

MedGRPO: Multi-Task Reinforcement Learning for Heterogeneous Medical Video Understanding

Yuhao Su^{1,2,*} Anwesa Choudhuri^{2,†} Zhongpai Gao^{2,†} Benjamin Planche²
Van Nguyen Nguyen² Meng Zheng² Yuhan Shen¹ Arun Innanje²
Terrence Chen² Ehsan Elhamifar^{1,‡} Ziyang Wu^{2,‡}
¹Northeastern University, Boston, MA, USA
²United Imaging Intelligence, Boston, MA, USA

Abstract

*Large vision-language models struggle with medical video understanding, where spatial precision, temporal reasoning, and clinical semantics are critical. To address this, we first introduce **MedVidBench**, a large-scale benchmark of 531,850 video-instruction pairs across 8 medical sources spanning video, segment, and frame-level tasks, curated through a rigorous quality assurance pipeline with expert-guided prompting and dual-model validation. While supervised fine-tuning on MedVidBench yields noticeable gains, standard Reinforcement Learning (RL) fails due to imbalanced reward scales across datasets, which destabilizes optimization and leads to training collapse. To overcome this, we introduce **MedGRPO**, a novel RL framework for balanced multi-dataset training with two key innovations: (1) cross-dataset reward normalization that maps each dataset’s median performance to a common reward value, ensuring fair optimization regardless of difficulty, and (2) a medical LLM judge that evaluates caption quality on five clinical dimensions through comparative similarity scoring. Supervised fine-tuning Qwen2.5-VL-7B on MedVidBench outperforms GPT-4.1 and Gemini-2.5-Flash across all tasks, while MedGRPO further improves the SFT baseline on grounding and captioning. Our work establishes a foundational benchmark and training methodology for advancing medical video understanding with VLMs. Our project website is available at: <https://uii-america.github.io/MedGRPO/>.*

1. Introduction

Large vision-language models (VLMs) have demonstrated remarkable capabilities in understanding and reasoning

about general-domain visual content [2, 26, 30, 56]. However, their performance significantly degrades when applied to high-stake, expert-driven domains such as medicine. Medical video understanding presents unique challenges beyond general video understanding problems such as temporal modeling [47] and object grounding [48]: it requires interpreting fine-grained surgical actions, understanding domain-specific terminology (e.g., correctly identifying a “grasper” rather than generic “tool”), assessing procedural safety, and reasoning about multi-phase temporal workflows—requirements far beyond the everyday activities on which most models are trained.

Current state-of-the-art VLMs, including GPT-4.1 [19], Gemini-2.5-Pro [11], and Qwen2.5-VL [4], struggle to provide adequate performance on medical video tasks. The primary bottleneck is the lack of suitable training data in instruction-following format. While existing medical video datasets such as CholecT50 [39], EgoSurgery [14], AVOS [16], and NurViD [36] contain rich annotations, ranging from frame-wise bounding boxes, action triplets to phase labels, they are not designed for VideoLLM which requires question-answer pairs in conversational format. Recent efforts like SurgLaVi [42] and SurgLLM [10] remain limited in task diversity and primarily focus on surgical procedures. Converting existing medical video annotations into high-quality instruction-following QA pairs at scale presents a fundamental challenge: medical annotations require expert-level understanding to transform into natural language that preserves clinical accuracy.

We introduce MedVidBench, a large-scale instructional dataset comprising 532K video-instruction pairs across 8 medical sources and 8 diverse tasks spanning three temporal granularities: video-level understanding (summarization, critical view of safety, next action prediction, skill assessment), segment-level reasoning (temporal action grounding, dense captioning, region captioning), and frame-level grounding (spatiotemporal localization). The core innova-

*This work was carried out during the internship of Yuhao Su at United Imaging Intelligence, Boston MA, USA. †Corresponding authors.

‡Co-last authors.

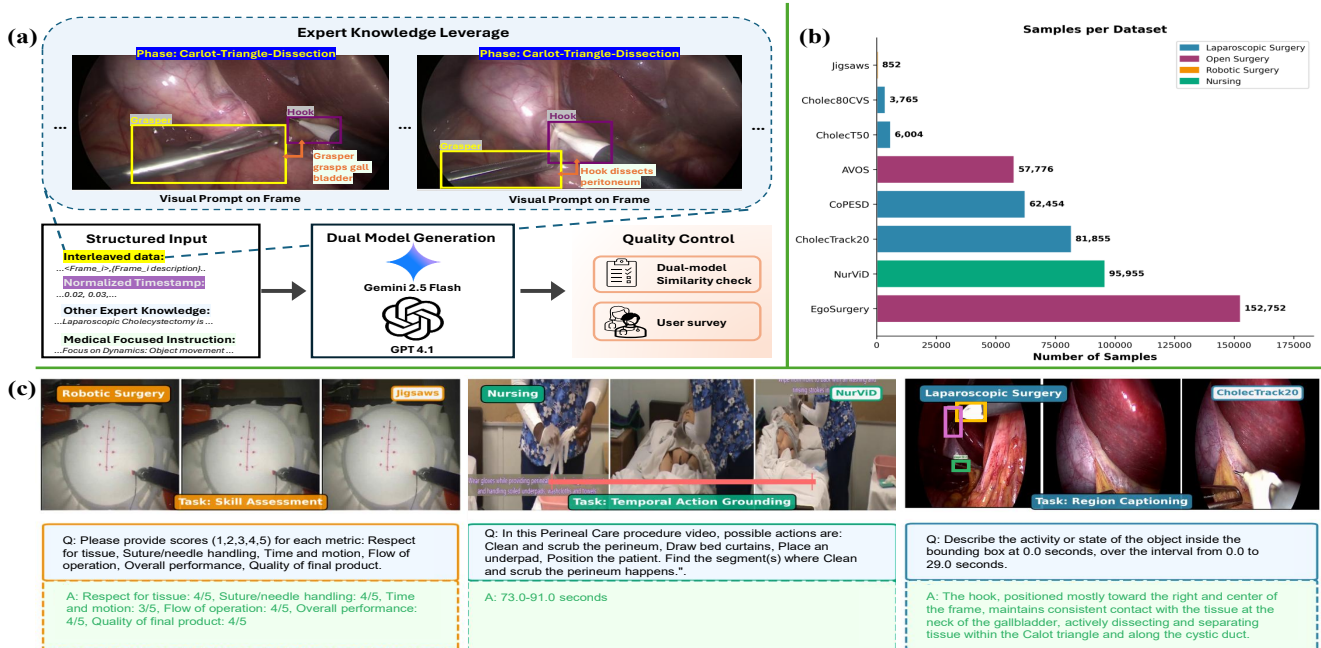


Figure 1. Overview of MedVidBench. (a) High quality data curation pipeline for MedVidBench. We leverage expert knowledge into prompt construction and generate high quality text using 2 VLMs (Gemini-2.5-Flash and GPT-4.1). (b) MedVidBench comprises of 8 different datasets, with 532k samples in total, spanning 4 different domains. (c) Examples of diverse tasks across different domains.

tion lies in our multi-perspective quality assurance pipeline that systematically converts existing expert annotations into instruction-following format. Rather than creating annotations from scratch, we leverage GPT-4.1 and Gemini-2.5-Flash to transform existing annotations—frame-wise bounding boxes, action triplets, phase labels—into natural language QA pairs. Our pipeline employs source-specific strategies: for datasets with dense frame-wise annotations (CholecT50, EgoSurgery), we use visual prompting by overlaying bounding boxes and labels on frames; for web-sourced videos (AVOS, NurViD), we extract high-quality transcripts using Whisper-X [5] and enrich prompts with video metadata (e.g., video titles). Dual-model validation ensures consistency and reduces generation errors.

While Supervised Fine-Tuning (SFT) a VLM on MedVidBench establishes a strong baseline, naively applying standard RL algorithms, such as GRPO [46], to our heterogeneous dataset to improve performance reveals a critical failure mode: training collapse. Models rapidly overfit to easy datasets—for instance, CoPESD [52] achieves median spatiotemporal grounding mIoU around 0.5—while performance on challenging datasets like EgoSurgery [14] (median mIoU around 0.12) degrades dramatically. This occurs because raw task metrics used as rewards create fundamentally unfair optimization: the model receives consistently higher rewards for easy dataset samples, causing gradient updates to prioritize easy sources while destabilizing learning on hard ones. A second challenge emerges in caption generation tasks: standard semantic similarity metrics fail

to capture medical correctness. Two captions may achieve high surface-level similarity yet differ critically in instrument specificity (“tool” vs. “grasper”), action precision (“grasps” vs. “dissects”), anatomical accuracy (“tissue” vs. “cystic duct”), and spatial detail (“upper area” vs. “upper right quadrant”).

We introduce MedGRPO, a reinforcement learning framework that addresses these challenges through two key innovations. First, cross-dataset reward normalization uses logistic functions centered on dataset-specific percentile statistics to map each dataset’s median performance to a fixed reward value, ensuring balanced optimization regardless of task difficulty. This normalization provides smooth gradients while being robust to outliers. Second, for caption generation tasks, we design a medical LLM judge that uses comparative similarity scoring to evaluate how closely generated captions match references across five clinical dimensions: medical terminology precision, instrument and anatomy identification, specificity versus vagueness, clinical procedure context, and action accuracy. This comparative framing avoids score inflation from absolute quality ratings. For reward computation during training, we combine LLM judge scores with semantic similarity in a hybrid design, leveraging both fine-grained clinical correctness and overall semantic coherence.

We first train Qwen2.5-VL-7B [4] using supervised fine-tuning (SFT) on MedVidBench and outperform closed-source alternatives including GPT-4.1 and Gemini-2.5-Flash. We then use reinforcement learning with MedGRPO

to train this SFT baseline and achieve substantial improvements over the SFT baseline across nearly all tasks while maintaining balanced learning across both easy and hard datasets. Our main contributions are:

- **MedVidBench**: A benchmark of 532K video-instruction pairs across 8 medical sources and 8 tasks spanning video/segment/frame-level understanding, curated through a multi-perspective quality assurance pipeline with source-specific strategies and dual-model validation.
- **MedGRPO**: A reinforcement learning framework enabling balanced multi-dataset training through (1) cross-dataset reward normalization using logistic functions for median-centered fairness, and (2) a medical LLM judge evaluating five clinical dimensions via by comparing how closely generated captions match references.
- **Comprehensive evaluation** demonstrating balanced improvements across heterogeneous datasets, superiority over closed-source models (GPT-4.1, Gemini-2.5-Flash), and ablations confirming that without reward normalization, training collapses with increasing entropy.

2. Related Work

Medical Video Datasets. Medical video understanding relies on specialized datasets capturing clinical procedures. Surgical datasets like CholecT50 [39] and EgoSurgery [14] provide frame-level annotations including action triplets and phase labels, but are limited in scale and diversity. Web-sourced datasets like AVOS [16] and NurViD [36] offer greater diversity with phase labels from which we extract audio, but lack frame-level precision. However, existing datasets provide annotations for traditional computer vision tasks rather than conversational question-answering format required by VideoLLMs, and remain isolated by procedure type. Our work systematically converts existing expert annotations into instruction-following format and enables joint training across 8 heterogeneous sources through reward normalization.

