

# 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 shows additional quantitative results on the baseline model Qwen3-VL. §D provides implementation details for SFT, GRPO training, and skill assessment evaluation. §E details our medical LLM judge rubrics across five clinical evaluation dimensions. §F 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 [18] and Gemini-2.5-Flash [11]). We employ a multi-perspective approach adapting prompting strategies to dataset characteristics: frame-annotated datasets (CholecT50 [38], EgoSurgery [14], CholecTrack20 [39], CoPESD [50]) receive rich contextual prompts incorporating frame-level annotations, while web-sourced datasets (AVOS [16], NurViD [35]) 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 segment 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<sub>i</sub> – video frame with bounding box overlay and per-frame annotation
- Timestamp<sub>i</sub> – 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 §E) 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 [50] 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:** captions 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

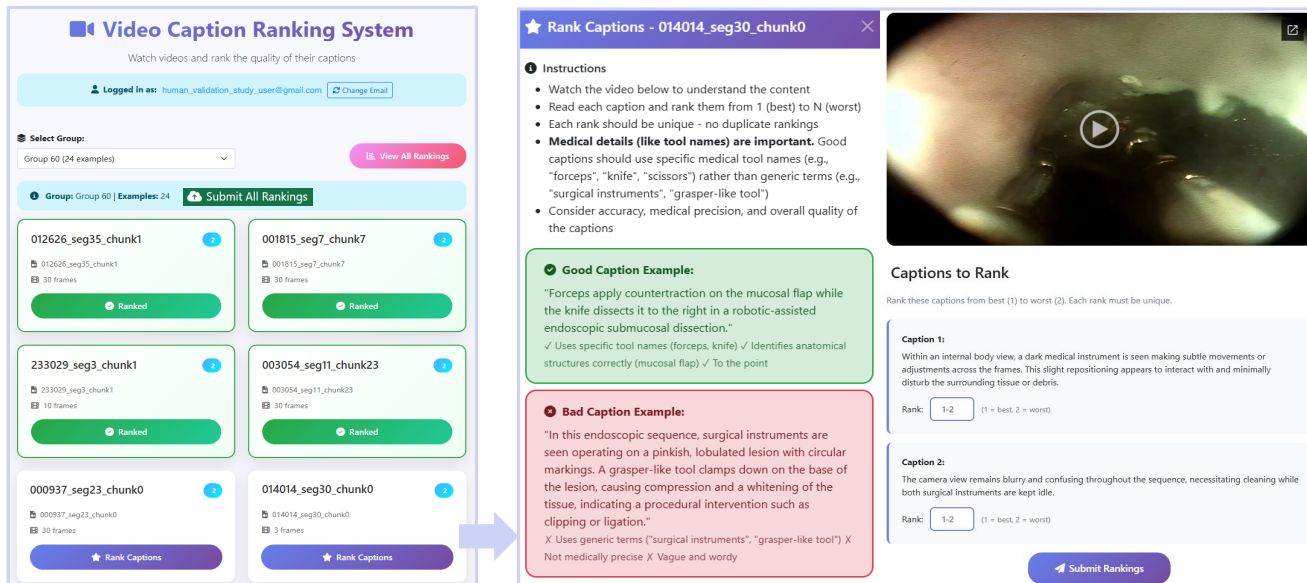


Figure 5. Interface for human validation study. Users were provided detailed instruction to rank caption after watching a short video. An instruction example for a good and bad caption was provided.

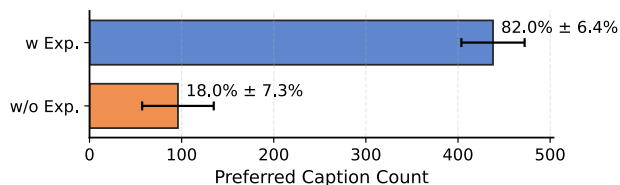


Figure 6. Human validation study results. User preference comparison with 12 participants on CoPESD dataset. “w/ Expert Prompt” refers to captions generated using our annotation-enriched prompting with overlaid bounding boxes, procedure context, and expert annotations. “w/o Expert Prompt” refers to captions 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.

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 words with a median of 21 words and mean of 41 words. Short answers ( $\leq 5$  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. Additional Quantitative Results

**Qwen3-VL.** Table 5 validates our framework’s generalizability by applying the same SFT and MedGRPO pipeline to Qwen3-VL-4B, 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<sub>F1</sub> (+0.045). These results demonstrate that MedGRPO generalizes across model architectures and scales, consistently improving upon SFT baselines regardless of the underlying model.

