SiMing-Bench: Evaluating Procedural Correctness from Continuous Interactions in Clinical Skill Videos

Xiyang Huang[1,2], Jiawei Lin[2,3], Keying Wu[1,2], Jiaxin Huang[4], Kailai Yang[5], Renxiong Wei[6], Cheng Zeng[1], Jiayi Xiang [1,2], Ziyan Kuang[1,2], Min Peng[1,2,†], Qianqian Xie[1,2,†], Sophia Ananiadou[5]

[1]School of Artificial Intelligence, Wuhan University,
[2]Cognitive Language and Information Lab, Wuhan University,
[3]Southwest Jiaotong University, [4]MBZUAI, [5]The University of Manchester,
[6]Zhongnan Hospital of Wuhan University

Correspondence: pengm@whu.edu.cn, xieq@whu.edu.cn

Abstract

Current video benchmarks for multimodal large language models (MLLMs) focus on event recognition, temporal ordering, and long-context recall, but overlook a harder capability required for expert procedural judgment: tracking how ongoing interactions update the procedural state and thereby determine the correctness of later actions. We introduce SiMing-Bench, the first benchmark for evaluating this capability from full-length clinical skill videos. It targets rubric-grounded process-level judgment of whether interaction-driven state updates preserve procedural correctness across an entire workflow. SiMing-Bench is instantiated with SiMing-Score, a physician-annotated dataset of real clinical skill examination videos spanning cardiopulmonary resuscitation, automated external defibrillator operation, and bag-mask ventilation, each paired with a standardized step-wise rubric and dual-expert labels. Across diverse open- and closed-source MLLMs, we observe consistently weak agreement with physician judgments. Moreover, weak performance on rubric-defined intermediate steps persists even when overall procedure-level correlation appears acceptable, suggesting that coarse global assessment substantially overestimates current models' procedural judgment ability. Additional analyses with binary step judgment and step-aligned clips indicate that the bottleneck is not merely fine-grained scoring or temporal localization, but modeling how continuous interactions update procedural state over time.

1 Introduction

Can multimodal large language models (MLLMs) track interaction-driven world evolution in video? Although MLLMs are increasingly studied as world models (Ge et al., 2024; Sreeram et al., 2025), it remains unclear whether they can follow how interactions drive a scene from one state to the next over time. Evaluating this ability requires testing whether a model can infer how each action updates the current scene and constrains what can happen next (Chi et al., 2025). Clinical skill assessment[1] is where this ability becomes directly judgeable. In this task, expert examiners determine from video whether a trainee is performing a procedure correctly by interpreting how interactions with instruments, the patient, and the surrounding environment unfold over time. In suturing, for example, a stitching motion may look correct in isolation, yet an examiner can still rule the procedure invalid if the wrong instrument was used or the steps occurred out of order. At the scale of medical training and licensing, automating even part of this expert judgment could remove millions of hours of manual review while turning the same judgments into scalable, fine-grained feedback for clinical training.

However, current MLLM evaluations still mainly assess perception, event understanding, or temporal organization, rather than whether unfolding interactions update procedural states that determine later-step validity (as shown in Table 1). General video benchmarks such as E.T.Bench (Liu et al., 2024) and MVBench (Li et al., 2024b) evaluate what happens, in what order, and for how long, but do not test whether earlier interactions alter the conditions for subsequent steps to be correct, feasible, or clinically valid. Medical video benchmarks such as Surgpub-video (Li et al., 2025b) and SurgVLM (Zeng et al., 2025) extend evaluation to domain-specific perception, yet still focus on localized findings or instrument-level recognition, rather than interaction-conditioned procedural correctness. Spatial reasoning benchmarks (Li et al., 2025a; Yang et al., 2025), such as ViewSpatial and MMSI, further test positions, viewpoints, and 3D structure, but not whether action-induced scene changes constrain the validity of subsequent ac-

[†]Corresponding authors.

[1]More details in Appendix A

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 50070–50103
October 24-29, 2026 ©2026 Association for Computational Linguistics

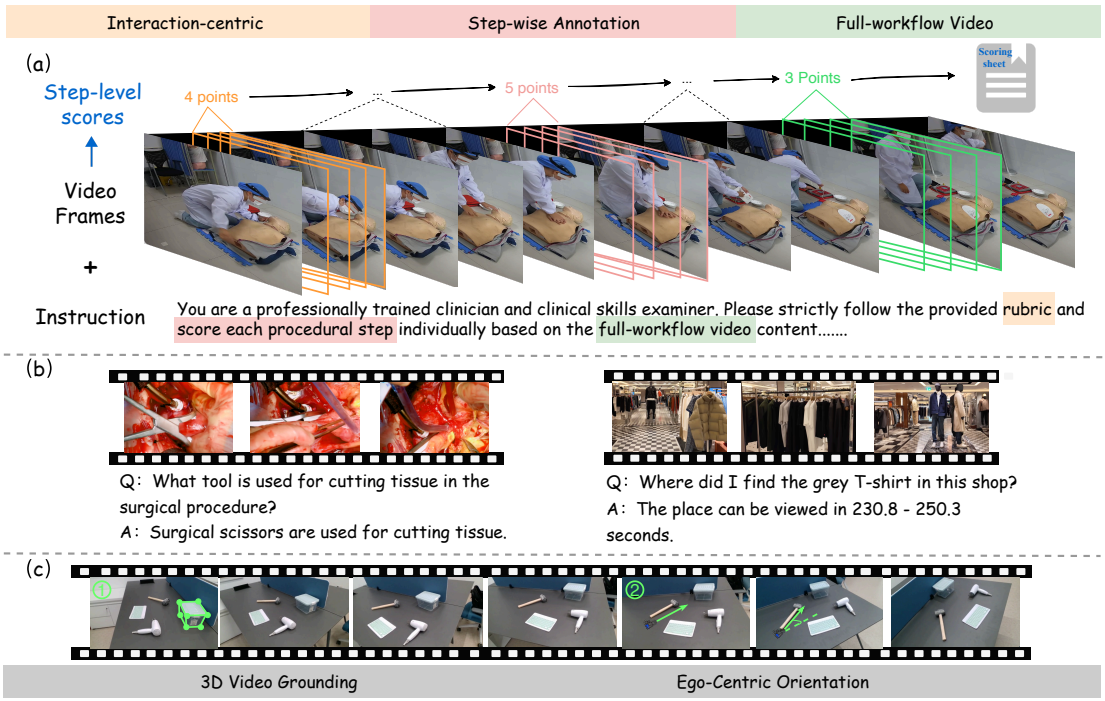


Figure 1: Comparison between SiMing-Bench and prior video benchmarks. (a) SiMing-Bench is built on full-workflow clinical skill videos and paired with expert-defined rubrics and step-wise scores, enabling evaluation of whether models can track continuous interactions throughout an entire procedure and make process-level procedural judgments. (b)(c) Existing benchmarks mainly evaluate localized video understanding, such as instrument recognition, temporal localization, and spatial reasoning.

tions in an ongoing procedure.

To address this gap, we introduce **SiMing-Bench**, the first benchmark built on clinical skill videos to test whether MLLMs can judge procedural correctness from continuous interactions in real-world environments[2]. It requires models to track interaction-driven state updates and make rubric-grounded procedural judgments over an entire workflow. SiMing-Bench is built on SiMing-Score, a manually annotated dataset of 200 full-length clinical skill videos spanning three skill categories, with a total duration of approximately 563 minutes and an average duration of 168.9 seconds per video. Each video preserves an entire workflow within a single scene. As shown in Figure 1 (a), collected from medical student examinations at three hospitals, the videos capture sustained interactions among the trainee, medical instruments, the patient, and the surrounding environment over time. Two physicians independently scored each video using standardized assessment forms, yielding step-level textual scoring criteria, ordinal ratings for multiple procedural steps, and an overall performance score. Given a full video and its scoring rubric, the task is to predict a discrete score for each predefined step as well as an overall score on the same scale used by human raters. Evaluation is conducted at two levels, with Pearson correlation(PLCC) and Spearman correlation(SRCC) measuring overall-score consistency and quadratically weighted Cohen's κ measuring step-level consistency.

Across 12 MLLMs with reported main results, we observe a persistent gap between model judgment and expert assessment. On overall procedure scoring, the highest observed PLCC is only 0.158, while still yielding zero step-level weighted Cohen's κ. In contrast, physician-physician agreement reaches 0.871 PLCC on overall scores and 0.732 Cohen's κ on step-level labels. This large gap shows that current MLLMs still struggle to convert continuous interactions in clinical videos into reliable procedural judgment. Moreover, weak overall-score correlation can coexist with near-zero agreement on rubric-defined intermediate steps, indicating that coarse global assessment substantially overestimates current models' process-level understanding. Additional diagnostic analyses further show

[2]The project is available at SiMing-Bench.