General Video LLMs. Large vision-language models have demonstrated impressive capabilities in general video understanding. Early frameworks like Flamingo [2] and BLIP-2 [26] established cross-modal alignment on images, with recent VideoLLMs like Video-ChatGPT [34], VideoLLaMA [61], and Video-LLaVA [28] extending these to videos through temporal modeling and video instruction tuning. State-of-the-art models including GPT-4.1 [19], Gemini-2.5-Flash [11], and Qwen2.5-VL [4] demonstrate strong performance on general-domain benchmarks [7, 57]. However, medical video understanding presents unique challenges: precise instrument identification, fine-grained action recognition, anatomical structure localization, and temporal reasoning about multi-step procedures.

Medical VideoLLMs. Medical video understanding

requires specialized knowledge beyond general vision-language models. Early medical VLMs focused on static radiology images (MedVInT [63], LLaVA-Med [25], Med-Flamingo [37]), while recent work addresses RGB surgical videos with temporal dynamics. SurgLaVi [42] introduces 240K surgical clip-caption pairs for CLIP-style contrastive pretraining but remains in descriptive format rather than instruction-following. Building on this, SurgLLM [10] fine-tunes Qwen2-VL on surgical instruction data from Cholec80 [50], a laparoscopic cholecystectomy dataset, demonstrating improved phase recognition and captioning. Similarly, Qwen2.5VL-7B_{Surg-CholecT50} [38] fine-tunes on CholecT50 for surgical workflow analyses. However, these approaches rely on single-dataset training, limiting cross-procedure generalization and task coverage, with no mechanisms to handle difficulty disparity across heterogeneous medical video datasets. Concurrent efforts advance medical AI through multimodal reasoning, surgical foundation models, and knowledge distillation [8, 9, 15, 17, 21–23, 49], yet multi-task learning in medical video understanding remains unexplored. Our work addresses this with a unified framework across 8 datasets and 8 tasks, enabling balanced learning via cross-dataset reward normalization.

Multi-Task Learning in VLMs. Multi-task learning leverages task synergies to improve generalization in vision-language models. Works like Unified-IO [33] and OFA [55] demonstrate that single transformers can handle diverse modalities through task-specific prompting. Multi-dataset training introduces optimization challenges including catastrophic forgetting [32, 35] and requires careful gradient management. Recent work applies RL to align VLMs with preferences [27, 60], but focuses on single-domain or single-task settings. Existing approaches either require architectural modifications (separate task heads [33], task-specific encoders [35], or task prompts [55]) or assume homogeneous difficulty distributions. In medical video understanding, no prior work addresses the challenge we identify: when datasets have vastly different difficulty (e.g., median performance 0.5 vs 0.12), standard RL causes training collapse due to reward magnitude disparities. Our cross-dataset reward normalization provides a simple yet effective solution without architectural changes, enabling balanced multi-dataset learning.

Evaluation Metrics for Video Captioning. Automatic evaluation of video captioning has evolved from n-gram metrics [6, 41, 51] to embedding-based approaches like BERTScore [62] and CLIPScore [18]. Recent LLM-as-judge methods [13, 20, 24, 29, 31] show strong correlation with human judgment, including in clinical settings [12]. However, medical video evaluation remains underexplored, and standard metrics fail to distinguish critical clinical differences like instrument specificity (‘tool’ vs. ‘grasper’),

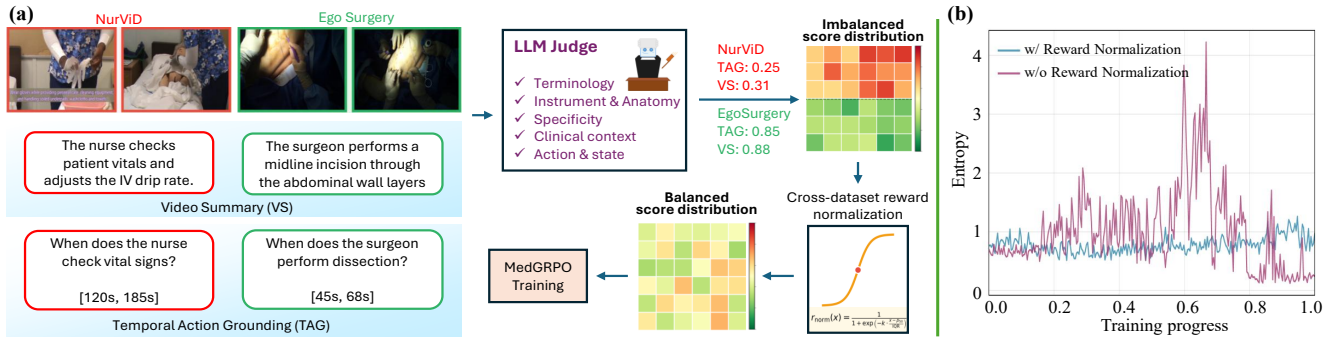


Figure 2. Overview of MedGRPO. (a) MedGRPO framework with cross-dataset reward normalization and medical LLM judge evaluation. (b) Training entropy comparison between models trained with and without reward normalization.

Table 1. Comparison with existing medical instruction datasets. Our MedVidBench provides superior scale, task coverage, multi-domain annotations, and quality assurance.

Dataset	Format	Domain	Tasks	Scale	Quality Assurance
SurgLaVi [42]	CLIP-style	Surgery	1	240K	Dual-model
Surg-396K [53]	Image-QA	Surgery	5	396K	Single model
SVU-31K [54]	Video-QA	Surgery	3	31K	Single model
MedVidBench (Ours)	Video-QA	Surgery & Instructional	8	532K	Dual-model validation

action precision (‘grasps’ vs. ‘dissects’), and anatomical accuracy (‘tissue’ vs. ‘cystic duct’). We introduce a medical LLM judge using comparative evaluation to assess caption quality on five clinical dimensions, combined with semantic similarity for robust RL training.

3. Method

We begin by introducing MedVidBench (§3.1), our large-scale benchmark created through a multi-perspective quality assurance pipeline (Figure 1). We then present MedGRPO (Figure 2), a reinforcement learning framework designed for fair multi-dataset training on MedVidBench: two-stage training paradigm (§3.2) and two core technical innovations of cross-dataset reward normalization (§3.3) that prevents catastrophic forgetting across heterogeneous datasets, and a medical LLM judge (§3.4) that captures domain-specific correctness beyond surface metrics.

3.1. MedVidBench: Multi-Granular Medical Video Benchmark

We construct MedVidBench, a unified benchmark spanning 8 medical video sources across 8 tasks with 531,850 video-instruction pairs in conversational QA format (Figure 1, Table 1). Our benchmark systematically transforms existing expert annotations—bounding boxes, procedure transcripts, action labels—into instruction-following format through a multi-perspective quality assurance pipeline with dual-model validation.

Data Sources and Task Coverage. MedVidBench

integrates 8 diverse medical video sources covering laparoscopic surgery (CholecT50 [39], CholecTrack20 [40], Cholec80-CVS [45], CoPESD [52]), open surgery (AVOS [16], EgoSurgery [14]), robotic surgery (JIGSAWS [1]), and nursing procedures (NurViD [36]). Our benchmark spans 626 unique videos with duration range 20s–1800s and adaptive FPS 0.1–3.0 optimized per task. We design tasks across three temporal granularities: 1) *Video-level tasks* include Video Summarization (VS), Critical View of Safety (CVS), Next Action Prediction (NAP), and Skill Assessment (SA), requiring holistic understanding of entire procedures; 2) *Segment-level tasks* include Temporal Action Grounding (TAG), Dense Video Captioning (DVC), and Region Captioning (RC), enabling event-level reasoning; 3) *Frame-level tasks* include Spatiotemporal Grounding (STG) for precise instrument localization. This multi-granular design enables models to understand *when* actions occur (temporal), *where* instruments are located (spatial), and *how* procedures should be performed (procedural context).

Multi-Perspective Quality Assurance Pipeline. Transforming existing medical annotations into instruction-following format requires expert-level precision (e.g., distinguishing “Maryland dissector” from generic “tool”). We develop a three-stage pipeline: (1) *Expert annotation prompting*: leverage dataset-specific strategies—for frame-annotated datasets (e.g., CholecT50, EgoSurgery), we overlay bounding boxes and labels directly on frames with procedure-specific context; for web-sourced datasets (AVOS, NurViD), we extract transcripts from high-quality audio using Whisper-X [5] and enrich with metadata; (2) *Dual-model generation*: generate captions independently with GPT-4.1 and Gemini-2.5-Flash, using both models for video summary and dense captioning tasks to prevent model-specific biases; (3) *Quality validation*: compute caption similarities between GPT-4.1 and Gemini-2.5-Flash outputs using sentence-transformers [44], filter low-quality pairs (similarity < 0.3), sample videos with 50–180 frames using adaptive FPS (0.1–3.0) suitable for different task re-

quirements, and create video-level train/test splits (test ratio 0.15). This rigorous pipeline produces 531,850 high-quality instances (461,413 training, 70,437 test).

Human Validation Study. We conduct user studies with medical professionals to empirically validate our expert annotation prompting approach. Medical experts compare captions generated using our annotation-enriched prompts (with overlaid bounding boxes, procedure context, and enriched transcripts) versus baseline generation from raw video frames only, rating on clinical accuracy and terminology precision. Results demonstrate significant preference for annotation-enriched generation, confirming that expert annotation prompting produces superior medical video QA pairs compared to naive frame-based generation. More dataset statistics, data curation details, LLM-judge prompts, and human study detail in the supplementary.

3.2. MedGRPO Training Paradigm

Our goal is to build a multi-task medical VideoLLM handling diverse tasks across heterogeneous datasets. We employ a two-stage approach: supervised fine-tuning (SFT) first adapts Qwen2.5-VL-7B to medical video understanding, injecting domain-specific knowledge while preserving the pretrained model’s general capabilities; this establishes baseline performance for percentile computation. Subsequently, reinforcement learning with GRPO enables multi-task improvement by aligning outputs with medical expertise through cross-dataset reward normalization (§3.3) and medical-specific LLM-judge evaluation (§3.4), without modifying the model architecture.

We apply GRPO [46], a policy gradient method that avoids value function training and naturally handles diverse reward scales through group-relative advantage estimation. For a given prompt q , we sample a group of $G = 8$ responses $\{o_i\}_{i=1}^G$ from our current policy π_θ . We compute the advantage \hat{A}_i for each response o_i by normalizing the response’s reward r_i within each group:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)} \quad (1)$$

The policy π_θ is optimized using the clipped surrogate objective against $\pi_{\theta_{\text{old}}}$ (the policy before the gradient update):

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E} \left[\sum_t \min \left(\rho_t \hat{A}_i, \text{clip}(\rho_t, 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}) \hat{A}_i \right) \right] \quad (2)$$

where $\rho_t = \pi_\theta(o_{i,t}|q, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})$ is the importance sampling ratio at token position t . Following DAPO [59], we adopt asymmetric clipping with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.3$ to allow larger positive updates while constraining negative ones, and remove the KL penalty term from standard GRPO for improved performance.

3.3. Cross-Dataset Reward Normalization

To enable fair multi-dataset multi-task optimization, we normalize rewards using dataset-task-specific logistic functions. While mapping rewards to $[0, 1]$ is common in GRPO, our key contribution is *stretching reward distributions* to have comparable spread across heterogeneous metrics. This addresses two sources of imbalance: datasets exhibit vastly different difficulty levels (e.g., CoPESD STG with median performance 0.5 vs. EgoSurgery STG at 0.12), and tasks use incomparable metrics (e.g., mIoU for grounding vs. our medical LLM judge scores (§3.4) for captioning). The key insight is *median fairness*: median-level performance receives equal normalized rewards across all dataset-task pairs, eliminating bias in gradient updates.

