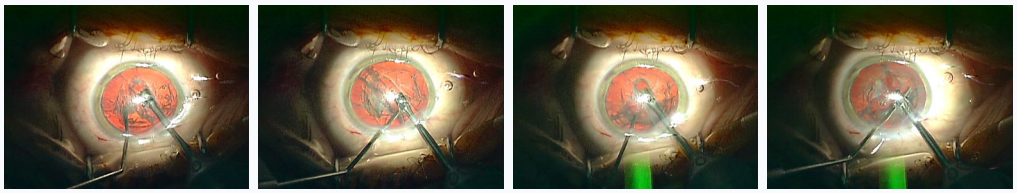
Figure 27 (continued): Example of the Workflow Repair task under the Workflow Reasoning category.

Description Correction

[Task]

You are solving a surgical video multi-field corruption and repair task. You must verify the procedure, phase, instruments, and anatomy in a structured description of the displayed Clip. Clip A contains four chronological frames. The described fields are consistent across all four frames; audit the whole Clip. The visual evidence for this question is provided below.

Clip A, frame 1 | Clip A, frame 2 | Clip A, frame 3 | Clip A, frame 4



[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.

[Question]

Audit the description below.
Description to audit:
{ "procedure": "Laparoscopic liver resection", "phase": "Irrigation/aspiration", "instruments": ["Spatula"], "anatomy": ["Cornea", "Pupil"] }

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Replace {ans} with the corrected Description. The answer must use exactly the same JSON structure as the Description to audit. Change only the incorrect values. "procedure" must be exactly one of: "Cataract surgery", "Laparoscopic cholecystectomy", "Laparoscopic liver resection". "phase" must be exactly one of: "Incision", "Viscoelastic", "Capsulorhexis", "Hydrodissection", "Phacoemulsification", "Irrigation/aspiration", "Capsule polishing", "Lens implantation", "Lens positioning", "Viscoelastic suction", "Anterior chamber flushing", "Tonifying/antibiotics". "instruments" must be a JSON array containing every visible instrument, selected from: "Phacoemulsification Tip", "Slit Knife", "Gauge", "Lens Injector", "Incision Knife", "Katena Forceps", "Capsulorhexis Forceps", "Capsulorhexis Cystotome", "Spatula", "Irrigation-Aspiration". Use [] when no instrument is visible, and list multiple instruments in the order shown above. "anatomy" must be a JSON array containing every visible anatomy category, selected from: "Cornea", "Lens", "Pupil". Use [] when none is visible, and list multiple categories in the order shown above.

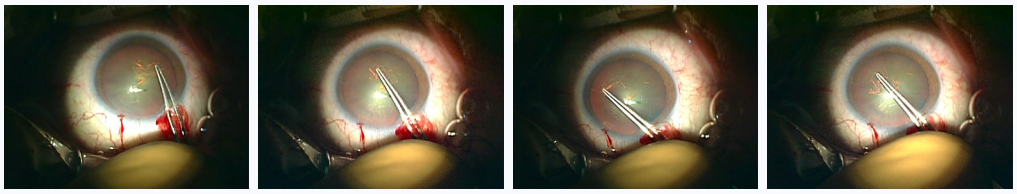
Figure 28: Example of the Description Correction task under the Description Correction category.

Description and Workflow Correction

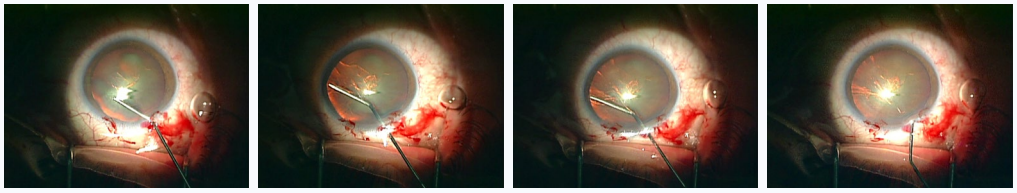
[Task]

You are solving a description and workflow correction task for cataract surgery. You must jointly audit the four displayed Clips as a local segment of the standard surgical workflow and audit the structured Claim paired with each Clip. The standard workflow is: 1. Incision 2. Viscoelastic 3. Capsulorhexis 4. Hydrodissection 5. Phacoemulsification 6. Irrigation/aspiration 7. Capsule polishing 8. Lens implantation 9. Lens positioning 10. Viscoelastic suction 11. Anterior chamber flushing 12. Tonifying/antibiotics Each Clip contains four chronological frames. All attributes covered by its Claim are consistent across those four frames; audit the whole Clip. The frames are successive annotated observations sampled approximately five seconds apart. The visual evidence for this question is provided below.