Benchmark	Domain	IC	FV	SA	IPJ	Major Evaluation Target
MVBench (Li et al., 2024b)	General Video	✗	✗	✗	✗	Temporal Understanding
Video-MME (Fu et al., 2025)	General Video	✗	✗	✗	✗	Compositional Video Understanding
LVBench (Wang et al., 2025a)	General Video	✗	✗	✗	✗	Long Video Understanding
EgoTempo (Plizzari et al., 2025)	Egocentric Video	✗	✗	✗	✗	Egocentric Temporal Understanding
OmniMMI (Wang et al., 2025c)	Streaming Video	✓	✗	✗	✗	Streaming Video Understanding
ScaleLong (Ma et al., 2025)	Long Video	✗	✓	✗	✗	Long-video Understanding
SurgVLM (Zeng et al., 2025)	Medical Video	✗	✗	✗	✗	Surgical Video Understanding
Surgpub-video (Li et al., 2025b)	Medical Video	✓	✗	✗	✗	Surgical Video Understanding
SiMing-Bench	Clinical Skill Video	✓	✓	✓	✓	**Interaction-driven State Validity Judgment**

Table 1: Comparison between SiMing-Bench and prior video benchmarks. **IC** denotes interaction-centric design, **FV** denotes full-workflow video, **SA** denotes step-wise annotation, and **IPJ** denotes interaction-driven procedural judgment. Together, these properties distinguish SiMing-Bench from prior benchmarks by enabling direct evaluation of interaction-driven state validity judgment in long, interaction-rich clinical workflows.

that this limitation is not resolved by simplifying the task. Under binary correctness judgment, the strongest observed result reaches only 0.132 MCC. Under manually segmented step-aligned clips, the highest observed step-level κ reaches only 0.075. Together, these results show that current MLLMs remain limited as world models because they do not yet reliably turn evolving interaction states into explicit judgments about procedural validity.

In summary, our main contributions are threefold. **First**, we introduce **SiMing-Bench**, the first benchmark built on clinical skill videos to test whether MLLMs can judge procedural correctness from continuous interactions. It targets a missing capability in current video evaluation, namely process-level judgment over long, interaction-rich procedures. **Second**, we formulate this problem as a multi-granularity assessment task and construct **SiMing-Score**, a manually annotated dataset of 200 full-length videos across three clinical skill categories. Each video is independently scored by two physicians with step-level scoring criteria, ordinal ratings, and an overall score. **Third**, experiments on diverse MLLMs reveal a persistent gap between model scoring and expert judgment. Further analyses show that this gap is not resolved by simplifying score granularity or temporal localization, suggesting a deeper limitation in how current MLLMs model interaction-driven procedural state.

2 Related Work

2.1 Video Understanding Benchmarks

Recent benchmarks have broadened MLLM evaluation for video understanding from event comprehension to temporal reasoning, long-context modeling, streaming perception, and open-ended multimodal understanding. Comprehensive benchmarks such as MVBench (Li et al., 2024b) and Video-MME (Fu et al., 2025) assess event recognition and multimodal comprehension across diverse domains, while LVBench (Wang et al., 2025a) and EgoTempo (Plizzari et al., 2025) emphasize long-video understanding and fine-grained temporal reasoning. OmniMMI (Wang et al., 2025c) and ScaleLong (Ma et al., 2025) further extend evaluation to streaming and hierarchical settings. Medical video benchmarks such as SurgVLM (Zeng et al., 2025) and Surgpub-video (Li et al., 2025b) move beyond frame-level recognition toward broader surgical intelligence. However, as shown in Table 1, existing benchmarks mainly focus on event recognition, temporal reasoning, or long-context understanding, rather than on whether continuous interactions progressively update scene states in ways that determine procedural correctness in subsequent steps throughout long workflows.

2.2 MLLMs for Video Understanding

Recent video MLLMs have shifted from video-adapted extensions to unified multimodal models with stronger long-context and spatiotemporal reasoning. LLaVA-Video (Zhang et al., 2024) and LLaVA-OneVision (Li et al., 2024a) extend general-purpose MLLMs to open-ended video understanding, while InternVL3.5 (Wang et al., 2025b), Qwen3-VL (Bai et al., 2025), Molmo2 (Clark et al., 2026), TimeSuite (Zeng et al., 2024), LLaVA-MR (Lu et al., 2024), and ReMoRa (Yashima et al., 2026) further improve long-context perception, grounding, temporal localization, and long-video modeling. Despite these advances, existing video MLLMs are still primarily optimized to understand what happened in a video, where it occurred, and how events unfold over

time. They are less explicitly designed to judge whether continuous interactions progressively update the scene state in ways that determine the correctness, feasibility, or validity of subsequent actions in a long procedure. This gap is closely related to recent vision-language world modeling studies (Ge et al., 2024; Chen et al., 2025), which move beyond event recognition toward modeling state transitions, action consequences, and temporally structured future evolution. However, it remains unclear whether current MLLMs can robustly track interaction-driven world-state updates for fine-grained procedural judgment, or instead rely on superficial sequential regularities in observed state changes.

3 SiMing-Bench

SiMing-Bench studies whether an MLLM can assess *procedural correctness* from a complete clinical skill video under an expert-defined scoring rubric. Unlike conventional video understanding tasks that ask what happened in a clip, the task requires mapping a long-form procedure to rubric-aligned judgments over clinically meaningful steps. Each step corresponds to a temporally extended interaction segment rather than a single visual event, and its correctness depends on whether the preceding interactions have established the procedural conditions required for the current action. The model must therefore evaluate the procedure as an evolving interaction process, rather than as a collection of isolated observations.

3.1 Task Definition

We formulate this setting as a structured prediction problem over long-form videos. Given a complete clinical skill video $V = \{V_t\}_{t=1}^{T}$, and a step-wise rubric $R = \{r_i\}_{i=1}^{N}$, the task requires a model F_θ predicting a sequence of step-level scores:

$$\hat{\mathbf{s}} = \{\hat{s}_i\}_{i=1}^{N} = F_\theta(V, R).$$

where V_t denotes the visual observation at time step t, r_i specifies the scoring criteria and point rules for step i. Here, each $\hat{s}_i \in \mathcal{Y}_i$ is a discrete score for step i, and $\mathcal{Y}_i$ is the valid score set defined by the rubric for that step. The overall procedure score is obtained by aggregating the predicted step-level scores:

$$\hat{y} = \sum_{i=1}^{N} \hat{s}_i,$$

Skill	Avg. Duration	#Rubric Steps
CPR	85.6 s	19
AED	50.2 s	9
BMV	33.1 s	5
Total videos	200	
Total duration	563 min	
Avg. duration per sample	168.9 s	
Total rubric steps	33	

Table 2: Dataset statistics of SiMing-Bench. Each video contains a complete assessment including CPR, AED, and BMV, rather than a single isolated skill. We report the average duration and rubric steps for each skill, together with the overall dataset scale after quality control.

Metric	**Value**
Overall-score agreement (PLCC)	0.871
Step-level agreement (Cohen's κ)	0.732
Step instances requiring adjudication	65 (0.98%)
Videos excluded during curation	33

Table 3: Annotation agreement and curation statistics before adjudication.

where $\hat{y}$ denotes the predicted total score for the full procedure.

3.2 Dataset Curation

Data source and collection. SiMing-Score is curated from real clinical skill assessment videos of medical students at three hospitals, forming, to our knowledge, the first full-length clinical skill video benchmark for rubric-grounded procedural assessment. Videos were collected in standardized training and examination settings, where students performed complete procedures under expert-defined protocols rather than expert demonstrations.

We collected 233 videos in total, each covering cardiopulmonary resuscitation (CPR), automated external defibrillator (AED) operation, and bag-mask ventilation (BMV), with average durations of 85.6s, 50.2s, and 33.1s, respectively. After quality control, the dataset contains approximately 563 minutes of video, averaging 168.9s per sample. Each video records one complete attempt within a single testing scene, preserving the full procedural trajectory rather than isolated clips or manually segmented events. Detailed statistics are provided in Table 2. The data collection and research protocols were approved by the Institutional Review Board (IRB)[3].

Because the performers are students in authentic assessments, the videos naturally contain clin-

[3]More details in Appendix H

ically meaningful variation in procedural quality, including incorrect ordering, incomplete execution, unstable coordination, and other deviations that affect rubric-based scoring. Thus, correctness must be judged from the full interaction process rather than assumed. The three skill categories instantiate complementary forms of long-horizon procedural interaction[4].

For each skill, expert rubrics define clinically meaningful sub-steps and explicit execution requirements, including action order, continuity, device handling, safety constraints, and coordination quality. All recordings were obtained under standardized examination conditions, with fixed protocols, comparable room layouts, and clear visibility of rubric-relevant interactions among the trainee, patient simulator, and instruments.

Physician Rubrics and Dual-Expert Annotation. Each skill category is paired with a standardized step-wise scoring rubric used in real clinical skill examinations, with full details provided in Appendix B. Developed by physicians for procedural assessment, these rubrics decompose each task into predefined sub-steps with explicit scoring criteria, ensuring that labels reflect expert judgments of procedural correctness rather than free-form descriptions or proxy supervision.

Two clinically trained physician annotators[5] independently review each full-length video and score every predefined step according to the rubric-defined requirements. The overall procedure score is obtained by aggregating these step-level scores under the rubric-specific scoring scheme. Thus, each sample contains dual-expert, rubric-aligned step annotations and a total score grounded in the full interaction process.

Quality control and adjudication. Quality control was conducted during both video curation and annotation. At the data level, we retained only videos with complete procedural coverage and clear visibility of rubric-relevant interactions, excluding recordings that were insufficient for rubric-based assessment due to interruption, severe occlusion, or incomplete capture. Procedurally incorrect or low-quality student performances were retained, as they are intrinsic to real clinical skill assessment and essential for evaluating rubric-grounded procedural judgment.

At the annotation level, two physicians first independently reviewed each full-length video and assigned step-level scores using the official rubric. Discrepant step labels were then jointly re-examined and resolved step by step through adjudication. The final consensus labels were used to derive the overall procedure score. Inter-annotator agreement before adjudication is reported in Table 3. Overall, 65 step instances (0.98%) across 31 videos (15.5%) required adjudication, and 33 videos were excluded because the recordings were insufficient for rubric-based assessment.