Normalization Function. For each dataset-task combination (d, t) , we apply logistic transformation to metric x :

$$r_{\text{norm}}^{(d,t)}(x) = \frac{1}{1 + \exp \left(-k \cdot \frac{x - p_{50}^{(d,t)}}{\text{IQR}^{(d,t)}} \right)} \quad (3)$$

where $p_{50}^{(d,t)}$ is the median, $\text{IQR}^{(d,t)} = p_{75}^{(d,t)} - p_{25}^{(d,t)}$ is the interquartile range, and $k = 3.0$ controls the sigmoid slope. The percentile statistics $\{p_{25}, p_{50}, p_{75}\}$ are computed from SFT baseline predictions on the training set for each dataset-task combination.

Design Properties. This design provides four advantages:

- (1) *Median fairness*—when $x = p_{50}^{(d,t)}$, the exponent becomes zero, yielding $r_{\text{norm}}^{(d,t)} = 0.5$ for all dataset-task pairs. For example, easy dataset-task pairs (CoPESD STG: $p_{50} \approx 0.5$) and hard pairs (EgoSurgery STG: $p_{50} \approx 0.12$) receive identical rewards at their respective medians, eliminating bias across both datasets and tasks.
- (2) *Smooth gradients*—the logistic function provides non-zero derivatives everywhere: $\frac{dr_{\text{norm}}}{dx} = \frac{k}{\text{IQR}} \cdot r_{\text{norm}}(1 - r_{\text{norm}})$, avoiding dead zones from hard clipping.
- (3) *Outlier robustness*—IQR-based scaling uses the middle 50% of the distribution, unlike min-max normalization sensitive to range extremes.
- (4) *Bounded output*—the logistic function maps to $(0, 1)$, compatible with GRPO’s group normalization, unlike unbounded z-score normalization.

Task-Specific Reward Design. We apply GRPO training to four representative tasks spanning different temporal granularities: VS (video-level), TAG (segment-level), RC (segment-level) and STG (frame-level). For grounding tasks (TAG, STG), we use multiplicative composite rewards as penalties rather than additive bonuses, since the model reliably outputs valid structures after SFT: $r_{\text{temporal}} = r_{\text{content}} \times r_{\text{format}}$, where r_{content} is logistic-normalized mIoU and $r_{\text{format}} = 1.0 - 0.6 \times (1 - \mathbb{I}_{\text{valid}})$ penalizes parsing failures. For captioning tasks (VS, RC), we combine semantic similarity and LLM judge evaluation (§3.4).

3.4. Medical LLM Judge for Evaluation

For caption generation tasks (VS, DVC, RC), standard embedding-based metrics (SentenceBERT [44], BERTScore [62]) capture overall paragraph-level semantic similarity but fail to assess fine-grained medical correctness. For example, captions “*The tool grasps tissue in the upper area*” vs. “*The grasper dissects the cystic duct in the upper right quadrant*” achieve cosine similarity ≈ 0.82 yet differ critically in *instrument specificity* (generic “tool” vs precise “grasper”), *action accuracy* (“grasps” vs “dissects”), *anatomical precision* (vague “tissue” vs specific “cystic duct”), and *spatial detail* (“upper area” vs “upper right quadrant”). Such distinctions are clinically significant but invisible to general-domain embeddings.

LLM Judge Design. We design a GPT-4.1-based judge using *comparative similarity scoring*: the judge evaluates “How closely does the generated caption match the reference?” rather than rating absolute quality. This comparative framing avoids score inflation [31] and provides better discrimination between model qualities. The judge evaluates five clinical dimensions with explicit 1–5 rubrics [34]: (1) *medical terminology precision* (clinical terms vs. lay language), (2) *instrument & anatomy identification* (specific tools and structures), (3) *specificity vs. vagueness* (precise details vs. generic descriptions), (4) *clinical procedure context* (workflow and safety awareness), and (5) *action & state accuracy* (surgical actions and tissue states). For each dimension, scores range from 1 (“completely different”) to 5 (“semantically equivalent”). We compute the mean score $\bar{s} = \frac{1}{5} \sum_{i=1}^5 s_i$ and apply logistic normalization: $r_{\text{LLM}} = r_{\text{norm}}(\bar{s})$.

Hybrid Evaluation Design. To capture both overall semantic coherence and fine-grained medical correctness, we combine semantic similarity and LLM judge evaluation. For caption generation tasks, we compute the final reward as an equal-weighted average of normalized semantic similarity and LLM judge scores. This design provides three advantages: (1) *complementary evaluation levels*—semantic similarity captures paragraph-level coherence while LLM judge assesses detail-level clinical correctness, (2) *medical accuracy*—domain-specific evaluation captures clinical nuances invisible to semantic metrics, (3) *robustness*—the 50% weight on semantic similarity ensures meaningful reward signals even when LLM evaluation occasionally fails.

4. Experiments

We conduct comprehensive experiments to evaluate MedGRPO across medical video understanding tasks. We demonstrate our approach achieves strong performance improvements over SFT baselines, enables dataset-fair optimization through reward normalization, and produces domain-specific improvements via LLM judge evaluation.

4.1. Experimental Setup

Datasets. MedVidBench (detailed in §3.1) provides two versions designed for different purposes. The **Large-Scale version** contains 531,850 samples (461,413 train and 70,437 test across 626 videos) leveraging all 8 medical datasets, maximizing available training data but with natural task imbalance favoring captioning tasks. This version is valuable for scaling law experiments and scenarios where maximum data utilization is prioritized. The **Standard version** contains 51,505 samples (45,260 train and 6,245 test across 611 videos, representing 9.81% of the large-scale version), carefully balanced across all tasks for efficient multi-task learning. Unless otherwise specified, all experiments use the Standard version for balanced evaluation.

Metrics. We evaluate across three temporal granularities. **Video-level tasks** include Video Summarization (VS) measured by LLM judge score; Critical View of Safety (CVS) assessed by accuracy; Next Action Prediction (NA) evaluated by accuracy; and Skill Assessment (SA) measured by accuracy. **Segment-level tasks** include Temporal Action Grounding (TAG) measured by mean IoU at thresholds 0.3 and 0.5; Dense Video Captioning (DVC) evaluated by LLM judge score and F1 score; and Region Captioning (RC) assessed by LLM judge score. **Frame-level tasks** include Spatiotemporal Grounding (STG) measured by mIoU.

Baselines. We compare our method against several baselines. **Off-the-shelf baselines** (GPT-4.1, Gemini-2.5-Flash, and Qwen2.5VL-7B) are evaluated using one-shot prompting with a format example but without any fine-tuning, performing significantly worse than fine-tuned models and demonstrating the difficulty of medical video understanding without domain adaptation. **Qwen2.5VL-7B_{SFT}** serves as our primary baseline, trained via supervised fine-tuning on MedVidBench. **Qwen2.5VL-7B_{MedGRPO}** represents our full method, which applies MedGRPO training to Qwen2.5VL-7B_{SFT} with cross-dataset reward normalization and medical LLM judge design. We further validate generalizability by applying the same SFT and MedGRPO pipeline to **Qwen3-VL-4B** [3], a smaller model with improved temporal modeling. We also compare against 2026 off-the-shelf models (**GPT-5.4**, **Gemini-3.1-Flash-Lite**), evaluated with the same one-shot prompting, and apply SFT to **Qwen3.5-4B** [3] to evaluate performance with the latest foundation models.

Implementation Details. We use Qwen2.5-VL-7B as the primary base model. During SFT training, video frames are sampled at 0.1-3 FPS with adaptive sampling for longer procedures. During GRPO training, frames are sampled at 1 FPS. Each frame contains $8 \times 28 \times 28$ to $48 \times 28 \times 28$ pixels. SFT training on the Standard version runs for 3 epochs with learning rate $5e-7$ and batch size 8, while scaling law experiments on the Large-Scale version use 1 epoch to demon-

Table 2. Main results on MedVidBench across 8 tasks. We compare off-the-shelf baselines from 2025 and 2026 (evaluated with one-shot prompting), our SFT baselines, and our full MedGRPO method on Qwen2.5VL-7B, Qwen3-VL-4B, and Qwen3.5-4B. We use accuracy for CVS/NAP/SA, mIoU for STG/TAG, LLM judge scores for DVC/VS/RC, and F1 score for DVC as metrics. The best scores within each group are highlighted with red and the second best with orange. MedGRPO results for Qwen3.5-4B will be added in a future update.

Model	CVS _{acc}	NAP _{acc}	SA _{acc}	STG _{mIoU}	TAG _{mIoU@0.3}	TAG _{mIoU@0.5}	DVC _{F1}	DVC _{llm}	VS _{llm}	RC _{llm}
2025 Off-the-shelf Baselines										
GPT-4.1 [19]	0.018	0.250	0.087	0.014	0.096	0.005	0.101	2.438	2.490	2.080
Gemini-2.5-flash [11]	0.101	0.228	0.107	0.047	0.045	0.021	0.084	2.387	2.352	1.912
VideoChat-R1.5-7B [58]	0.000	0.270	0.006	0.000	0.009	0.005	0.026	1.723	3.034	3.086
2025 Qwen2.5VL-7B										
Qwen2.5VL-7B [4]	0.105	0.151	0.010	0.020	0.006	0.068	0.075	2.512	2.452	2.090
Qwen2.5VL-7B _{Surg-CholecT50} (NVIDIA) [38]	0.000	0.302	0.000	0.000	0.019	0.013	0.051	1.945	2.101	2.986
Qwen2.5VL-7B _{SFT} (Ours)	0.894	0.442	0.218	0.177	0.142	0.091	0.165	3.665	3.596	2.757
Qwen2.5VL-7B _{MedGRPO} (Ours)	0.896	0.405	0.254	0.202	0.216	0.156	0.214	3.797	4.184	3.442
2025 Qwen3-VL-4B										
Qwen3VL-4B [3]	0.000	0.178	0.006	0.000	0.039	0.034	0.128	1.939	2.926	2.853
Qwen3VL-4B _{SFT} (Ours)	0.895	0.466	0.270	0.133	0.465	0.403	0.435	3.862	4.180	3.752
Qwen3VL-4B _{MedGRPO} (Ours)	0.898	0.473	0.285	0.176	0.504	0.441	0.480	3.950	4.227	3.861
2026 Off-the-shelf Baselines & Qwen3.5-4B										
GPT-5.4 [19]	0.164	0.393	0.267	0.004	0.086	0.055	0.178	3.403	3.976	3.714
Gemini-3.1-flash-lite [11]	0.242	0.406	0.225	0.059	0.072	0.049	0.174	3.198	3.737	3.492
Qwen3.5-4B [43]	0.309	0.231	0.276	0.051	0.074	0.040	0.142	2.699	3.491	3.037
Qwen3.5-4B _{SFT} (Ours)	0.897	0.576	0.354	0.190	0.482	0.429	0.451	3.741	4.238	3.746

Table 3. Ablation study on Qwen2.5VL-7B for reward normalization and LLM judge. Row A: our full MedGRPO method with four task rewards. Rows B-E isolate individual contributions. Notation: task combinations (e.g., TAG+STG) indicate which task rewards are used during GRPO training.