## D. Implementation Details

**Qwen2.5VL SFT Training.** We use Qwen2.5-VL-7B-Instruct as our base model. Training is conducted on 8× H100 GPUs using distributed training with DeepSpeed ZeRO-3 offload. The per-device batch size is 6 with gradi-

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	
<b>CholecT50</b> [38]	Laparoscopic	50	✓	-	✓	-	✓	-	✓	-	7.1K
<b>CholecTrack20</b> [39]	Laparoscopic	20	-	-	-	-	-	✓	-	✓	102.7K
<b>Cholec80-CVS</b> [43]	Laparoscopic	80	-	-	-	✓	-	-	-	-	4.4K
<b>CoPESD</b> [50]	Laparoscopic	40	✓	-	✓	-	✓	✓	✓	✓	70.3K
<b>AVOS</b> [16]	Open Surgery	25	-	-	✓	-	✓	-	✓	-	62.5K
<b>EgoSurgery</b> [14]	Open Surgery	21	-	-	-	-	-	✓	-	✓	154.3K
<b>JIGSAWS</b> [1]	Robotic Surgery	103	-	✓	-	-	-	-	-	-	1.0K
<b>NurViD</b> [35]	Nursing	287	✓	-	✓	-	✓	-	✓	-	129.5K
<b>Total samples</b>		<b>626</b>	<b>6.8K</b>	<b>1.0K</b>	<b>9.5K</b>	<b>4.4K</b>	<b>73.3K</b>	<b>210.3K</b>	<b>158.5K</b>	<b>68.0K</b>	<b>531.8K</b>

Table 5. Generalization to Qwen3-VL-4B on MedVidBench across 8 tasks. We apply the same SFT and MedGRPO training pipeline to Qwen3-VL-4B. Metrics: accuracy for CVS/NAP/SA, mIoU for STG/TAG, LLM judge scores for DVC/VS/RC, and F1 for DVC. Best in red, second best in orange.

Model	CVS <sub>acc</sub>	NAP <sub>acc</sub>	SA <sub>acc</sub>	STG <sub>mIoU</sub>	TAG <sub>mIoU@0.3</sub>	TAG <sub>mIoU@0.5</sub>	DVC <sub>llm</sub>	DVC <sub>F1</sub>	VS <sub>llm</sub>	RC <sub>llm</sub>
Qwen3VL-4B [3]	0.000	0.178	0.006	0.000	0.039	0.034	1.939	0.128	2.926	2.853
Qwen3VL-4B <sub>SFT</sub> (Ours)	<span style="background-color: #fff3cd;">0.895</span>	<span style="background-color: #fff3cd;">0.466</span>	<span style="background-color: #fff3cd;">0.270</span>	<span style="background-color: #fff3cd;">0.133</span>	<span style="background-color: #fff3cd;">0.465</span>	<span style="background-color: #fff3cd;">0.403</span>	<span style="background-color: #fff3cd;">3.862</span>	<span style="background-color: #fff3cd;">0.435</span>	<span style="background-color: #fff3cd;">4.180</span>	<span style="background-color: #fff3cd;">3.752</span>
Qwen3VL-4B <sub>MedGRPO</sub> (Ours)	<span style="background-color: #f8d7da;">0.898</span>	<span style="background-color: #f8d7da;">0.473</span>	<span style="background-color: #f8d7da;">0.285</span>	<span style="background-color: #f8d7da;">0.176</span>	<span style="background-color: #f8d7da;">0.504</span>	<span style="background-color: #f8d7da;">0.441</span>	<span style="background-color: #f8d7da;">3.950</span>	<span style="background-color: #f8d7da;">0.480</span>	<span style="background-color: #f8d7da;">4.227</span>	<span style="background-color: #f8d7da;">3.861</span>

ent accumulation steps of 1. We train for 3 epochs with differentiated learning rates:  $5 \times 10^{-7}$  for the language model,  $1 \times 10^{-6}$  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. 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  $5 \times 10^{-7}$  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.

## E. 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). §E.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 §E.2 with scoring rubrics.

### E.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

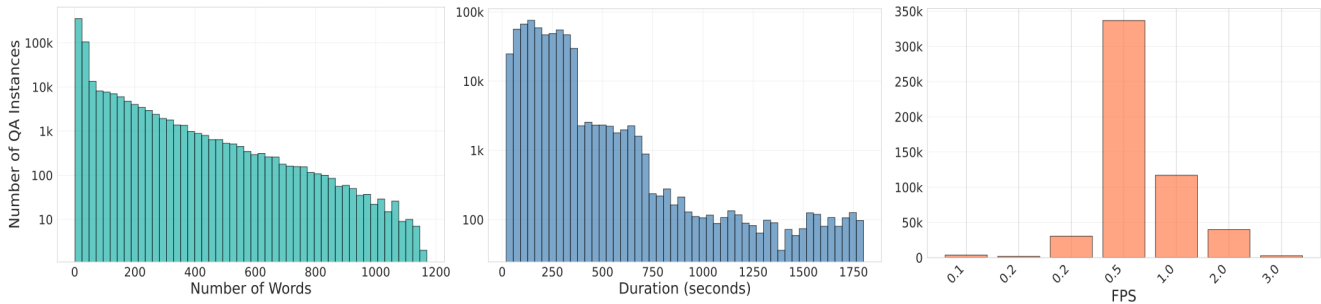


Figure 7. Dataset distribution analysis. Dataset distribution across 532K QA instances from 8 medical video datasets. (Left) Answer length distribution showing word counts ranging from 1 to 1,170 words (median: 21, mean: 41). Short answers ( $\leq 5$  words, 28.1%) are predominantly from temporal action 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.