SiMing-Bench provides dense expert annotations. Across 200 videos, it yields 6,600 step-level procedural judgments and 13,200 independent physician step-level scores before adjudication. Despite its focused scale and procedure coverage, it offers a rigorous diagnostic setting for state-conditioned procedural judgment beyond ordinary video QA or coarse action recognition. SiMing-Bench is being actively expanded through ongoing data collection. Details of the expansion plan are provided in Appendix I.

3.3 Evaluation

We evaluate MLLMs on SiMing-Bench under a rubric-grounded prompting protocol, where each structured clinical scoring rubric is converted into a textual prompt and the model predicts step-level scores for the full procedure. Because clinical skill assessment requires both global judgment over the entire procedure and local judgment over clinically meaningful sub-steps, we evaluate model performance at two complementary levels.

Overall-score consistency. We first measure whether model predictions agree with expert judgments on the quality of the complete procedure. Let $\hat{y}$ and y denote the predicted and expert overall scores, respectively. We report Pearson correlation(PLCC) and Spearman correlation(SRCC) over the evaluation set. PLCC measures agreement in absolute scoring tendency, while SRCC measures whether the model preserves the expert ranking of procedural quality across samples. The formulations are provided in Appendix E.1.

Step-level consistency. We further measure whether the model agrees with experts on the correctness of individual procedural steps. Since step-level ratings are ordinal rather than interval-scaled, we evaluate agreement using quadratic weighted

[4]Appendix A.1 provides detailed descriptions of CPR, AED operation, and BMV, including their procedural characteristics and representative action sequences.

[5]More details in Appendix C

Models	Parameters	PLCC	SRCC	Cohen's κ	MCC	P	R	F_1
Open-source Video-MLLMs								
LLaVA-Video (Zhang et al., 2024)	7B	0.104	-0.095	0.033	-0.033	0.157	0.135	0.145
LLaVA-Video (Zhang et al., 2024)	72B	-0.047	-0.015	0.044	0.039	0.207	0.369	0.266
LLaVA-OneVision (Li et al., 2024a)	7B	0.038	0.123	0.038	0.029	0.145	0.059	0.084
Qwen3-VL (Bai et al., 2025)	30B	0.037	0.099	-0.010	0.043	0.174	0.659	0.276
Omni-MLLMs								
Qwen3-Omni (Xu et al., 2025)	30B	-0.002	0.011	0.000	0.000	0.183	1.000	0.310
GPT-4o (OpenAI, 2024)	–	0.158	0.169	0.000	-0.023	0.156	0.070	0.097
Medical Video-MLLMs								
UniMed-VL (Ning et al., 2025)	14B	-0.035	-0.053	-0.006	0.053	0.257	0.107	0.152
MedGemma (Sellergren et al., 2025)	27B	-0.059	-0.074	0.007	0.028	0.201	0.393	0.266
MediX-R1 (Mullappilly et al., 2026)	30B	-0.046	-0.058	-0.048	-0.039	0.176	0.461	0.254
Commercial MLLMs								
Doubao-1.6-Vision (ByteDance, 2025)	230B	-0.055	-0.034	0.037	-0.031	0.150	0.081	0.105
Gemini-3 Pro (Google, 2025)	–	0.022	-0.116	0.018	-0.020	0.157	0.106	0.126
GPT-5.2 (OpenAI, 2025)	–	-0.053	0.016	0.048	0.132	0.212	0.909	0.344

Table 4: Main results on SiMing-Bench. We evaluate diverse MLLMs under the rubric-grounded full-video setting at both the overall-score and step levels. MCC, Precision, Recall, and F_1 are additionally reported for binary step correctness judgment (more details in Section 4.3).

Cohen's κ. Both expert and predicted step scores are mapped to four ordered categories, corresponding to fully correct, basically correct with minor issues, clearly deficient, and severely incorrect or failed. This metric penalizes larger disagreements more heavily and is therefore appropriate for rubric-based procedural assessment. Implementation details are provided in Appendix E.2.

Unless otherwise specified, overall-score metrics are computed against the expert consensus score, and step-level agreement is computed over all rubric-defined steps in the evaluation set.

4 Experiments

4.1 Experimental Settings

Baseline models. To comprehensively evaluate model performance on SiMing-Bench, we compare a diverse set of representative baselines spanning general-purpose open-source Video-LLMs (Zhang et al., 2024; Li et al., 2024a; Bai et al., 2025), omni-modal large models (Xu et al., 2025; OpenAI, 2024), medical-domain MLLMs (Ning et al., 2025; Sellergren et al., 2025; Mullappilly et al., 2026), and commercial proprietary MLLMs (ByteDance, 2025; Google, 2025; OpenAI, 2025). This selection allows us to assess procedural understanding ability across models with different architectures and domain specialization[6].

4.2 Main Results Analysis

Current MLLMs fail to achieve expert-aligned process-level judgment. Table 4 shows that all evaluated models perform poorly on SiMing-Bench, with uniformly weak agreement with expert ratings at both the overall-score and step levels. The highest observed overall-score result reaches only 0.158 PLCC and 0.169 SRCC, while yielding zero step-level κ. Among open-source models, LLaVA-OneVision achieves the highest SRCC, but no model produces consistent gains across evaluation levels. The dominant pattern is therefore not isolated weakness in particular systems, but a systematic failure to convert long-horizon interaction history into expert-aligned, rubric-grounded process-level judgment. This failure also does not diminish reliably with scale: performance does not improve monotonically, and both open- and closed-source systems remain well below expert agreement under zero-shot evaluation.

Overall-score agreement masks failures in step-level procedural assessment. A central finding is the persistent mismatch between overall-score consistency and step-level consistency. GPT-

[6]Our main experiments focus on zero-shot, rubric-conditioned MLLM evaluation. We discuss alternative approaches that require task-specific adaptation in Appendix F.

Models	PLCC	SRCC	Cohen's κ
LLaVA-Video-7B	-0.150 (-0.254)	-0.170 (-0.075)	0.009 (-0.024)
LLaVA-Video-72B	0.057 (+0.104)	0.098 (+0.113)	-0.003 (-0.047)
LLaVA-OneVision	0.000 (-0.038)	0.000 (-0.123)	0.000 (-0.038)
Qwen3-VL	0.135 (+0.098)	0.125 (+0.026)	0.000 (+0.010)
Qwen3-Omni	0.092 (+0.094)	0.110 (+0.099)	0.075 (+0.075)
GPT-4o	-0.018 (-0.176)	0.004 (-0.165)	0.019 (+0.019)
UniMed-VL	0.084 (+0.119)	0.061 (+0.114)	0.016 (+0.022)
MedGemma	0.030 (+0.089)	0.079 (+0.153)	0.002 (-0.005)
MediX-R1	0.108 (+0.154)	0.121 (+0.179)	0.010 (+0.058)
Doubao-1.6-Vision	0.055 (+0.110)	0.032 (+0.066)	0.006 (-0.031)
Gemini-3 Pro	0.331 (+0.309)	0.296 (+0.412)	0.015(-0.003)
GPT-5.2	0.147 (+0.200)	0.157 (+0.141)	0.011 (-0.037)

Table 5: Comparison between the full-video setting and the step-aligned clips setting on PLCC, SRCC and Cohen's κ. Improvements are highlighted with red backgrounds.

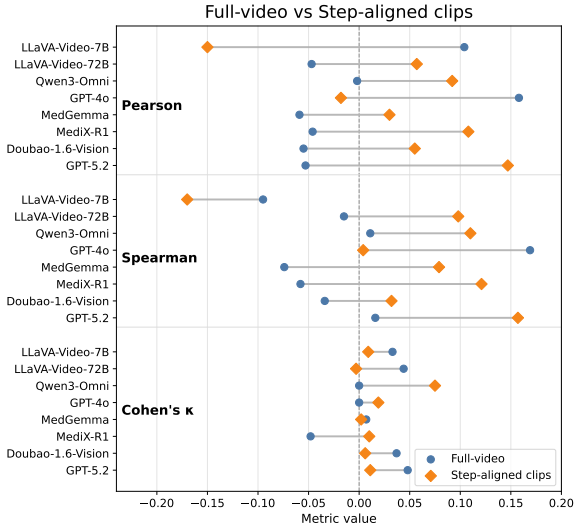


Figure 2: Comparison between the full-video setting and the step-aligned clip setting. Each dumbbell connects the same model under the two evaluation settings.

4o, for example, achieves the strongest correlation with expert overall scores, yet fails to obtain positive agreement on rubric-defined steps. Similar discrepancies appear across other models. This indicates that weak agreement at the full-procedure level can arise without correct judgment of the intermediate interactions that determine procedural validity. Current models may capture coarse signals about global procedure quality while still failing to identify which earlier interactions changed the procedural state and why later actions became correct or incorrect. This result directly supports the multi-granularity design of SiMing-Bench: evaluating only the overall score would substantially overestimate current models' ability to perform expert-like process-level assessment.

Neither scale nor medical specialization resolves the underlying failure. Increasing model scale does not reliably improve performance on SiMing-Bench. For example, larger model variants(LLaVA-Video) do not consistently yield higher agreement across metrics, and larger commercial models also remain near zero in agreement. Medical-domain models likewise fail to show a consistent advantage. UniMed-VL, MedGemma, and MediX-R1 remain near or below zero on the reported metrics, indicating that medical specialization alone is insufficient for rubric-grounded procedural assessment. This is consistent with the task design of SiMing-Bench, which does not primarily test recognition of medical objects or terminology, but whether a model can represent how continuous interactions update procedural conditions and determine the validity of subsequent actions.