Model	CVS _{acc}	NAP _{acc}	SA _{acc}	STG _{mIoU}	TAG _{mIoU@0.3}	TAG _{mIoU@0.5}	DVC _{F1}	DVC _{llm}	VS _{llm}	RC _{llm}
A TAG+STG+VS+RC (full)	0.896	0.405	0.254	0.202	0.216	0.156	0.214	3.797	4.184	3.442
B TAG+STG+VS+RC w/o normal	0.020	0.267	0.234	0.010	0.004	0.003	0.002	1.061	3.805	3.469
C TAG+STG w/o VS+RC	0.914	0.394	0.257	0.193	0.202	0.142	0.225	3.718	3.776	3.425
D VS+RC w/ llm-judge	0.894	0.434	0.239	0.183	0.149	0.096	0.165	3.688	3.824	3.235
E VS+RC w/o llm-judge	0.894	0.363	0.259	0.183	0.140	0.090	0.161	3.628	3.733	2.984

strate performance improvements as training data increases from 0 to 461K samples (Figure 4). GRPO training uses 5000 gradient updates with learning rate $5e-7$ and batch size 5. We adopt technical improvements from DAPO [59] in our GRPO implementation with group size $G = 8$ and clipping parameters $\epsilon_{low} = 0.2$ and $\epsilon_{high} = 0.28$. Logistic normalization uses slope $k = 3.0$ and format penalty $\beta = 0.6$. All experiments run on $8 \times$ H100 GPUs.

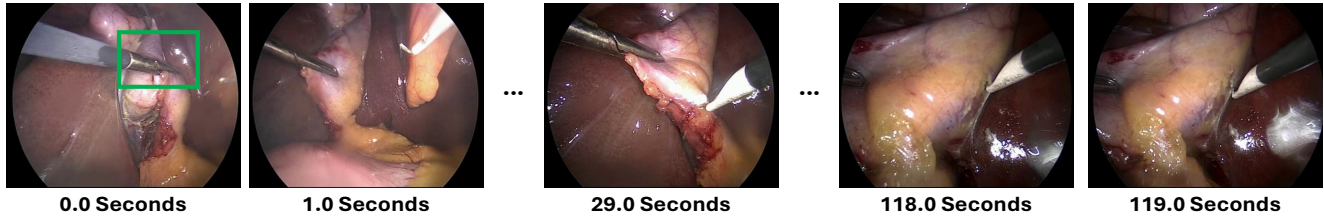
4.2. Main Results

Table 2 presents our main results across all 8 tasks. We compare against off-the-shelf baselines from 2025 (GPT-4.1, Gemini-2.5-Flash, VideoChat-R1.5-7B, Qwen2.5VL-7B, Qwen2.5VL-7B_{Surg-CholecT50}, and Qwen3-VL-4B) and 2026 (GPT-5.4, Gemini-3.1-Flash-Lite), our SFT baselines, and our full MedGRPO method on Qwen2.5VL-7B, Qwen3-VL-4B, and Qwen3.5-4B.

Comparison with Off-the-shelf Baselines. Our SFT baseline significantly outperforms all off-the-shelf baselines, demonstrating the importance of domain adaptation. GPT-4.1 achieves only 0.018 on CVS and 0.014 on STG, while Gemini-2.5-Flash reaches 0.101 on CVS and 0.047 on STG.

VideoChat-R1.5-7B, a recent RL-trained general-domain VideoLLM, scores 0.000 on CVS and STG, showing that general video RL training does not transfer to medical tasks. Off-the-shelf Qwen2.5VL-7B similarly performs poorly across grounding tasks (STG: 0.020, TAG@0.3: 0.006). Qwen2.5VL-7B_{Surg-CholecT50}, a recent medical VideoLLM fine-tuned on laparoscopic data, also fails to generalize (CVS: 0.000, STG: 0.000), highlighting that single-dataset surgical specialization is insufficient for the diverse tasks in MedVidBench. In contrast, our SFT baseline achieves 0.894 on CVS and 0.177 on STG, establishing the importance of large-scale multi-source training on MedVidBench.

MedGRPO Performance. Our full method (Qwen2.5VL-7B_{MedGRPO}) achieves consistent improvements over SFT across most tasks: CVS (+0.002 to 0.896), STG (+0.025 to 0.202), SA (+0.036 to 0.254), TAG@0.3 (+0.074 to 0.216), TAG@0.5 (+0.065 to 0.156), DVC_{llm} (+0.132 to 3.797), DVC_{F1} (+0.049 to 0.214), VS_{llm} (+0.588 to 4.184), and RC_{llm} (+0.685 to 3.442). NAP slightly decreases from 0.442 to 0.405, as it was not included in the reward optimization (we use TAG, STG, VS, and RC for computing rewards). Caption generation tasks benefit most from our hybrid LLM



Question: You are an expert surgical analyst specializing in minimally invasive procedures. This video shows an endoscopic view of a laparoscopic cholecystectomy, a keyhole surgery to remove the gallbladder. *Describe the activity or state of the object inside the bounding box at 0.0 seconds, over the interval from 0.0 to 29.0 seconds.*

GT: The grasper consistently grips and retracts the gallbladder towards the top left of the surgical field, providing counter-traction and exposure.

GPT-4.1: At 0.0 seconds and throughout the interval from 0.0 to 29.0 seconds, the endoscopic view depicts the initial stages of a laparoscopic cholecystectomy. The video shows surgical instruments dissecting tissues around the gallbladder. The operating surgeon is using these tools to carefully separate and manipulate tissues for a clearer view and safer dissection.

Gemini-2.5-flash: The electrocautery hook is actively used for the dissection of the gallbladder from the liver bed during the laparoscopic cholecystectomy. Over the interval, the instrument systematically advances along the gallbladder-liver interface, utilizing electrocautery to separate the gallbladder tissue and achieve hemostasis.

Qwen2.5VL-7B: The hook on the right side is dissecting tissue in the lower right hepatocystic triangle, making contact with the omentum and pericholecystic connective tissue. The hook is moving back and forth, carefully separating the tissues around the gallbladder.

SFT: The grasper on the right side consistently retracts the gallbladder or surrounding tissue to provide exposure throughout the sequence.

MedGRPO: The grasper, positioned primarily on the upper left, steadily holds and maintains exposure of the surgical field, retracting the gallbladder to facilitate dissection.

Figure 3. Qualitative comparison of region captioning generation.

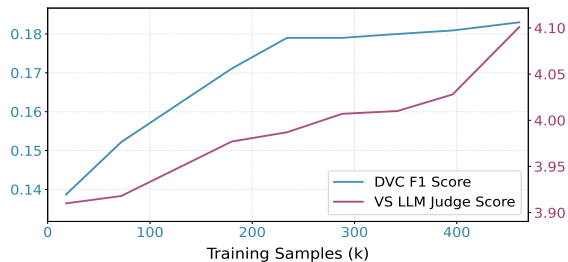


Figure 4. Scaling law analysis. Performance on Dense Video Captioning (DVC F1 score) and Video Summarization (VS LLM judge score) improves consistently as training samples from the Large-Scale version increase from 0 to 461K.

judge design, with VS and RC showing substantial improvements. Temporal grounding tasks also show strong gains, with TAG@0.3 and TAG@0.5 improving significantly.

Generalization to Qwen3-VL-4B. To validate framework generalizability, we apply the same SFT and MedGRPO pipeline to Qwen3-VL-4B [3], a smaller model with improved temporal modeling. Off-the-shelf Qwen3-VL-4B achieves near-zero performance on grounding tasks (STG: 0.000, CVS: 0.000), confirming that architectural advances alone cannot address medical video understanding without domain adaptation. Our SFT training yields strong gains (CVS: 0.895, TAG@0.3: 0.465, TAG@0.5: 0.403), and MedGRPO further improves across all tasks, with notable gains on STG (+0.043), TAG@0.3 (+0.039), and DVC_{F1} (+0.045). These results demonstrate that MedGRPO generalizes across model architectures and scales, consistently

improving upon SFT baselines regardless of the underlying model.

Comparison with 2026 Models. We further compare against the latest 2026 off-the-shelf models: GPT-5.4, Gemini-3.1-Flash-Lite, and Qwen3.5-4B. While these models show substantial improvements over their 2025 counterparts (e.g., GPT-5.4 reaches 0.164 on CVS vs. 0.018 for GPT-4.1, and 3.976 on VS vs. 2.490), they still fall significantly short of our fine-tuned models on grounding tasks (STG: 0.004, 0.059, and 0.051, TAG@0.3: 0.086, 0.072, and 0.074). Notably, off-the-shelf Qwen3.5-4B achieves 0.309 on CVS and 3.491 on VS, substantially outperforming its 2025 predecessor Qwen3-VL-4B (0.000 and 2.926), yet still far below our SFT models. Applying SFT to Qwen3.5-4B achieves the highest NAP accuracy (0.576) across all models and strong performance on grounding (TAG@0.3: 0.482, TAG@0.5: 0.429), further confirming that domain adaptation on MedVidBench remains essential even as foundation models advance. MedGRPO training on Qwen3.5-4B will be added in a future update.

Qualitative Analysis. Figure 3 illustrates representative improvements in region captioning generation. Ground Truth describes: “The grasper consistently grips and retracts the gallbladder towards the top left of the surgical field, providing counter-traction and exposure.” GPT-4.1 produces generic descriptions without specific instruments. Gemini-2.5-Flash misidentifies the tool as “electrocautery hook” with incorrect actions. Our SFT baseline identifies “grasper” but uses vague spatial terms (“right side”). Med-

GRPO generates clinically accurate captions: “The grasper, positioned primarily on the upper left, steadily holds and maintains exposure of the surgical field, retracting the gallbladder to facilitate dissection,” demonstrating precise instrument identification, accurate spatial localization, specific action descriptions, and clinical context understanding.

4.3. Ablation Studies

We conduct ablation studies on Qwen2.5VL-7B to validate each component.

Reward Normalization. Table 3 Row B shows that the removal of reward normalization causes catastrophic training collapse: performance drops from 0.894 (SFT baseline) to 0.020 in *CVS*, 0.177 to 0.010 in *STG* and 0.142 to 0.004 in *TAG@0.3*. Figure 2(b) visualizes this collapse through training entropy, showing that without reward normalization, entropy becomes highly unstable with dramatic spikes, while our reward normalization maintains stable entropy throughout training. This occurs because unnormalized rewards create high magnitude differences between dataset-task pairs, causing the optimizer to focus exclusively on high magnitude tasks while neglecting others. In contrast, Row A shows that our full method with reward normalization maintains balanced improvements across all tasks, confirming that proper reward scaling is essential for stable multi-dataset RL training.

Multi-Task Learning Benefits. Comparing Row A (full method with all four types of task) versus Row C (only *TAG+STG* without *VS+RC*), we observe that training with caption generation tasks improves grounding performance: *STG* improves from 0.193 to 0.202 (+4.7%), *TAG@0.3* from 0.202 to 0.216 (+6.9%), and *TAG@0.5* from 0.142 to 0.156 (+9.9%). This demonstrates beneficial task synergy in multi-task learning: caption generation requires understanding video content and temporal dynamics, which creates richer visual representations that benefit spatial-temporal localization tasks. Similarly, grounding tasks help captioning by providing better spatial awareness.