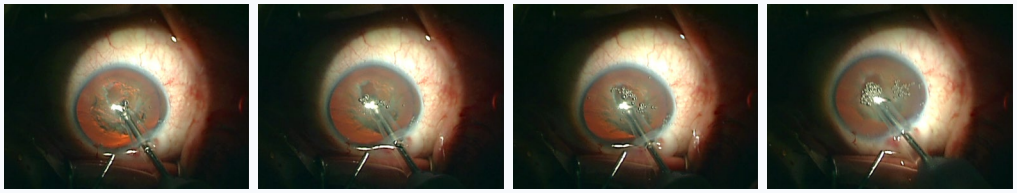
Clip A, frame 1 | Clip A, frame 2 | Clip A, frame 3 | Clip A, frame 4



Clip B, frame 1 | Clip B, frame 2 | Clip B, frame 3 | Clip B, frame 4



Clip C, frame 1 | Clip C, frame 2 | Clip C, frame 3 | Clip C, frame 4



Clip D, frame 1 | Clip D, frame 2 | Clip D, frame 3 | Clip D, frame 4

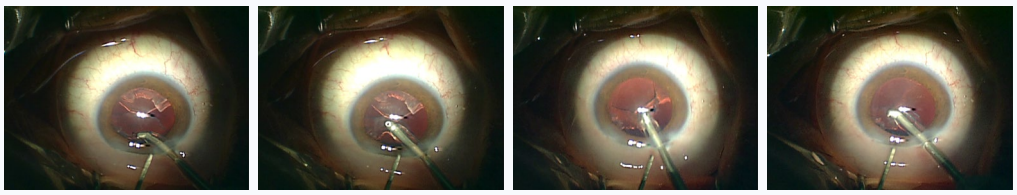


Figure 29: Example of the Description and Workflow Correction task under the Description and Workflow Correction category.

[Rules]

1. Use only the images and text provided in this prompt.
2. If answer options are provided, choose only from the provided options.
3. Follow the required answer format exactly.

[Question]

Each Claim is paired one-to-one with the Clip carrying the same label. Audit the displayed workflow and every Claim.
Claims:
Claim A (Clip A):
{ "phase": "Capsulorhexis", "instruments": ["Capsulorhexis Forceps"] }
Claim B (Clip B):
{ "phase": "Hydrodissection", "instruments": ["Gauge"], "spatial_relation": { "subject": "Gauge", "object": "Cornea", "relation_group": "vertical", "relation": "below" } }
Claim C (Clip C):
{ "phase": "Phacoemulsification", "instruments": ["Phacoemulsification Tip", "Spatula"], "spatial_relation": { "subject": "Phacoemulsification Tip", "object": "Cornea", "relation_group": "horizontal", "relation": "right_of" } }
Claim D (Clip D):
{ "phase": "Capsule polishing", "instruments": ["Spatula", "Irrigation-Aspiration"], "spatial_relation": { "subject": "Irrigation-Aspiration", "object": "Cornea", "relation_group": "vertical", "relation": "below" } }

[Answer Format]

Output exactly one JSON object:
{"think": "{think}", "answer": {ans}}
and nothing else. Replace {think} with your reasoning. Positions are one-based: Clip A = 1, Clip B = 2, Clip C = 3, and Clip D = 4. Replace {ans} with a JSON object containing only the required repairs. The only permitted keys are: "workflow", "Claim A", "Claim B", "Claim C", and "Claim D". Do not include a key when the corresponding workflow or Claim is correct. Use {} when no repair is needed. For a workflow error, use: "workflow": { "operation": "{operation}", "position": [{position}], "parameter": [{parameter}] } Use "insert" when one or more required workflow stages are missing. For "insert", give the two positions surrounding the gap and list the missing stages in "parameter". Use "swap" when two adjacent Clips appear in reversed procedural order. For "swap", give the two positions to swap and use [] for "parameter". Use "delete" when a workflow stage is duplicated or an unrelated Clip has been inserted. For "delete", give the position to delete and use [] for "parameter". Use "replace" when an incorrect Clip occupies the position of a required workflow stage. For "replace", give the position to replace and list the expected stage in "parameter". For an incorrect Claim, use its displayed Claim label as the key and return the complete corrected Claim. Use exactly the same JSON structure as the displayed Claim. Keep every correct value unchanged and modify only the incorrect values. For "phase", use exactly one of: "Incision", "Viscoelastic", "Capsulorhexis", "Hydrodissection", "Phacoemulsification", "Irrigation/aspiration", "Capsule polishing", "Lens implantation", "Lens positioning", "Viscoelastic suction", "Anterior chamber flushing", "Tonifying/antibiotics". For "instruments", return every annotated instrument category visible consistently across all four frames of the Clip. Use a JSON array containing zero or more of: "Phacoemulsification Tip", "Slit Knife", "Gauge", "Lens Injector", "Incision Knife", "Katena Forceps", "Capsulorhexis Forceps", "Capsulorhexis Cystotome", "Spatula", "Irrigation-Aspiration". In every displayed "spatial_relation", "subject", "object", and "relation_group" are correct. Only "relation" may be incorrect. For "horizontal", use "left_of" or "right_of". For "vertical", use "above" or "below". For "topological", use "overlapping" or "disjoint". List multiple values in the order shown in their corresponding Answer Space.

Figure 29 (continued): Example of the Description and Workflow Correction task under the Description and Workflow Correction category.