SiMing-Bench therefore exposes a concrete, interaction-centered world-modeling challenge for current video MLLMs: they do not reliably infer how earlier interactions reshape the conditions under which later actions become correct, incorrect, feasible, or invalid.

4.3 Diagnostic Analysis of Step-level Failure

To diagnose the source of step-level failure, we conduct two additional analyses that separately reduce the difficulty of score assignment and temporal evidence search. To further examine whether the observed performance gap can be attributed to alternative factors beyond interaction-grounded procedural state modeling, we provide additional diagnostic analyses in Appendix G.

Binary correctness judgment does not resolve the failure. We first simplify step-level evaluation from ordinal scoring to binary correctness classification, where the model only decides whether each interaction step is correctly performed. This removes the need to map visual evidence to fine-grained score levels and isolates a more basic capability, judging whether a clinically meaningful interaction satisfies the procedural requirement. Performance nevertheless remains weak across all metrics[7]. As shown in Table 4, most models achieve MCC values close to zero, and some are negative; even the highest observed MCC is only 0.132. Moreover, some seemingly stronger F1 scores are

[7]Appendix E.3 details the Binary Error-detection Metrics. To evaluate whether models can identify incorrect procedural execution, erroneous steps are treated as the positive class when computing Precision, Recall, F1, and MCC.

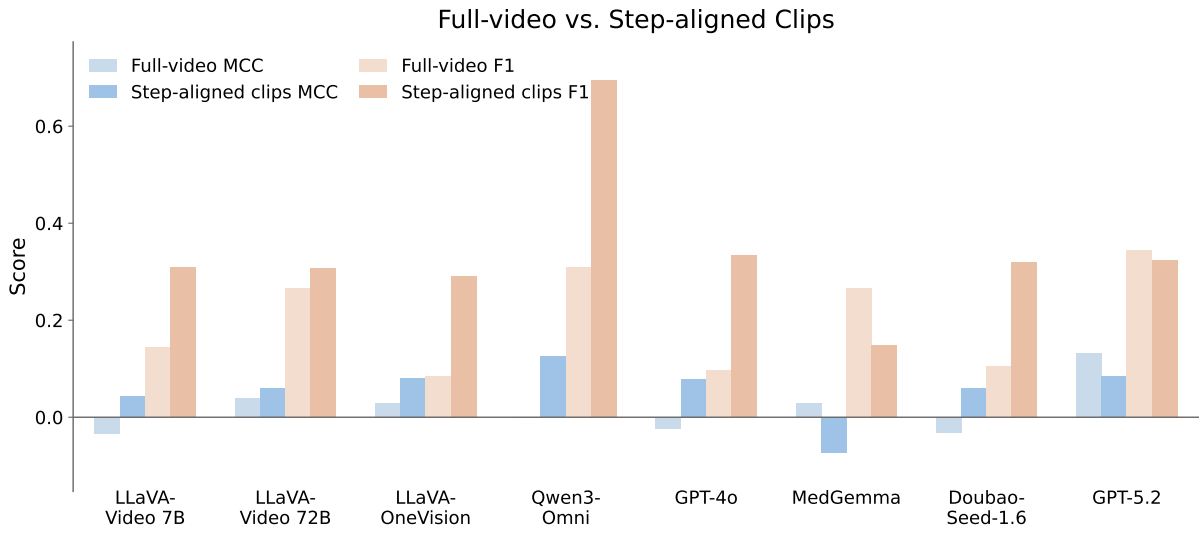


Figure 3: Comparison between the full-video setting and the step-aligned clips setting on binary step judgment. For each model, we report MCC and F1 under both settings.

driven by heavily biased prediction behavior rather than reliable step discrimination. For example, Qwen3-Omni attains perfect recall but extremely low precision and zero MCC, indicating a strong tendency to predict most steps as incorrect. This suggests that the bottleneck is not merely fine-grained score assignment, but the underlying ability to distinguish correct from incorrect interactions.

Step-aligned temporal localization yields only limited gains. We next evaluate models on manually segmented step-aligned clips, where each input contains only a short clip and the corresponding step requirement. This removes the need to search over the full procedure for relevant evidence. Figure 2 illustrates that while some models show modest improvement in overall-score correlations, step-level agreement remains extremely weak. As shown in Table 5, κ stays near zero for almost all models, with the highest observed value reaching only 0.075. Binary correctness evaluation shows a similar pattern. Figure 3 shows that, although step segmentation improves performance for some models, the gains remain limited and often coexist with strong positive prediction bias. These results indicate that long-horizon temporal search is only part of the challenge. Even when relevant evidence is pre-localized, current MLLMs still struggle to judge whether the observed interaction satisfies the procedural criterion.

Taken together, these diagnostic analyses show that the main bottleneck is neither fine-grained score mapping alone nor temporal localization alone. Instead, current MLLMs fail at the more fundamental problem of converting observed interactions into expert-aligned procedural judgment, consistent with a broader failure to maintain interaction-updated procedural state over the course of a procedure.

5 Conclusion

We introduce SiMing-Bench, a benchmark for evaluating expert-aligned process-level judgment from full-length clinical skill videos. Beyond event recognition or long-context recall, SiMing-Bench tests whether MLLMs can model how ongoing interactions update procedural conditions and thereby determine the correctness of later actions. Across overall-score, step-level, binary judgment, and step-aligned clip settings, current MLLMs consistently show weak agreement with physician experts. Moreover, weak performance on rubric-defined intermediate steps persists even when overall procedure-level correlation appears acceptable, suggesting that coarse agreement substantially overestimates true procedural assessment ability. The limited gains from binary simplification and pre-localized clips further indicate that the core bottleneck is not merely scoring difficulty or temporal evidence search, but interaction-grounded procedural state modeling. We hope SiMing-Bench encourages future multimodal systems to move beyond event recognition toward expert-aligned judgment over continuously evolving real-world processes.

Limitations

SiMing-Bench focuses on clinical skill assessment in standardized educational examination settings. This design provides high-quality rubric-grounded supervision and controlled procedural comparisons, but it does not cover the full variability of real clinical practice, where patient conditions, environment, team coordination, and unexpected events may be substantially more complex. Our conclusions should therefore be interpreted as evidence of a core failure in process-level procedural judgment under structured assessment, rather than as a complete characterization of video understanding in all clinical environments.

The benchmark currently covers three clinical skill categories, including cardiopulmonary resuscitation, automated external defibrillator operation, and bag-mask ventilation. These tasks were chosen because they involve long-form, interaction-dependent procedures with expert-defined rubrics, but they do not exhaust the space of clinical procedural reasoning. Our findings should therefore be interpreted as evidence about interaction-grounded procedural judgment in structured clinical skill assessment, rather than as a complete characterization of MLLM world modeling across all procedural domains. Other settings, such as surgery, bedside examination, multi-actor team procedures, or non-medical assembly tasks, may involve different objects, temporal dependencies, interaction structures, and sources of uncertainty.

Our evaluation is also rubric-grounded by design. This is a strength for standardized assessment, but it means that the benchmark emphasizes agreement with expert scoring protocols rather than all possible forms of clinical reasoning. In particular, successful performance on SiMing-Bench would not by itself imply broader competence in diagnosis, treatment planning, or open-ended clinical decision making.

Finally, although we include diagnostic analyses with binary judgment and step-aligned clips, these experiments do not fully disentangle all possible sources of model failure. The observed performance gaps are consistent with a limitation in interaction-grounded procedural state modeling, but future work is still needed to more explicitly probe memory, localization, causal attribution, and latent world-state tracking in long-form multimodal reasoning.

Ethical Considerations

This work studies MLLMs in the context of clinical skill assessment, a domain in which errors can have educational and safety implications. SiMing-Bench is intended strictly as a research benchmark for evaluating process-level multimodal reasoning. It is not designed or validated for direct deployment in medical education scoring, competency certification, or clinical decision support without substantial additional validation, oversight, and regulatory review.

The dataset is curated from standardized clinical skill training and examination scenarios involving medical students. The data collection and research use protocols were reviewed and approved by the Institutional Review Board (IRB)[8] of the participating institution. Because the videos concern educational assessment in a medically themed setting, careful handling of privacy and institutional permissions is essential. Data collection, storage, annotation, and release should comply with the relevant institutional review, consent, and de-identification requirements applicable to the participating sites. Any public release should avoid personally identifying information and should preserve only the information necessary for benchmark evaluation.

A further ethical risk is overinterpretation of model capability. Since the benchmark measures rubric-grounded procedural judgment, strong performance would not imply that a model is safe for real patient-facing use. Conversely, weak performance should not be misread as a judgment about student competence outside the specific scoring rubric and examination context. We therefore frame all results as model-evaluation findings rather than educational or clinical decisions about individuals.

Finally, the benchmark may contribute to future automated assessment tools, which creates both opportunity and risk. Such systems could support training and feedback, but they could also amplify annotation biases, institutional scoring conventions, or unjustified trust in model outputs. For this reason, any downstream use should preserve human expert oversight, make scoring criteria transparent, and treat model outputs as assistive rather than authoritative.