LLM Judge Design. Rows D-E isolate the contribution of our medical LLM judge for caption generation tasks. Row D (*VS+RC* with LLM judge) improves *VS* from 3.596 (SFT baseline) to 3.824 and *RC* from 2.757 to 3.235, demonstrating that domain-specific evaluation captures medical precision beyond semantic similarity. Row E (*VS+RC* without LLM judge) achieves only 3.733 on *VS* and 2.984 on *RC*, showing the importance of medical-specific evaluation. The full method (Row A) combining all four task rewards achieves the strongest performance: *VS* 4.184 and *RC* 3.442, improving 16.4% and 24.8% over SFT baseline.

5. Conclusion

We present MedVidBench, a large-scale heterogeneous benchmark for medical video understanding, and MedGRPO, a multi-dataset RL framework. We find that standard RL collapses on this data due to unbalanced reward scales across datasets. MedGRPO prevents this using two innovations: cross-dataset reward normalization to balance optimization, and a medical LLM judge for domain-specific evaluation. Our method substantially outperforms supervised fine-tuning, establishing that multi-dataset RL requires reward balancing and domain-adapted evaluation.

References

- [1] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017. 4, 15
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3
- [3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 6, 7, 8
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 3, 7
- [5] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023. 2, 4, 13
- [6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 3
- [7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 3
- [8] Chengan Che, Chao Wang, Xinyue Chen, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. A stitch in time: Learning procedural workflow via self-supervised plackett-luce ranking. *arXiv preprint arXiv:2511.17805*, 2025. 3
- [9] Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. Lemon: A large endoscopic monocular dataset and foundation model for perception in surgical settings. *arXiv preprint arXiv:2503.19740*, 2025. 3
- [10] Zhen Chen, Xingjian Luo, Kun Yuan, Jinlin Wu, Danny Chan, Nassir Navab, Hongbin Liu, Zhen Lei, and Jiebo Luo.

- Surgllm: A versatile large multimodal model with spatial focus and temporal awareness for surgical video understanding. *arXiv preprint arXiv:2509.00357*, 2025. 1, 3
- [11] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasapat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 3, 7, 13
- [12] Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640, 2025. 3
- [13] Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, 2024. 3
- [14] Ryo Fujii, Masashi Hatano, Hideo Saito, and Hiroki Kajita. Egosurgery-phase: A dataset of surgical phase recognition from egocentric open surgery videos. In *MICCAI*, 2024. 1, 2, 3, 4, 13, 15
- [15] Haozhen Gong, Xiaozhong Ji, Yuansen Liu, Wenbin Wu, Xiaoxiao Yan, Jingjing Liu, Kai Wu, Jiazhen Pan, Bailiang Jian, Jiangning Zhang, et al. Med-cmr: A fine-grained benchmark integrating visual evidence and clinical logic for medical complex multimodal reasoning. *arXiv preprint arXiv:2512.00818*, 2025. 3
- [16] Emmett D. Goodman, Krishna K. Patel, Yilun Zhang, William Locke, Chris J. Kennedy, Rohan Mehrotra, Stephen Ren, Melody Guan, Orr Zohar, Maren Downing, Hao Wei Chen, Jevin Z. Clark, Margaret T. Berrigan, Gabriel A. Brat, and Serena Yeung-Levy. Analyzing surgical technique in diverse open surgical videos with multi-task machine learning. *JAMA Surgery*, 2024. 1, 3, 4, 13, 15
- [17] Yongxin Guo, Hao Lu, Onur C Koyun, Zhengjie Zhu, Muhammet Fatih Demir, and Metin Nafi Gurcan. Momentum memory for knowledge distillation in computational pathology. *arXiv preprint arXiv:2602.21395*, 2026. 3
- [18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 3
- [19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 3, 7, 13
- [20] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023. 3
- [21] Qizhen Lan and Qing Tian. Acam-kd: Adaptive and cooperative attention masking for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3957–3966, 2025. 3
- [22] Qizhen Lan, Aaron Choi, Jun Ma, Bo Wang, Zhaoming Zhao, Xiaoqian Jiang, and Yu-Chun Hsu. From performance to practice: Knowledge-distilled segmentator for on-premises clinical workflows. *arXiv preprint arXiv:2601.09191*, 2026.
- [23] Qizhen Lan, Yu-Chun Hsu, Nida Saddaf Khan, and Xiaoqian Jiang. Reco-kd: Region-and context-aware knowledge distillation for efficient 3d medical image segmentation. *arXiv preprint arXiv:2601.08301*, 2026. 3
- [24] Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666, 2024. 3
- [25] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 3
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 3
- [27] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. VLFeedback: A large-scale AI feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246, 2024. 3
- [28] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 3
- [29] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 3
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1
- [31] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023. 3, 6
- [32] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 3
- [33] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Motlaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3
- [34] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 3, 6
- [35] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 3
- [36] Hu Ming, Wang Lin, Yan Siyuan, Ma Don, Ren Qingli, Xia Peng, Feng Wei, Duan Peibo, Ju Lie, and Ge Zongyuan. Nurvid: A large expert-level video database for nursing procedure activity understanding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1, 3, 4, 13, 15
- [37] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 3
- [38] NVIDIA. Qwen2.5-VL-7B-Surg-CholecT50. <https://huggingface.co/nvidia/Qwen2.5-VL-7B-Surg-CholecT50>, 2025. HuggingFace model card. 3, 7
- [39] 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, 2022. 1, 3, 4, 13, 15
- [40] Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L Lavanchy, and Nicolas Padoy. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8942–8952, 2025. 4, 13, 15
- [41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 3
- [42] Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. Surglavi: Large-scale hierarchical dataset for surgical vision-language representation learning. *arXiv preprint arXiv:2509.10555*, 2025. 1, 3, 4
- [43] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. 7
- [44] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 4, 6
- [45] Manuel Sebastián Ríos, María Alejandra Molina-Rodríguez, 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. 4, 15
- [46] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 5
- [47] Yuhao Su and Ehsan Elhamifar. Two-stage active learning for efficient temporal action segmentation. In *European Conference on Computer Vision (ECCV)*, pages 161–183. Springer, 2024. 1
- [48] Yuhao Su and Ehsan Elhamifar. Regionaligner: Bridging ego-exo views for object correspondence via unified text-visual learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3265–3274, 2026. 1
- [49] Ziyu Su, Abdul Rehman Akbar, Usama Sajjad, Anil V Parwani, and Muhammad Khalid Khan Niazi. Streamline pathology foundation model by cross-magnification distillation. *arXiv preprint arXiv:2509.23097*, 2025. 3
- [50] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016. 3
- [51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3
- [52] Guankun Wang, Han Xiao, Renrui Zhang, Huxin Gao, Long Bai, Xiaoxiao Yang, Zhen Li, Hongsheng Li, and Hongliang Ren. Copesd: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. *arXiv preprint arXiv:2410.07540*, 2024. 2, 4, 13, 14, 15
- [53] Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*, 2025. 4
- [54] Guankun Wang, Wenjin Mo, Junyi Wang, Long Bai, Kun Yuan, Ming Hu, Jinlin Wu, Junjun He, Yiming Huang, Nicolas Padoy, et al. Surgvidlm: Towards multi-grained surgical video understanding with large language model. *arXiv preprint arXiv:2506.17873*, 2025. 4
- [55] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 3
- [56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1
- [57] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3
- [58] Ziang Yan, Xinhao Li, Yanan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025. 7

- [59] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. [5](#), [7](#)
- [60] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. [3](#)
- [61] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multi-modal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. [3](#)
- [62] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. [3](#), [6](#)
- [63] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277, 2024. [3](#)

Supplementary Material:

MedGRPO: Multi-Task Reinforcement Learning for Heterogeneous Medical Video Understanding

This supplementary material provides comprehensive details on MedVidBench dataset construction and MedGRPO training methodology. §A describes our data curation pipeline, including prompt design strategies for web-sourced and frame-annotated datasets, QA generation procedures, and human validation study results. §B presents detailed dataset statistics covering task distribution, temporal characteristics, and annotation quality patterns. §C provides implementation details for SFT, GRPO training, and skill assessment evaluation. §D details our medical LLM judge rubrics across five clinical evaluation dimensions. §E presents additional qualitative results and failure analysis. All materials support reproducibility and provide insights beyond the main paper’s scope constraints.

A. Dataset Curation Pipeline

Our data curation pipeline transforms existing expert annotations from 8 medical video datasets into high-quality instruction-following format using dual multi-modal large language models (GPT-4.1 [19] and Gemini-2.5-Flash [11]). We employ a multi-perspective approach adapting prompting strategies to dataset characteristics: frame-annotated datasets (CholecT50 [39], EgoSurgery [14], CholecTrack20 [40], CoPESD [52]) receive rich contextual prompts incorporating frame-level annotations, while web-sourced datasets (AVOS [16], NurViD [36]) utilize high-quality audio transcripts (Whisper-X [5]) and video metadata to supplement visual understanding.

A.1. Prompt Design Principles

Our prompts follow a consistent six-component structure: (1) **role definition** establishing domain expertise (Expert medical analyst), (2) **background knowledge** providing procedure-specific context (anatomy, key structures, workflow), (3) **input data specification** enumerating available information (frames, timestamps, annotations), (4) **task definition** clarifying the objective (generate temporal summary, describe region), (5) **guiding principles** enforcing quality standards (be visually grounded, use precise terminology, avoid verbatim copying), and (6) **output format** specifying structure (one sentence, emphasize dynamics). We instantiate this template differently based on available annotations, as shown below.

A.2. Web-Sourced Datasets

For Web-sourced datasets, we compensate for limited expert annotations through multi-modal context integration. We enrich prompts with video metadata (title) and seg-

ment level annotation (action labels), high-quality temporal aligned ASR transcripts extracted using Whisper-X to provide comprehensive context. The prompt template is as below:

Role: Expert video analyst specializing in medical procedures

Background Knowledge:

- Video Title
- Video Description

Input Data:

- Frame_i – raw video frame
- Timestamp_i – normalized value in [0, 1]
- Action Label
- Transcript Segments: [timestamp interval: text]
- Context: action label

Task: Analyze frames and generate concise summary describing temporal evolution

Guiding Principles:

- Be Visually Grounded: Focus on observable events only
- Use Precise Naming: Specific surgical terminology
- Avoid Verbatim Copying: No prompt phrases in output
- Be Concise and Direct: No generic filler
- Focus on Dynamics: Object movement and instrument actions

Output Format: One sentence describing what happens over time, emphasizing motion, interaction, and anatomical changes.