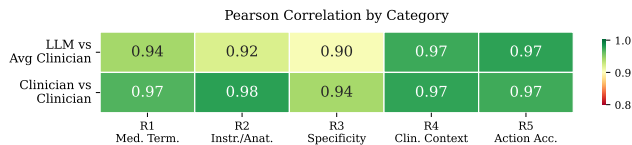


Figure 8. Pearson’s correlation between average clinician-clinician ratings and LLM-Clinician ratings across five evaluation dimensions (§ E.1). Experiment was performed with 10 board certified clinicians. The five dimensions are detailed in § E.2.

$r=0.977$ , 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.

## E.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?

*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



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).

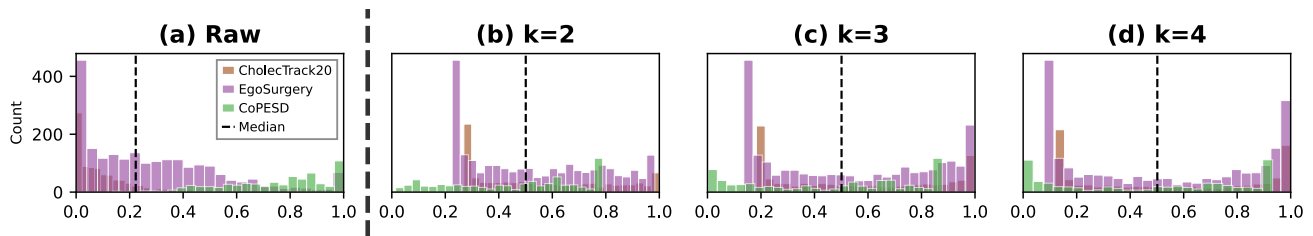


Figure 10. Reward distributions of w/o and w/ normalization  $k \in \{2, 3, 4\}$ .

- 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?

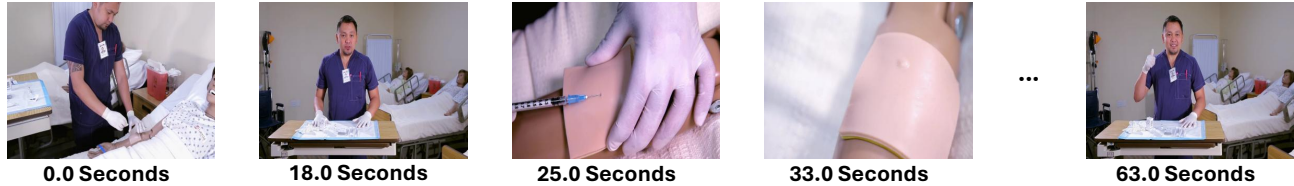
*Scoring Rubrics:*

- 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

## F. 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



**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.

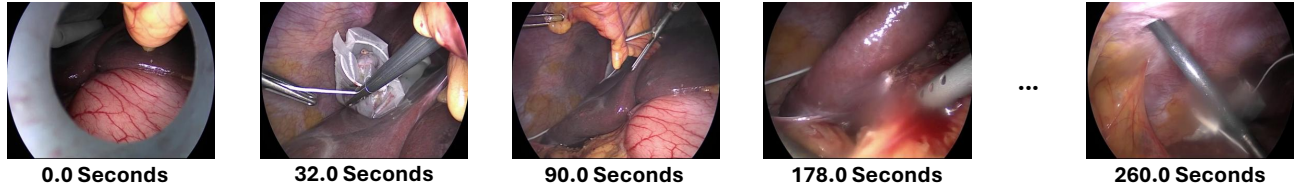
Figure 11. Qualitative examples on dense video captioning.

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 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 “vis-

ible 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 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



**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.

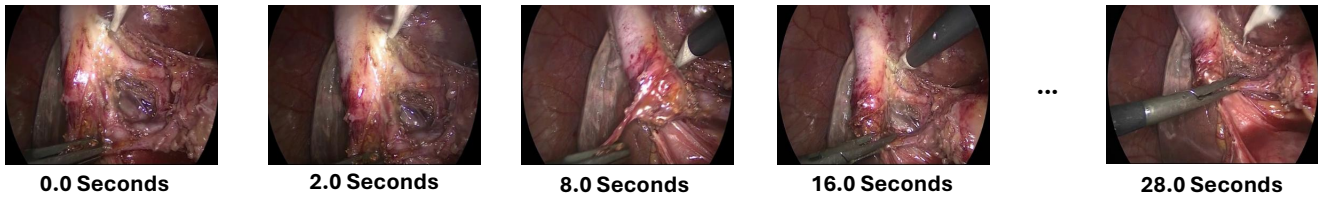
**Gemini-2.5-flash:** This laparoscopic cholecystectomy video demonstrates the final stages of the procedure. After the gallbladder has been excised, an endobag is introduced into the abdominal cavity. The resected gallbladder is then carefully maneuvered and placed inside this bag using laparoscopic 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.

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 en-

tirely; 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.