[8] More details in Appendix H

Acknowledgments

This work is supported by the Key Program of the National Natural Science Foundation of China (Grant No. U23A20316), the General Program of the National Natural Science Foundation of China (Grant No. 62676290), CCF-Tencent Rhino-Bird Open Research Fund (Grant No. CCF-Tencent RAGR20250115), Wuhan Natural Science Foundation Exploratory Program (Morning Light Program) Project (Grant No. 2026040301020029), the Fundamental Research Funds for the Central Universities (Grant No. 2042026rw219) and the Intelligent Computing Center of the National Cybersecurity Talent and Innovation Base, Wuhan. SA acknowledges support from the British Heart Foundation Centre of Research Excellence at The University of Manchester (RE/24/130017).

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, and 1 others. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.

ByteDance. 2025. Seed1.6 tech introduction. https://seed.bytedance.com/en/seed1_6. Accessed: December 2025.

Delong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. 2025. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722*.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, and 1 others. 2025. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*.

Lauren Chong, Silas Taylor, Matthew Haywood, Barbara-Ann Adelstein, and Boaz Shulruf. 2017. The sights and insights of examiners in objective structured clinical examinations. *Journal of educational evaluation for health professions*, 14.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, and 1 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Michael D Cusimano, Robert Cohen, William Tucker, John Murnaghan, Ron Kodama, and Richard Reznick. 1994. A comparative analysis of the costs of administration of an osce (objective structured clinical examination). *Academic Medicine*, 69(7):571–6.

Ronald M Epstein and Edward M Hundert. 2002. Defining and assessing professional competence. *Jama*, 287(2):226–235.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. 2024. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7346–7355.

Google. 2025. A new era of intelligence with gemini 3. https://blog.google/products-and-platforms/products/gemini/gemini-3/. Accessed: 2026-03-15.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Ronald M Harden, Mary Stevenson, W Wilson Downie, and GM Wilson. 1975. Assessment of clinical competence using objective structured examination. *Br Med J*, 1(5955):447–451.

David Hope and Helen Cameron. 2015. Examiners are most lenient at the start of a two-day osce. *Medical Teacher*, 37(1):81–85.

Xiyang Huang, Yingjie Han, Yaoxu Li, Runzhi Li, Pengcheng Wu, and Kunli Zhang. 2025. Cmeaa: Cross-modal enhancement and alignment adapter for radiology report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8546–8556.

Xiyang Huang, Renxiong Wei, Yihuai Xu, Zhiyuan Chen, Keying Wu, Jiayi Xiang, Buzhou Tang, Yanqing Ye, Jinyu Chen, Cheng Zeng, and 1 others. 2026. Overview of the clinicalskillqa 2026 shared task on continuous perception and procedural reasoning in

clinical skill assessment. In *BioNLP 2026*, pages 1101–1108.

Kamran Z Khan, Sankaranarayanan Ramachandran, Kathryn Gaunt, and Piyush Pushkar. 2013. The objective structured clinical examination (osce): Amee guide no. 81. part i: an historical and theoretical perspective. *Medical teacher*, 35(9):e1437–e1446.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, and 1 others. 2025a. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Yaoqian Li, Xikai Yang, Dunyuan Xu, Yang Yu, Litao Zhao, Xiaowei Hu, Jinpeng Li, and Pheng-Ann Heng. 2025b. Surgpub-video: A comprehensive surgical video dataset for enhanced surgical intelligence in vision-language model. *arXiv preprint arXiv:2508.10054*.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. 2024. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110.

Weiheng Lu, Jian Li, An Yu, Ming-Ching Chang, Shengpeng Ji, and Min Xia. 2024. Llava-mr: Large language-and-vision assistant for video moment retrieval. *arXiv preprint arXiv:2411.14505*.

David Ma, Huaqing Yuan, Xingjian Wang, Qianbo Zang, Tianci Liu, Xinyang He, Yanbin Wei, Jiawei Guo, Ni Jiahui, Zhenzhu Yang, and 1 others. 2025. Scalelong: A multi-timescale benchmark for long video understanding. *arXiv preprint arXiv:2505.23922*.

Kevin McLaughlin, Martha Ainslie, Sylvain Coderre, Bruce Wright, and Claudio Violato. 2009. The effect of differential rater function over time (drift) on objective structured clinical examination ratings. *Medical education*, 43(10):989–992.

George E Miller. 1990. The assessment of clinical skills/competence/performance. *Academic medicine*, 65(9):S63–7.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. 2026. Medix-r1: Open ended medical reinforcement learning. *arXiv preprint arXiv:2602.23363*.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Lihao Liu, Yuandong Pu, Huihui Xu, Chenhui Gou, Ziyan Huang, Yi Xin, Qi Qin, Zhongying Deng, Diping Song, Bin Fu, and 8 others. 2025. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. *Preprint*, arXiv:2510.15710.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2026-03-15.

OpenAI. 2025. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/. Accessed: 2026-03-15.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. 2025. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24129–24138.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Shiva Sreeram, Tsun-Hsuan Wang, Alaa Maalouf, Guy Rosman, Sertac Karaman, and Daniela Rus. 2025. Probing multimodal llms as world models for driving. *IEEE Robotics and Automation Letters*.

Cees PM Van Der Vleuten and Lambert WT Schuwirth. 2005. Assessing professional competence: from methods to programmes. *Medical education*, 39(3):309–317.

Merrilyn Walton, Helen Woodward, Samantha Van Staalduinen, Claire Lemer, Felix Greaves, Douglas Noble, Benjamin Ellis, Liam Donaldson, Bruce Barraclough, and as Expert Lead for the Sub-Programme Expert Group convened by the World Alliance of Patient Safety. 2011. Republished paper: The who patient safety curriculum guide for medical schools. *Postgraduate medical journal*, 87(1026):317–321.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. 2025c. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18925–18935.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2025. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine*, 8(1):141.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36:33469–33484.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, and 1 others. 2025. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*.

Daichi Yashima, Shuhei Kurita, Yusuke Oda, and Komei Sugiura. 2026. Remora: Multimodal large language model based on refined motion representation for long-video understanding. *arXiv preprint arXiv:2602.16412*.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, and 1 others. 2024. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*.

Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, and 1 others. 2025. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer.

Kunli Zhang, Xiyang Huang, Hongying Zan, Yutuan Ma, Qianxiang Gao, and Yaoxu Li. 2023. Fgkf: Fine-grained knowledge fusion for radiology report generation. In *China Health Information Processing Conference*, pages 238–253. Springer.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

A Clinical Skill Assessment

A.1 What is clinical skills assessment?

Clinical skills assessment refers to performance-based approaches in medical education used to evaluate whether learners can integrate communication, knowledge, technical skills, and clinical reasoning in practice-relevant tasks (Epstein and Hundert, 2002). Rather than assessing only what learners know, such assessment is commonly understood as targeting higher levels of performance beyond knowledge alone (Miller, 1990). One of the most widely used formats for this purpose is the Objective Structured Clinical Examination (OSCE) (Khan et al., 2013), which has become widely adopted in both undergraduate and postgraduate clinical education. In an OSCE, learners rotate through a series of stations organized around specific clinical tasks, with predefined scoring approaches and standardized administration procedures intended to improve the structure and objectivity of clinical performance assessment. In our study, we focus on procedural clinical skills in emergency care, specifically cardiopulmonary resuscitation (CPR), automated external defibrillator (AED) operation, and bag-mask ventilation (BMV). As shown in Figure 4, we present multi-frame examples of these skills to illustrate how each procedure unfolds over time through a sequence of coordinated actions. These skills are highly structured, time-sensitive, and interaction-intensive, requiring learners to complete a full sequence of clinically meaningful actions rather than isolated subtasks. This processual nature makes them well suited for evaluating fine-grained procedural competence, including step-wise execution quality, action ordering, and consistency across the entire workflow.

A.2 Why is it important?

Clinical skills assessment is important for at least three reasons. First, it is closely tied to patient safety and quality of care, as medical training must ultimately ensure that learners can translate knowledge into safe and effective clinical practice rather than merely reproduce factual information in written examinations (Walton et al., 2011). Second, it addresses a fundamental limitation of written tests. As emphasized in the classic framework of clinical competence, written examinations mainly assess whether learners "know" or "know how," whereas clinical skills assessments are better suited to evaluating whether learners can actually demonstrate performance in practice, namely whether they can "show how" and, to some extent, "do" (Miller, 1990). This distinction is particularly important in medical education, where the ultimate goal is not merely knowledge retention, but the ability to function effectively in authentic clinical encounters (Harden et al., 1975). Third, clinical skills assessment supports not only individual learner development but also the continuous improvement of the curriculum and the assessment program itself. Contemporary medical education increasingly views assessment as a programmatic process that informs feedback, guides learning, and provides evidence for improving educational design at the program level (Van Der Vleuten and Schuwirth, 2005).

A.3 What challenges does it face?

Clinical skills assessment faces several persistent challenges. First, it is resource-intensive, particularly when implemented in the form of OSCEs. OSCEs are widely recognized as difficult to organize, requiring substantial human and material resources and often involving greater time and cost than traditional or less structured assessment formats (Cusimano et al., 1994). Second, scoring consistency remains limited despite the standardized design of OSCEs. Although OSCEs were developed to improve the structure and objectivity of clinical performance assessment, examiner judgments may still be affected by contextual factors, examiner examinee interactions, and differences among examiners, indicating that standardization does not fully eliminate scoring variability (Chong et al., 2017). Third, examiner severity may vary over the course of an examination. Temporal variation in OSCE scoring has been observed, including differential rater function over time and greater leniency at the beginning of an examination (McLaughlin et al., 2009; Hope and Cameron, 2015). This suggests that examiner severity may shift systematically even within the same assessment session, introducing an additional source of rating variability beyond case content or candidate performance alone.