A.3. Frame-Annotated Datasets

For datasets with rich expert annotations, we maximize information utilization through two complementary annotation strategies: (1) frame-text interleaved input. Frame-wise texts including triplet annotations (CholecT50) providing surgical action labels as (instrument, verb, target) triplets and textual descriptions (CoPESD) providing detailed per-frame narrative annotations. (2) bounding box visual prompts (CholecTrack20, EgoSurgery) overlaying spatial object locations with object labels directly on frames. The prompt template is as below:

Role: Expert surgical analyst

Background Knowledge: Per-surgery background info for each dataset, e.g., Laparoscopic Cholecystectomy

- Anatomy: Gallbladder anatomy and position
- Key Structures: Cystic duct, common bile duct, cystic artery
- Critical Landmark: Hepatocystic triangle (Calot triangle)

Input Data:

- (Interleaved) Frame_i – video frame with bounding box overlay and per-frame annotation
- Timestamp_i – normalized [0, 1]
- Action Label

Task: Analyze sequence and generate concise summary describing temporal evolution

Guiding Principles:

- Be Visually Grounded: Observable events only
- Use Precise Naming: Specific instrument names from annotations
- Use Preferred Verbs: Verbs from annotation vocabulary
- Be Concise and Direct: No filler
- Focus on Dynamics: Object movement, instrument causation

Output Format: One sentence describing what visibly happens over time

For regional captioning, we adapt the same prompt template by modifying the task specification to generate per-object descriptions with emphasis on spatial location and object-specific movements.

A.4. QA Generation and Quality Assurance

After caption generation, we create diverse QA pairs by combining dataset-specific context prefixes with task-specific question templates (3–6 template variants per task), producing instruction-following instances for all 8 tasks spanning video-level, segment-level, and frame-level understanding. To ensure annotation quality, we employ dual-model validation: for caption generation tasks (video summary, dense captioning, region captioning), we independently generate captions using both GPT-4.1 and Gemini-2.5-Flash, compute semantic similarity using sentence-transformers, and filter low-quality pairs with similarity <0.3. This dual-model approach prevents model-specific biases and hallucinations while ensuring consistent high-quality annotations. For evaluating caption quality during both dataset validation and RL training, we design a medical LLM judge (detailed rubrics in §D) that assesses captions across five clinical dimensions through comparative similarity scoring.

A.5. Human Validation Study

To validate our annotation-enriched prompting approach, we conducted a user study with 12 participants who are experts, work in medical data analysis, to compare captions generated using our expert prompts versus a frames-only baseline, both of which are described next. For CoPESD [52] dataset, we generated two types of captions: (1) **with expert prompt:** captions using our full pipeline with overlaid bounding boxes, procedure-specific context, and expert annotations; (2) **without expert prompt:** cap-

tions generated from raw video frames only using a minimal prompt (“Describe what you see in this healthcare procedure video in one sentence”), without procedural context, annotation overlays, timestamps, or domain knowledge. We developed a web interface, shown in Figure 5, to allow participants to rank the caption-pairs. Participants were provided with detailed instructions and examples to select the superior caption based on clinical accuracy and terminological precision. Figure 6 shows the results: participants strongly preferred captions generated with expert prompts (82.0%) over frames-only captions (18.0%), confirming that our annotation-enriched prompting strategy produces superior medical video descriptions compared to naive frame-based generation.

B. MedVidBench: Dataset Statistics

Task and Domain Distribution. Table 4 shows the statistical breakdown of MedVidBench by task and dataset. MedVidBench covers 8 dataset sources and 8 tasks spanning 4 domains: laparoscopic surgery (184.5K samples, 34.7%), open surgery (216.8K, 40.8%), robotic surgery (1.0K, 0.2%), and nursing (129.5K, 24.4%). The task distribution in MedVidBench reflects annotation granularity: frame-level annotations (e.g., spatial boxes) enable abundant region captioning samples (210.3K, 39.5%) as each frame contains multiple annotated regions, while specialized tasks at video level requiring expert holistic assessment remain rare (skill assessment: 1.0K, 0.2%, CVS: 4.4k, 0.8%). Segment-level tasks like temporal action grounding (158.5K, 29.8%) and dense captioning (73.3K, 13.8%) fall between these extremes.

Temporal Characteristics and Frame Sampling. Figure 7 (middle and right) shows MedVidBench exhibits substantial temporal diversity. Video durations range from 20 seconds to 1,800 seconds (30 minutes) with a median of 182 seconds and mean of 212 seconds, displaying a long-tail distribution where most videos fall within typical medical procedure segment lengths. Frame sampling rates vary from 0.1 to 3.0 FPS, with the majority of instances (63.3%) using 0.5 FPS, followed by 1.0 FPS (22.0%) and 2.0 FPS (7.5%). This distribution reflects two key factors: (1) source datasets have varying native frame rates, and (2) task-specific temporal requirements differ substantially. Video-level tasks (e.g. video summary) analyze longer durations and thus use low sampling rates to maintain manageable frame sequence lengths while capturing procedural evolution. This adaptive sampling strategy accommodates both dataset constraints and task-specific temporal granularity requirements.

Annotation Quality and Word Counting. Figure 7 (left) shows the answer length distribution ranges from 1 to 1,170

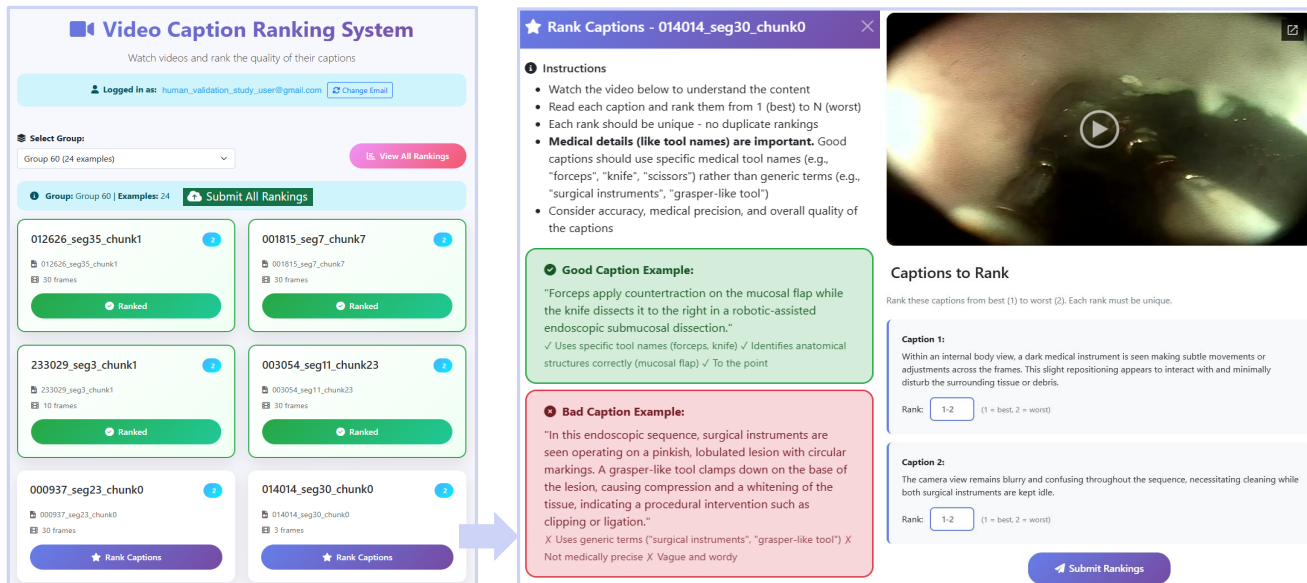


Figure 5. Interface for human validation study. Users were provided detailed instruction to rank caption after watching a short video. An instruction example for a good and bad caption was provided.

Table 4. MedVidBench statistics by dataset and task. Our benchmark covers 8 medical video sources with 532K video-instruction pairs across 8 tasks spanning video-level, segment-level, and frame-level understanding. Task abbreviations: VS (Video Summarization), SA (Skill Assessment), NAP (Next Action Prediction), CVS (Critical View of Safety), DVC (Dense Video Captioning), RC (Region Captioning), TAG (Temporal Action Grounding), STG (Spatiotemporal Grounding).

Dataset	Domain	Videos	Video-Level				Segment-Level			Frame-Level	Total
			VS	SA	NAP	CVS	DVC	RC	TAG	STG	
CholecT50 [39]	Laparoscopic	50	✓	-	✓	-	✓	-	✓	-	7.1K
CholecTrack20 [40]	Laparoscopic	20	-	-	-	-	-	✓	-	✓	102.7K
Cholec80-CVS [45]	Laparoscopic	80	-	-	-	✓	-	-	-	-	4.4K
CoPESD [52]	Laparoscopic	40	✓	-	✓	-	✓	✓	✓	✓	70.3K
AVOS [16]	Open Surgery	25	-	-	✓	-	✓	-	✓	-	62.5K
EgoSurgery [14]	Open Surgery	21	-	-	-	-	-	✓	-	✓	154.3K
JIGSAWS [1]	Robotic Surgery	103	-	✓	-	-	-	-	-	-	1.0K
NurViD [36]	Nursing	287	✓	-	✓	-	✓	-	✓	-	129.5K
Total samples		626	6.8K	1.0K	9.5K	4.4K	73.3K	210.3K	158.5K	68.0K	531.8K

words with a median of 21 words and mean of 41 words. Short answers (≤ 5 words, 28.1%) are predominantly from temporal action grounding tasks providing concise timestamps. Long answers (> 20 words, 51.8%) come mainly from descriptive tasks, with dense video captioning generating the longest responses due to detailed narration of multiple sequential actions, followed by region captioning describing surgical instrument movements. This distribution reflects the fundamental task heterogeneity in medical video understanding: grounding tasks require precise localization with minimal text, while captioning tasks demand rich and accurate clinical descriptions.

C. Implementation Details

Qwen2.5VL SFT Training. We use Qwen2.5-VL-7B-Instruct as our base model. Training is conducted on $8 \times$ H100 GPUs using distributed training with DeepSpeed ZeRO-3 offload. The per-device batch size is 6 with gradient accumulation steps of 1. We train for 3 epochs with differentiated learning rates: 5×10^{-7} for the language model, 1×10^{-6} for both the vision encoder and multimodal projector. A cosine learning rate scheduler is applied with a warmup ratio of 0.03. Weight decay is set to 0.01 and maximum gradient norm is clipped at 1.0. All training uses bfloat16 mixed precision. Video per-frame min and max pixels are set between $8 \times 28 \times 28$ to $48 \times 28 \times 28$ pixels.

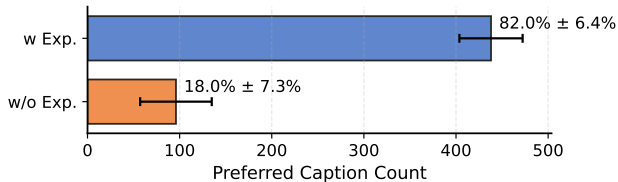


Figure 6. Human validation study results. User preference comparison with 12 participants on CoPESD dataset. “w/ Expert Prompt” refers to captions generated using our annotation-enriched prompting with overlaid bounding boxes, procedure context, and expert annotations. “w/o Expert Prompt” refers to captions generated from raw frames only with minimal prompting. Participants strongly prefer captions generated with expert prompts (82.0% vs 18.0%), validating our multi-perspective quality assurance pipeline.

We fine-tune all model components including the vision encoder, multimodal projector, and language model, and enable gradient checkpointing to reduce memory usage.

Qwen2.5VL GRPO Training. We implement GRPO training using the EasyR1 framework built on veRL. Training is conducted on $8 \times$ H100 GPUs with the SFT checkpoint as initialization. We use a group size of $G = 8$ responses per prompt with temperature 0.8 and top-p sampling at 0.95. The learning rate is 5×10^{-7} and maximum gradient norm clipped at 0.5. Videos are sampled at 1.0 FPS to simplify exploration during rollout. Following DAPO practices, we use asymmetric PPO clipping with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$ and disable KL divergence penalty.

Skill Assessment Evaluation. We average the 6 OSATS dimension scores and apply thresholds to derive 3 classes (Novice/Intermediate/Expert). We report macro-averaged accuracy (mAcc) and MAE (lower is better): zero-shot Qwen2.5VL-7B achieves MAE=2.440, mAcc=0.000; SFT improves to MAE=1.262, mAcc=0.197; MedGRPO further improves to MAE=1.246, mAcc=0.254, demonstrating the effectiveness of our method.

D. Medical LLM Judge

As described in §3.4 of the main paper, we employ an LLM-as-a-judge approach using GPT-4.1 to evaluate caption quality through comparative assessment across five medical-domain-specific dimensions. Each dimension uses a 1–5 scale measuring how closely the generated caption matches the reference: **5** (very close match, minor phrasing differences), **4** (good match, minor omissions), **3** (partial match, notable omissions), **2** (significant differences, missing important information), **1** (very different, major errors or missing content). §D.1 details how the LLM Judge

scores correlates with scores from board certified clinicians on these five dimensions. We describe these five dimensions in details in §D.2 with scoring rubrics.

D.1. Quality Assurance

To establish the validity of our LLM-as-a-judge approach using GPT-4.1 to evaluate caption quality, we conduct a rigorous human study with **10 board-certified clinicians**. The clinicians were asked to score 30 samples across the same 5 clinical dimensions used by our LLM Judge (paper lines 404-410). Results show strong correlation: Pearson $r=0.977$, Cohen’s Kappa =0.817, confirming our automated metric effectively proxies human clinical preference. We further show the correlation of the LLM Judge with clinicians across all the five evaluation dimensions ($R1-R5$) in Fig. 8. This highlights that the LLM Judge very closely agrees with clinicians in all of the evaluation dimensions.

D.2. Detailed Rubrics

These five evaluation dimensions used by the LLM Judge and board certified clinicians to score captions is detailed in this section.

Medical Terminology Precision (R1). *Definition:* Does the generated caption use the same medical terms as the reference?

Scoring Rubric:

- Score 5: medical terms match reference precisely (instruments, anatomy, actions)
- Score 4: most terms match reference, minor substitutions acceptable
- Score 3: some terms match reference, some generic or imprecise
- Score 2: many terms don’t match reference, often generic
- Score 1: terms mostly don’t match reference or are incorrect

Instrument and Anatomy Identification (R2). *Definition:* Are the instruments and anatomical structures identified the same as in the reference?

Scoring Rubrics:

- Score 5: all instruments and anatomy match reference identifications
- Score 4: most instruments and anatomy match reference
- Score 3: some instruments and anatomy match reference, some missing
- Score 2: many instruments and anatomy don’t match reference
- Score 1: instruments and anatomy mostly wrong or missing vs reference

Specificity vs Vagueness (R3). *Definition:* Is the level of specificity/vagueness similar to the reference?

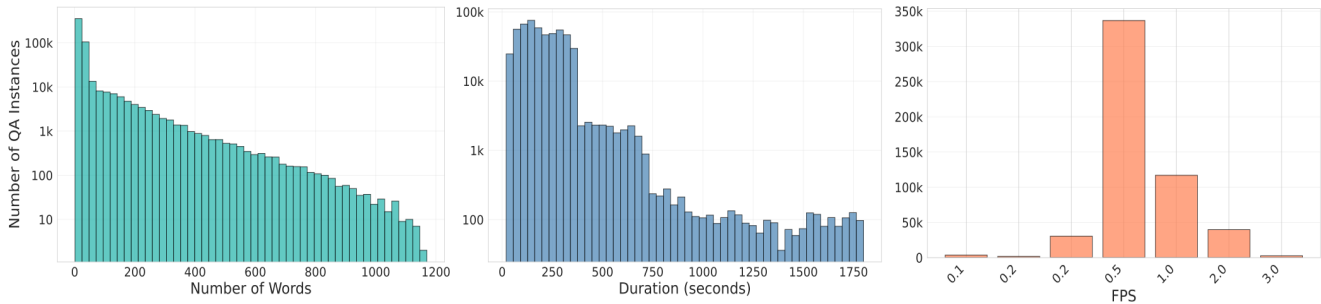


Figure 7. Dataset distribution analysis. Dataset distribution across 532K QA instances from 8 medical video datasets. (Left) Answer length distribution showing word counts ranging from 1 to 1,170 words (median: 21, mean: 41). Short answers (≤ 5 words, 28.1%) are predominantly from temporal action grounding tasks, while long answers (> 20 words, 51.8%) come mainly from dense video captioning and region captioning tasks. (Middle) Video duration distribution showing durations from 20 to 1,800 seconds (median: 182s, mean: 212s), exhibiting a long-tail pattern. (Right) FPS distribution showing that most instances use 0.5 FPS (63.3%), followed by 1.0 FPS (22.0%) and 2.0 FPS (7.5%). Left and middle panels use logarithmic scale on y-axis; right panel uses linear scale.

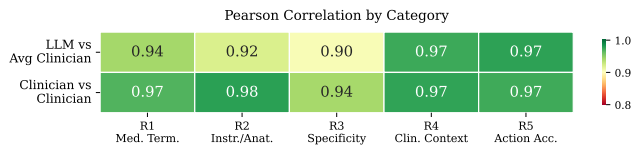


Figure 8. Pearson’s correlation between average clinician-clinician ratings and LLM-Clinician ratings across five evaluation dimensions (§ D.1). Experiment was performed with 10 board certified clinicians. The five dimensions are detailed in § D.2.

Scoring Rubrics:

- Score 5: specificity level matches reference (specific when reference is specific)
- Score 4: specificity level mostly matches reference
- Score 3: specificity level sometimes differs from reference
- Score 2: specificity level often differs from reference (too vague or too specific)
- Score 1: specificity level doesn’t match reference at all

Clinical Procedure Context (R4). *Definition:* Does the generated caption convey the same procedural understanding as the reference?

Scoring Rubrics:

- Score 5: procedural context matches reference (workflow, steps, purpose)
- Score 4: most procedural context matches reference
- Score 3: some procedural context matches reference, some missing
- Score 2: procedural context differs significantly from reference
- Score 1: procedural context mostly missing or wrong vs reference

Action and State Accuracy (R5). *Definition:* Are the actions and states described the same as in the reference?

Scoring Rubrics:

- Score 5: all actions and states match reference (active/idle, grasping/releasing, etc.)
- Score 4: most actions and states match reference
- Score 3: some actions and states match reference, some differ
- Score 2: many actions and states differ from reference
- Score 1: actions and states mostly wrong vs reference

E. Additional Qualitative Results

Task Examples. Figure 1 (c) in the main paper shows three tasks (Skill Assessment, Region Captioning, Temporal Action Grounding) across nursing, laparoscopic and robotic surgery. Figure 9 provides five additional examples (Dense Video Captioning, Spatio-Temporal Grounding, Critical View Safety, Video Summary, Next Action Prediction) spanning nursing, laparoscopic and open surgery. Together, these examples showcase MedVidBench’s coverage of 8 diverse tasks across 4 medical domains (nursing, robotic, laparoscopic and open surgery).

Sensitivity of Slope k : Figure 10 shows reward distributions for STG across datasets. **w/o normalization** (left), distributions are dramatically imbalanced—easy datasets yield consistently higher rewards than hard ones, causing training collapse. **w/ normalization** (right), all datasets achieve balanced, centered distributions. Crucially, varying $k \in \{2, 3, 4\}$ produces nearly identical normalized distributions, showing our method is relatively insensitive to k .

Dense Video Captioning. Figure 11 shows qualitative comparisons for dense video captioning on a Penicillin Skin



Figure 9. Examples of diverse tasks. 5 diverse tasks from MedVidBench (Dense Video Captioning, Spatio-Temporal Grounding, Critical View Safety, Video Summary, and Next Action Prediction) spanning 3 domains (Nursing, Laparoscopic Surgery and Open Surgery).

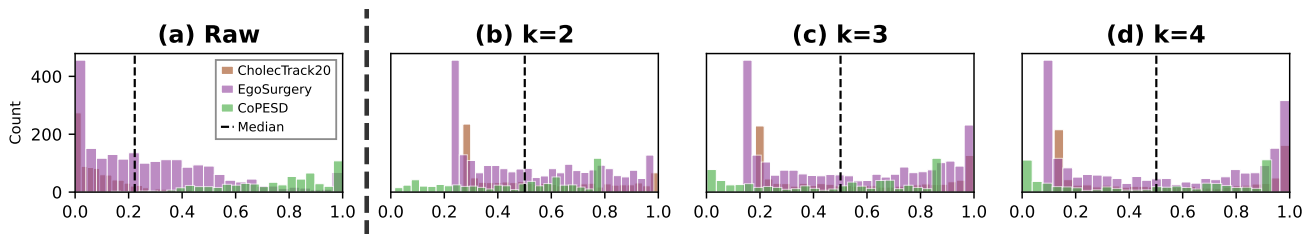
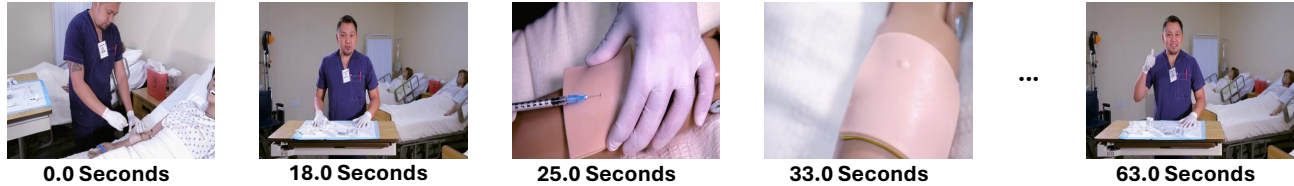


Figure 10. Reward distributions of w/o and w/ normalization $k \in \{2, 3, 4\}$.

Testing procedure. MedGRPO demonstrates superior performance over off-the-shelf models and SFT baseline: (1) **action identification**—correctly identifies both key actions (disinfect skin at 0.0–9.0s, perform intradermal puncture at 19.0–27.0s) matching ground truth (GT: 0.0–5.0s and 13.0–34.0s), while GPT-4.1 completely misses both actions and invents non-existent “prepare skin test solution” and “observe results”, Gemini-2.5-flash captures only partial disinfection (0.0–1.0s), and Qwen2.5VL-7B generates highly repetitive, non-specific descriptions across 23.0–58.0s without naming explicit actions; (2) **precise terminology**—uses specific clinical terms (“intradermal puncture”, “volar aspect of forearm”, “circular motions”, “shallow angle”, “epidermis”, “visible wheal”) matching GT vocabulary, versus Gemini-2.5-flash’s vague “cleaning a skin-like surface” and Qwen2.5VL-7B’s generic “holds a syringe...preparing

for injection”; (3) **technical detail**—captures procedural specifics including “alcohol swab”, “air-dry”, “needle at shallow angle into epidermis”, “brief retraction”, and “visible wheal formation”, closely aligning with GT descriptions; and (4) **temporal accuracy**—provides reasonable temporal boundaries with minor deviations (4s extension for disinfection, 6s shift for puncture), while SFT slightly underestimates disinfection duration (0.0–4.5s) and off-the-shelf models show severe temporal misalignment or excessive repetition.