Recent advances in large language models (LLMs) (Xie et al., 2023, 2024) have substantially improved general reasoning and language understanding capabilities. Building on these advances, multimodal large language models (MLLMs) have shown strong potential for interpreting multimodal

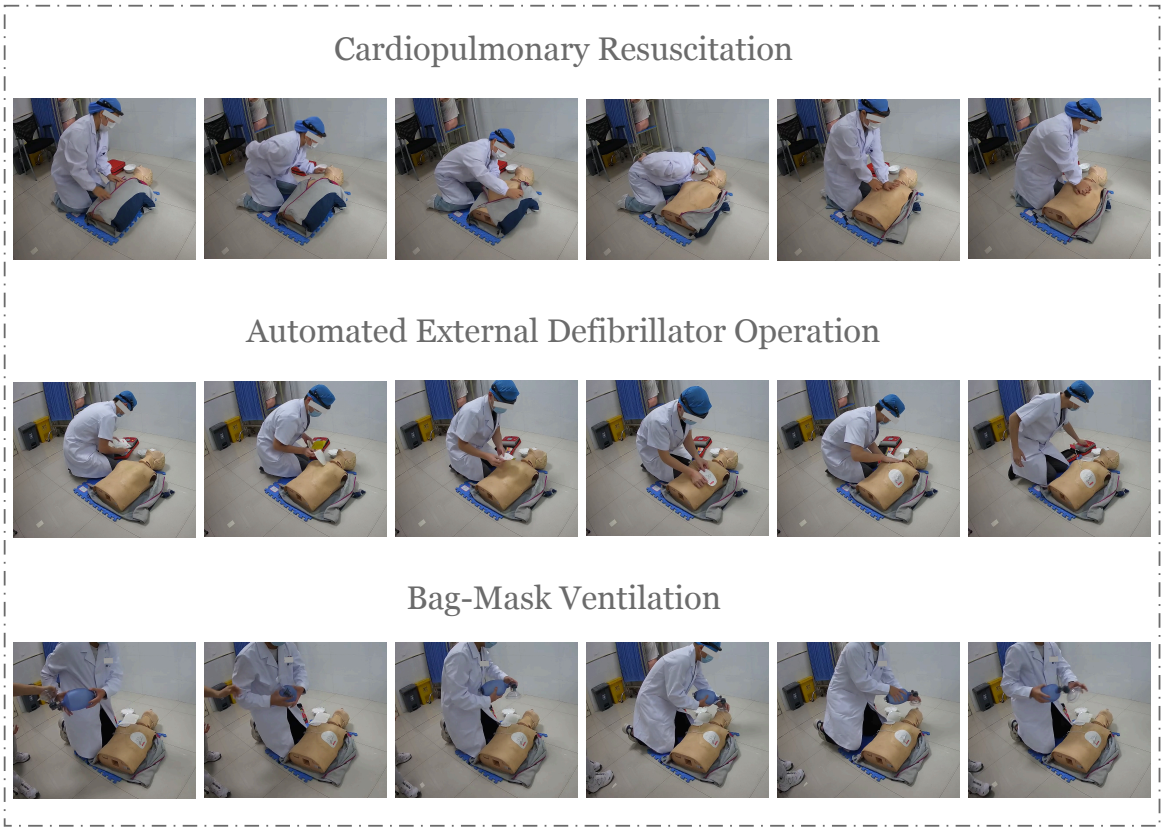


Figure 4: Multi-frame examples of the three emergency care skills considered in this study are shown: cardiopulmonary resuscitation (CPR), automated external defibrillator (AED) operation, and bag-mask ventilation (BMV). The figure illustrates how each procedure unfolds as a temporally ordered sequence of coordinated actions, highlighting the highly structured and interaction-intensive nature of these tasks.

inputs across a wide range of domains (Huang et al., 2025; Xie et al., 2025; Wang et al., 2023; Zhang et al., 2023; Huang et al., 2026). In clinical settings, these capabilities create new opportunities for automated skill assessment, where MLLMs hold promise for providing scalable and consistent evaluation of procedural performance. Such systems therefore offer a promising direction for reducing assessment costs and mitigating variability in human scoring.

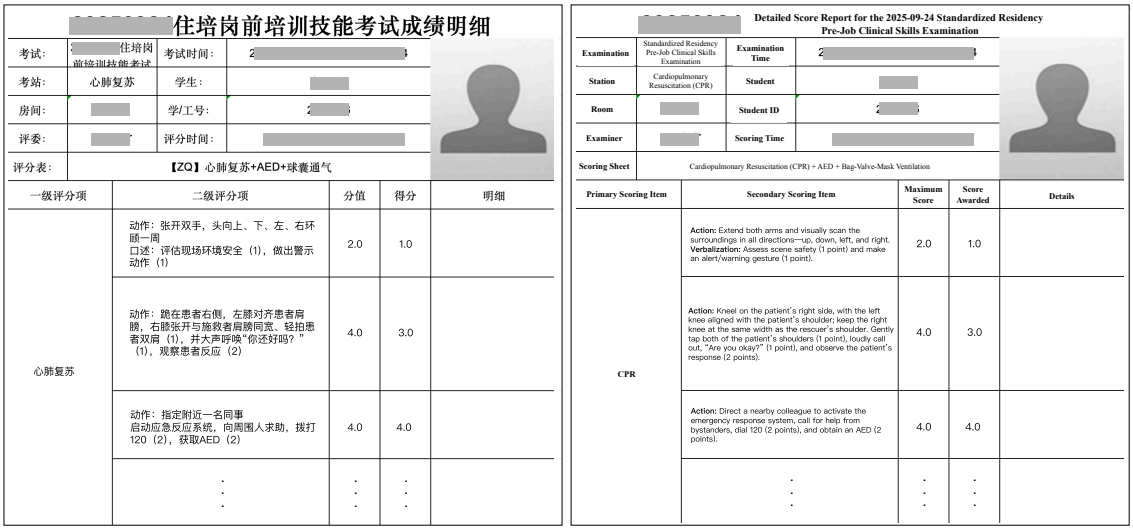
B Physician-Defined Scoring Rubric

Each clinical skill category is paired with an official physician-defined scoring rubric used in standardized clinical skill examinations. As shown in Figure 5, the rubric decomposes the full procedure into clinically meaningful sub-steps and specifies, for each step, the maximum score and the scoring criteria. The resulting annotations are therefore rubric-grounded judgments of procedural correctness rather than free-form descriptions. Across the three skill categories, the benchmark contains 33 rubric-defined steps in total: 19 for CPR, 9 for AED, and 5 for BMV. The complete set of rubric items is shown in Figure 6, Figure 7 and Figure 8. As reported in the main paper, each video contains the complete workflow covering all three skills. The step-level labels are assigned independently by two clinically trained physician annotators according to the rubric and then adjudicated when disagreements occur.

C Privacy and Annotation

The student videos were collected in the context of routine clinical skills training and assessment rather than through open recruitment. During data collection, students wore surgical caps and masks throughout video recording, and facial regions in the collected videos were additionally masked to further protect identity information. In addition, personally identifying information in the scoring forms, including names, institutional affiliations, and student ID numbers, was masked.

Both annotators are physicians from the collabo-



[redacted]住培岗前培训技能考试成绩明细

考试：	[redacted]住培岗前培训技能考试	考试时间：	[redacted]
考站：	心肺复苏	学生：	[redacted]
房间：	[redacted]	学/工号：	[redacted]
评委：	[redacted]	评分时间：	[redacted]
评分表：	【ZQ】心肺复苏+AED+球囊通气		

一级评分项	二级评分项	分值	得分	明细
心肺复苏	动作：张开双手，头向上、下、左、右环顾一周 口述：评估现场环境安全（1），做出警示动作（1）	2.0	1.0	
	动作：跪在患者右侧，左膝对齐患者肩膀，右膝张开与施救者肩膀同宽，轻拍患者双肩（1），并大声呼唤“你还好吗？”（1），观察患者反应（2）	4.0	3.0	
	动作：指定附近一名同事启动应急反应系统，向周围人求助，拨打120（2），获取AED（2）	4.0	4.0	
	...	...	...	

[redacted] Detailed Score Report for the 2025-09-24 Standardized Residency Pre-Job Clinical Skills Examination

Examination	Standardized Residency Pre-Job Clinical Skills Examination	Examination Time	[redacted]
Station	Cardiopulmonary Resuscitation (CPR)	Student	[redacted]
Room	[redacted]	Student ID	[redacted]
Examiner	[redacted]	Scoring Time	[redacted]
Scoring Sheet	Cardiopulmonary Resuscitation (CPR) + AED + Bag-Valve-Mask Ventilation		

Primary Scoring Item	Secondary Scoring Item	Maximum Score	Score Awarded	Details
CPR	**Action:** Extend both arms and visually scan the surroundings in all directions—up, down, left, and right. **Verbalization:** Assess scene safety (1 point) and make an alert/warning gesture (1 point).	2.0	1.0	
	Action: Kneel on the patient's right side, with the left knee aligned with the patient's shoulder; keep the right knee at the same width as the rescuer's shoulder. Gently tap both of the patient's shoulders (1 point), loudly call out, "Are you okay?" (1 point), and observe the patient's response (2 points).	4.0	3.0	
	Action: Direct a nearby colleague to activate the emergency response system, call for help from bystanders, dial 120 (2 points), and obtain an AED (2 points).	4.0	4.0	
	...	...	...	

Figure 5: Representative example of the physician-defined scoring rubric in SiMing-Bench. The left panel presents the original Chinese examination form, and the right panel presents its English-translated version. English translations are provided for readability; the original Chinese rubrics are used as model inputs in the main experiments.

rating hospital involved in data collection, with substantial experience in clinical teaching and skills examination. Prior to formal annotation, they jointly participated in the development and refinement of the rubric. All videos were annotated independently. No crowdsourcing platform was used.

D Experimental Settings

D.1 Prompts

To ensure a fair and consistent comparison across all models, we employed a standardized evaluation protocol, where each model was evaluated in a single run under the same prompting setup. For each task type, we used a unified prompt template, modifying only the task-specific instructions and expected output formats when necessary. This design minimizes prompt-induced variation and helps ensure that observed performance differences are attributable to the models themselves rather than to differences in prompt construction. The four prompt templates correspond to the four evaluation settings studied in the paper: full-video step-wise scoring(Figure 9), full-video step-wise error detection(Figure 10), step-aligned clip scoring(Figure 11), and step-aligned clip error detection(Figure 12). Taken together, these four templates form a controlled prompt suite for the main and diagnostic experiments, enabling a more interpretable analysis of whether current MLLMs fail because of scoring granularity, temporal localization, or deeper limitations in converting continuous interaction processes into expert-aligned procedural assessments.

Models	Frames
LLaVA-Video	64
LLaVA-OneVision	64
Qwen3-VL	64
Qwen3-Omni	60
GPT-4o	50
UniMed-VL	30
MedGemma	30
MediX-R1	30
Doubao-1.6-Vision	240
Gemini-3 Pro	120
GPT-5.2	50

Table 6: The evaluated models and the number of input frames used for each model.

D.2 Temporal Sampling Density

For each model, we used its official default inference settings. For models that do not natively support video input, we uniformly sampled frames from each video according to the maximum number of frames allowed by the model. The number of input frames used for each model is reported in Table 6.

Our full-video evaluation follows the frame budget supported by each model. For example, LLaVA-Video uses 64 uniformly sampled frames, while GPT-4o and GPT-5.2 use 50 frames. Since each input video contains a complete CPR–AED–BMV workflow with an average duration of 168.9 seconds, full-video sampling provides a sparse tempo-

ral view of the procedure.

The 168.9-second duration refers to the complete assessment sample rather than CPR alone. The average durations of CPR, AED, and BMV are 85.6, 50.2, and 33.1 seconds, respectively. Thus, the effective sampling density differs across skills. Sparse sampling can be limiting for temporally dense CPR criteria, such as compression rhythm, rate, depth, and continuity.

The step-aligned clip setting partially reduces this limitation. In this setting, each input contains a manually localized clip for the target rubric step, while using the same model-specific frame budget. Based on the average skill durations and the number of rubric steps, the average step-level temporal window is approximately 5.12 seconds across all skills, with CPR, AED, and BMV corresponding to 4.51, 5.58, and 6.62 seconds per step on average, respectively. Under a 50-frame budget, this corresponds to an effective sampling density of approximately 9.77 FPS on average, and 11.10 FPS for CPR steps, compared with only 0.30 FPS over the full 168.9-second video and 0.58 FPS over the average 85.6-second CPR segment. For LLaVA-Video with 64 frames, the corresponding densities are approximately 12.50 FPS on average and 14.21 FPS for CPR steps, compared with 0.38 FPS over the full video and 0.75 FPS over the CPR segment. Thus, step-aligned clips substantially increase the effective frame density around the relevant action and remove the need to search over the entire procedure.

As shown in Table 5, however, step-level agreement remains weak under this setting. This suggests that sparse full-video sampling is not the primary factor behind the weak step-level agreement. As discussed in Section G.3, the weak performance on AED and BMV further suggests that the difficulty is not limited to CPR items requiring high-frequency motion estimation, since many AED and BMV steps involve action order, device handling, safety checks, airway preparation, and continuity of care.

Future work should evaluate denser video inputs, motion-aware encoders, and hybrid pipelines that combine temporal action analysis with rubric-conditioned procedural judgment.

E Metric

E.1 Overall-score Consistency

Formally, given expert overall scores $\{x_i\}_{i=1}^n$ and model-predicted overall scores $\{y_i\}_{i=1}^n$, Pearson correlation is computed as

$$r_p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (1)$$

where $\bar{x}$ and $\bar{y}$ are the sample means of expert and predicted scores, respectively.

Let $u_i = \mathrm{rank}(x_i)$ and $v_i = \mathrm{rank}(y_i)$ denote the ranks of expert and predicted overall scores. Spearman correlation is then computed as

$$r_s = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^n (u_i - \bar{u})^2}\sqrt{\sum_{i=1}^n (v_i - \bar{v})^2}}, \quad (2)$$

where $\bar{u}$ and $\bar{v}$ are the mean ranks of $\{u_i\}_{i=1}^n$ and $\{v_i\}_{i=1}^n$, respectively.

E.2 Step-level scoring consistency

To improve the interpretability of the score-to-category conversion, we adopt a **rubric-aligned mapping scheme**. Let $s_{\max}$ denote the maximum score for a procedural step, d the number of deducted points, and $s = s_{\max} - d$ the resulting raw score. The ordinal label $g(s)$ is defined as

$$g(s) = \begin{cases} 3, & d = 0, \\ 2, & 0 < d \leq 1 \text{ and } s > 0, \\ 1, & 1 < d < s_{\max}, \\ 0, & s = 0. \end{cases} \quad (3)$$

This mapping follows the logic of the original clinical rubric: label 3 corresponds to fully correct execution, label 2 to execution with only a minor deduction, label 1 to clearly deficient execution with partial credit retained, and label 0 to severely incorrect or failed execution. Given these ordinal labels, we compute **quadratically weighted Cohen's κ** to measure agreement between expert annotations and model predictions. Let O_{ij} denote the observed proportion of samples assigned to expert label i and model label j, and let E_{ij} denote the corresponding expected proportion under statistical independence. With $K = 4$ ordinal categories, the quadratic weight is defined as

$$w_{ij} = \frac{(i-j)^2}{(K-1)^2}. \quad (4)$$

The weighted observed disagreement and expected disagreement are given by

$$D_o = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} w_{ij} O_{ij}, \tag{5}$$

and

$$D_e = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} w_{ij} E_{ij}. \tag{6}$$

The final quadratically weighted Cohen's κ is computed as

$$\kappa = 1 - \frac{D_o}{D_e}. \tag{7}$$

A higher κ indicates stronger agreement between model predictions and expert ratings on ordered step-level judgments, while assigning larger penalties to disagreements that are farther apart on the ordinal scale.

E.3 Binary Error-detection Metrics

For the binary step judgment setting, the prompt asks the model to output a correctness indicator for each rubric step, where 1 denotes that the step is correctly performed and 0 denotes that the step contains an error. For metric computation, we convert this correctness indicator into an error indicator by label inversion, such that erroneous steps are treated as the positive class. Specifically, if $c_i \in \{0, 1\}$ denotes the model-predicted correctness label for step i, we define the corresponding error label as $e_i = 1 - c_i$. The same conversion is applied to the expert correctness labels. All binary metrics are then computed between the predicted error labels and the expert error labels.

This convention is intentional. Clinical skill assessment is primarily concerned with detecting incorrect or unsafe procedural execution, rather than confirming the majority of correctly performed steps. We therefore report Precision, Recall, F1, and Matthews correlation coefficient (MCC) under this error-positive convention. Under this convention, high recall indicates that the model detects most erroneous steps, while low precision indicates that many steps flagged as erroneous are in fact correct. This also prevents the binary metrics from being dominated by the majority class of correctly performed steps.

F Future Directions: From Task-Specific Video Understanding to Expert-aligned Process-level Judgment

Task-interface mismatch. Our main experiments focus on zero-shot, rubric-conditioned MLLM evaluation, where a model receives a full-workflow clinical skill video together with a natural-language expert rubric and predicts step-level ordinal scores as well as an overall procedure score. This formulation differs from the standard interfaces of many task-specific video understanding and world-modeling approaches. Temporal action localization models typically predict temporal boundaries and action categories (Zhang et al., 2022). Action-recognition models, including SlowFast-style video backbones, usually predict action labels from clips or videos (Feichtenhofer et al., 2019). World-modeling methods often focus on predicting future states, modeling state transitions, or reasoning about action-conditioned consequences (Hafner et al., 2023; Bruce et al., 2024). These outputs are related to procedural video analysis, but they do not directly correspond to rubric-conditioned step-level ordinal scores.

This interface mismatch is not merely a representational inconvenience. It reflects a shift from recognizing procedural events to judging their state-conditioned validity. In SiMing-Bench, procedural correctness is not determined by action identity alone, but by whether an observed interaction satisfies an expert-defined criterion under the procedural state induced by preceding interactions. For example, recognizing that a trainee attaches AED pads is insufficient, since the step may be scored differently depending on whether earlier requirements, such as device preparation, patient exposure, and safety checks, have been properly completed. Similarly, detecting chest compression does not by itself determine whether the action satisfies rubric criteria involving hand position, depth, rhythm, recoil, and continuity. Thus, SiMing-Bench requires models to go beyond descriptive video understanding and perform normative, rubric-grounded procedural judgment over temporally evolving interactions.