Video Summary. Figure 12 presents qualitative comparisons for video summarization on a final part of laparoscopic cholecystectomy. MedGRPO produces a clinically accurate summary closely matching ground truth: grasper retracts gallbladder and positions specimen bag, bipolar



Question: You are an expert in healthcare workflow understanding. The footage captures a segment of a nursing procedure in a realistic clinical setting. The video shows part of the Penicillin Skin Testing procedure. Identify and describe ONLY segments where the video shows actions from this list: Check, Disinfect skin, Handwashing, Observe results of skin test, Perform intradermal puncture, Prepare skin test solution, Release trapped air. Skip all other content. **For each qualifying segment, give [start, end] and a concise description that explicitly names the action from the list.**

GT: 0.0-5.0 seconds: disinfect skin: A clinician wearing gloves retrieves an alcohol pad, then uses it to methodically swab the designated injection site on the patient's forearm model using circular motions, ensuring thorough antiseptic preparation of the skin surface. 13.0-34.0 seconds: perform intradermal puncture: The nurse dons gloves, removes the needle cap, and stretches the skin on the manikin's forearm before inserting the needle at a shallow angle; a close-up reveals needle advancement just beneath the epidermis, followed by the formation of a distinct bleb, after which the needle is withdrawn, leaving a visible wheal at the injection site.

GPT-4.1: 15.0-22.0 seconds: Prepare skin test solution.\n35.0-36.0 seconds: Observe results of skin test.

Gemini-2.5-flash: 0.0-1.0 seconds: Disinfect skin. A gloved hand is seen cleaning a skin-like surface with a white swab. Skin-like surface (likely a practice arm/mannequin) with a white pad.

Qwen2.5VL-7B: 23.0-26.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, indicating preparation for an injection.\n28.0-30.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, preparing for an injection.\n34.0-37.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, indicating preparation for an injection.\n40.0-42.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, preparing for an injection.\n44.0-46.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, indicating preparation for an injection.\n48.0-50.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, preparing for an injection.\n52.0-54.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, indicating preparation for an injection.\n56.0-58.0 seconds: A person wearing gloves holds a syringe on a mannequin's arm, preparing

SFT: 0.0-4.5 seconds: disinfect skin: A gloved hand applies antiseptic solution to the patient simulator's forearm using a circular scrubbing motion, preparing the skin for penicillin skin testing.

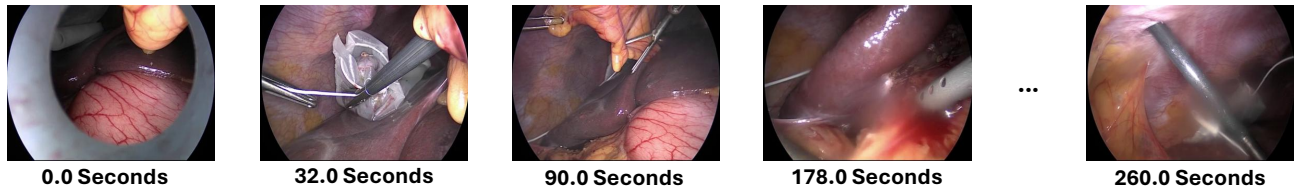
MedGRPO: 0.0-9.0 seconds: disinfect skin: Wearing gloves, the clinician uses an alcohol swab to thoroughly disinfect the volar aspect of the patient's forearm, performing circular motions over the designated injection site before withdrawing the swab to allow the skin to air-dry, preparing it for the penicillin skin test.\n19.0-27.0 seconds: perform intradermal puncture: Gloved hands position a syringe needle at a shallow angle into the epidermis of the forearm, followed by a brief retraction and lateral release of the plunger to create a visible wheal, after which the syringe is withdrawn, leaving the injected site exposed for observation.

Figure 11. Qualitative examples on dense video captioning.

instrument coagulates gallbladder bed, irrigator aspirates fluid for visualization, followed by specimen bag packaging and extraction. Key advantages: (1) **procedural accuracy**—correctly identifies the post-excision extraction phase matching GT (gallbladder positioning into specimen bag, liver retraction, omentum retraction, fluid aspiration, extraction), while GPT-4.1 hallucinates entire dissection workflow (“dissected and removed”, “dissects surrounding tissues”, “applies clips before cutting”) not present in video, and Qwen2.5VL-7B invents “scissors cutting the gallbladder from attachments” despite no cutting occurring; (2) **instrument identification**—accurately names specific instruments (grasper, bipolar instrument, irrigator) with their functions, whereas Gemini-2.5-flash uses generic terms (“laparoscopic instruments”) and GPT-4.1 describes non-existent “clips”; (3) **workflow completeness**—captures the full sequence (positioning, liver/omentum retraction, fluid aspiration, coagulation, extraction) matching GT, while SFT omits critical irrigation and fluid aspiration steps entirely; and (4) **temporal focus**—correctly focuses on specimen removal phase, unlike GPT-4.1’s fabricated early-stage dissection narrative spanning “66 frames at 0.25 FPS”.

Failure Analysis. Figure 13 illustrates a representative failure case on Critical View of Safety (CVS) assessment from Cholec80-CVS, evaluating three criteria with scores

0–2 per criterion. Ground truth scores: two structures (1), cystic plate (0), hepatocystic triangle (1). Despite strong performance on other tasks, all models struggle with CVS scoring: **Scoring calibration issues**—MedGRPO and SFT both score conservatively (0, 0, 0), systematically underestimating versus ground truth, while GPT-4.1 and Gemini-2.5-flash consistently overestimate (2, 2, 2), and Qwen2.5VL-7B shows erratic scoring (0, 2, 0). No model correctly identifies the intermediate ground truth pattern (1, 0, 1), suggesting fundamental difficulty in calibrating to surgical assessment rubrics rather than model-specific failure. **Intermediate score challenge**—with only 3.8K CVS training samples, models struggle to distinguish between partial achievement (score 1) versus absent (0) or complete (2), defaulting to extreme scores. CVS requires nuanced anatomical judgment: partial structure identification, subtle tissue plane visualization, and incomplete clearance assessment. **Multi-criteria reasoning**—simultaneous evaluation of three interrelated surgical safety criteria demands integrated anatomical knowledge and spatial reasoning that current models lack. Future work should explore specialized scoring calibration mechanisms, confidence-aware predictions for ambiguous cases, and expanded training data for underrepresented surgical evaluation tasks.



Question: You are an expert in surgical video interpretation. The following clip depicts a laparoscopic cholecystectomy, a minimally invasive operation for gallbladder removal, captured via an endoscopic camera. *Summarize the video content in a short paragraph, focusing on the main surgical activities.*

GT: Graspers initially position the gallbladder into a specimen bag and secure its enclosure. Subsequently, a grasper retracts the liver, followed by extensive omentum retraction and fluid aspiration using an irrigator. Finally, graspers maneuver the specimen bag containing the gallbladder by its closure string, progressively extracting it upwards and out of the abdominal cavity until the endoscope withdraws.

GPT-4.1: The video captures a laparoscopic cholecystectomy, where the gallbladder is dissected and removed in a minimally invasive procedure. Initial steps involve the insertion of surgical instruments and identification of the gallbladder's anatomy. The surgeon carefully dissects surrounding tissues, isolates the cystic duct and artery, and applies clips before cutting them. The gallbladder is then separated from the liver bed and retracted into a retrieval bag for removal from the abdominal cavity, concluding with inspection of the surgical site for hemostasis and closure of incisions. The procedure is performed over 66 frames at a frame rate of 0.25 FPS, offering a detailed view of each surgical step, emphasizing precision and care in this common operation to ensure patient safety and effectiveness of the removal process.

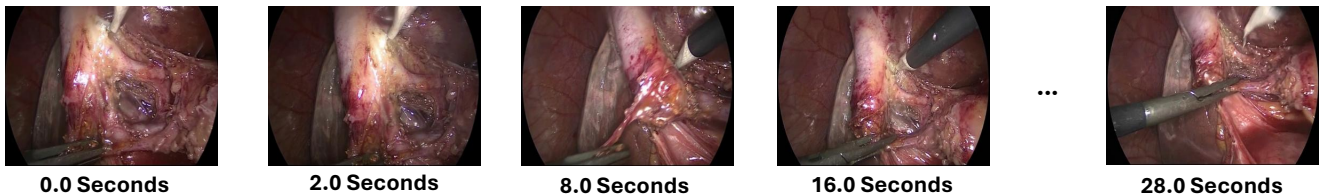
Gemini-2.5-flash: This laparoscopic cholecystectomy video demonstrates the final stages of the procedure. After the gallbladder has been excised, an endobag is introduced into the abdominal cavity. The resected gallbladder is then carefully maneuvered and placed inside this bag using laparoscopic instruments. Following this, the drawstring of the endobag is secured, and the bag containing the gallbladder is carefully extracted from the abdomen through a port site. The surgical field is then inspected, and the video concludes with the initiation of wound closure, specifically the suturing of a port site to ensure proper fascial closure.

Qwen2.5VL-7B: The video showcases a laparoscopic cholecystectomy where the surgeon uses various instruments to dissect and remove the gallbladder. Initially, a grasper retracts the liver, followed by another instrument retracting and manipulating the gallbladder. The surgeon then uses scissors to cut the gallbladder from its attachments, carefully separating it from the surrounding tissue. Throughout the procedure, the camera captures close-up views of the surgical site, highlighting the meticulous dissection and the use of forceps and scissors to ensure a clean separation of the gallbladder from the liver and other structures.

SFT: A grasper initially retracts the liver to expose the gallbladder fossa and surrounding tissue. Subsequently, graspers manipulate a specimen bag to fully enclose the gallbladder within it. Following this, a grasper retracts the liver while a bipolar instrument coagulates the liver surface. Finally, a grasper grasps the specimen bag containing the gallbladder, preparing it for extraction from the abdominal cavity.

MedGRPO: The grasper retracts the gallbladder superiorly and laterally to expose the surgical field, enabling the introduction and positioning of a specimen bag. Once secured, the grasper maintains retraction as the bipolar instrument coagulates the gallbladder bed and the irrigator aspirates fluid, ensuring clear visualization. The grasper then assists in packaging the gallbladder into the bag, which is subsequently withdrawn from the abdominal cavity.

Figure 12. Qualitative examples on video summary.



Question: You are an expert surgical analyst. The video comes from Cholec80-CVS and is for evaluating Strasberg's Critical View of Safety. For this laparoscopic cholecystectomy procedure, evaluate the Critical View of Safety based on the three essential criteria: proper identification of two structures, adequate cystic plate exposure, and complete hepatocystic triangle clearance. Provide scores (0,1,2) for each criterion.

GT: Two structures: 1, Cystic plate: 0, Hepatocystic triangle: 1.

GPT-4.1: Two structures: 2, Cystic plate: 2, Hepatocystic triangle: 2.

Gemini-2.5-flash: Two structures: 2, Cystic plate: 2, Hepatocystic triangle: 2.

Qwen2.5VL-7B: Two structures: 0, Cystic plate: 2, Hepatocystic triangle: 0.

SFT: Two structures: 0, Cystic plate: 0, Hepatocystic triangle: 0.

MedGRPO: Two structures: 0, Cystic plate: 0, Hepatocystic triangle: 0.

Figure 13. Failure case examples on Critical View of Safety (CVS) assessment.