Relevance to interaction-grounded procedural state tracking. Although these approaches cannot be directly applied to the current SiMing-Bench formulation without task-specific adaptation, they remain relevant to the underlying capability studied

in this work. SiMing-Bench operationalizes one concrete aspect of video world modeling: whether a model can track how continuous interactions update procedural conditions and use the updated state to judge the validity of later actions. Existing approaches may therefore provide useful diagnostic lenses and architectural ingredients for studying and improving this capability.

Temporal localization methods may help identify where rubric-relevant evidence occurs in a long procedure. Action-recognition models and video backbones may provide stronger visual representations for recognizing fine-grained procedural actions and execution quality. World-modeling methods may support explicit modeling of how each interaction changes the procedural state and constrains subsequent actions. These directions suggest that future MLLMs may benefit from mechanisms that combine long-horizon evidence localization, fine-grained procedural perception, and explicit state-update modeling. Therefore, these approaches should be viewed not as direct replacements for the zero-shot MLLM formulation evaluated in our main experiments, but as promising routes for strengthening interaction-grounded procedural state tracking in future rubric-grounded assessment systems.

G Supplemental Diagnostic Analyses

The main experiments show that current MLLMs achieve weak agreement with physician scores in rubric-grounded clinical skill assessment, especially at the step level. This section provides additional diagnostic analyses to examine whether the observed gap can be attributed to several alternative factors, including missing medical knowledge, rubric-text complexity, lack of explicit reasoning, fine-grained domain perception, label imbalance, statistical uncertainty, and score-prior baselines.

These analyses are not intended to isolate a single exclusive cause of failure. Instead, they test whether reducing specific sources of difficulty leads to substantial and consistent improvements. Across the diagnostic settings, model performance improves only in limited and model-dependent cases, while step-level agreement remains weak. This pattern suggests that the gap is not fully explained by any single confounder, and is more consistent with difficulty in converting observed clinical interactions into rubric-defined procedural judgments.

Model	PLCC	SRCC	Cohen's κ
Medical-knowledge-enhanced prompt			
LLaVA-Video-7B	0.020	0.029	0.005
LLaVA-Video-72B	0.046	0.060	0.013
LLaVA-OneVision	-0.079	-0.147	-0.011
Qwen3-VL	-0.028	-0.006	0.058
Qwen3-Omni	-0.175	-0.150	-0.004
GPT-4o	0.019	0.066	0.008
UniMed-VL	-0.022	0.054	0.008
MedGemma	0.019	0.018	-0.011
MediX-R1	-0.048	-0.047	-0.037
Doubao-1.6-Vision	0.134	0.157	0.010
Gemini-3 Pro	0.149	0.088	0.027
GPT-5.2	0.141	0.178	0.016
Rubric-simplified prompt			
LLaVA-Video-7B	-0.002	0.024	0.012
LLaVA-Video-72B	0.010	0.044	0.013
LLaVA-OneVision	-0.016	0.005	0.022
Qwen3-VL	-0.116	0.044	-0.003
Qwen3-Omni	-0.016	-0.050	0.004
GPT-4o	0.123	0.080	0.034
UniMed-VL	-0.055	-0.089	0.015
MedGemma	-0.015	-0.039	0.022
MediX-R1	-0.182	-0.233	-0.021
Doubao-1.6-Vision	-0.011	0.018	0.001
Gemini-3 Pro	-0.039	-0.013	0.015
GPT-5.2	0.148	0.194	0.019

Table 7: Diagnostic results for medical-knowledge-enhanced and rubric-simplified prompting. Additional clinical background and simplified rubrics produce limited and model-dependent changes in overall-score correlation, while step-level agreement remains close to zero across models.

G.1 Medical Knowledge and Rubric Understanding

One possible explanation for the weak results is that MLLMs lack the medical knowledge needed to interpret clinical skill videos. Another is that physician-defined rubrics are linguistically complex, making it difficult to map textual criteria to visual evidence. To examine these factors, we introduce two controlled prompting variants. Figure 14 and 15 show the prompt used in this setting.

Medical-knowledge-enhanced prompting. We prepend concise clinical background to the original scoring instruction. The added information summarizes the purpose of each procedure, key observable cues, and common failure modes. The video input, step-wise rubric, and scoring protocol are unchanged. This setting reduces the need to recall external clinical knowledge while preserving the requirement to judge whether the observed execution satisfies the rubric.

Rubric-simplified prompting. We rewrite the physician-defined rubric into shorter and more operational criteria while preserving the original scoring semantics. The simplified rubric removes redundant phrasing and makes the expected visual evidence more explicit. The video input and scoring protocol remain unchanged. This setting tests whether reducing rubric-comprehension difficulty improves assessment performance.

We compare these prompting variants against the original full-video, full-rubric setting reported in the main experiments. Table 7 shows that neither intervention leads to consistent improvement across models. With medical-knowledge-enhanced prompting, a few recent proprietary models obtain modest positive overall-score correlations, such as Gemini-3 Pro, GPT-5.2, and Doubao-1.6-Vision, but several other models remain near zero or negative. Rubric simplification shows an even more mixed pattern: GPT-5.2 improves in overall-score correlation, whereas many open-source and medical-domain models do not show reliable gains. Across both settings, step-level Cohen's κ remains close to zero for almost all models, with the highest values still indicating weak agreement. These results suggest that missing medical knowledge and rubric-text complexity may contribute to the task difficulty, but reducing these factors alone does not resolve the core failure in producing expert-aligned, step-level procedural judgments.

G.2 Explicit Reasoning Prompting

Another possible explanation is that the original zero-shot prompt requires complex multi-constraint rubric following. We therefore evaluate whether explicit reasoning prompts improve performance under the full-video setting with the full rubric.

We consider two prompting variants. The first is a generic reasoning prompt, where the model is instructed to reason step by step before assigning scores. The second is a state-aware reasoning prompt tailored to SiMing-Bench. For each rubric-defined step, the model is asked to identify: (1) the procedural condition required by the rubric, (2) the observed interaction, (3) the procedural state update induced by the interaction, and (4) whether the updated state satisfies the rubric criterion. This prompt encourages the model to organize its prediction around interaction-induced state changes rather than isolated action recognition. Figure 16 and 17 shows the prompt used in this setting.

Full Video + Full Rubric + Generic Reasoning			
Model	PLCC	SRCC	Cohen's κ
LLaVA-Video-7B	0.052	0.035	0.077
LLaVA-Video-72B	0.106	0.097	0.137
LLaVA-OneVision	-0.019	-0.043	-0.037
Qwen3-VL	-0.071	-0.026	0.002
Qwen3-Omni	0.034	0.046	0.037
GPT-4o	-0.121	-0.058	0.017
UniMed-VL	-0.081	-0.095	0.005
MedGemma	0.104	0.122	0.014
MediX-R1	0.037	-0.064	-0.039
Doubao-1.6-Vision	-0.075	-0.019	-0.024
Gemini-3 Pro	0.091	0.079	0.028
GPT-5.2	0.129	0.114	0.034
Full Video + Full Rubric + State-aware Reasoning			
Model	PLCC	SRCC	Cohen's κ
LLaVA-Video-7B	0.115	0.081	0.126
LLaVA-Video-72B	0.049	0.057	0.013
LLaVA-OneVision	0.117	0.083	-0.036
Qwen3-VL	0.176	0.214	0.019
Qwen3-Omni	-0.175	-0.105	-0.023
GPT-4o	-0.074	-0.055	0.018
UniMed-VL	-0.001	-0.022	0.008
MedGemma	-0.089	-0.084	-0.003
MediX-R1	0.133	0.192	-0.043
Doubao-1.6-Vision	0.081	0.156	0.010
Gemini-3 Pro	-0.001	-0.009	0.017
GPT-5.2	0.146	0.155	0.035

Table 8: Diagnostic results with explicit reasoning prompts under the full-video and full-rubric setting. State-aware reasoning improves overall-score correlation for some models, but the gains are not consistent across models. Step-level agreement remains weak under both prompting variants.

As shown in Table 8, generic reasoning does not reliably improve performance across models. Although a few models obtain modest positive overall-score correlations, the results remain generally weak and unstable. For example, GPT-5.2 achieves the highest PLCC under generic reasoning, while MedGemma obtains the highest SRCC; however, both correlations are still low. Similarly, the relatively higher step-level agreement of LLaVA-Video (72B) under generic reasoning remains limited, with Cohen's κ reaching only 0.137.

State-aware reasoning leads to clearer gains for some models, particularly in overall-score correlation. Qwen3-VL improves substantially from negative or near-zero correlation under generic reasoning to 0.176 PLCC and 0.214 SRCC with state-aware reasoning. Similar improvements are observed for LLaVA-Video (7B), LLaVA-OneVision, MediX-R1, Doubao-1.6-Vision, and GPT-5.2. However, the effect is not consistent:

Skill	Mean κ	Best κ	Top-3 κ	MCC	F1
CPR	0.014	0.063	0.044	0.007	0.157
AED	0.029	0.097	0.068	0.007	0.174
BMV	0.013	0.230	0.114	0.007	0.207

Table 9: Skill-wise diagnostic results across CPR, AED, and BMV. Agreement remains weak across all three skills, suggesting that the observed failure is not limited to CPR-specific fine-grained perception.

Skill	Corr.	Incorr.	Corr. Rate	Incorr. Rate
CPR	3249	551	85.5%	14.5%
AED	1352	448	75.1%	24.9%
BMV	772	228	77.2%	22.8%
Overall	5373	1227	81.4%	18.6%

Table 10: Label distribution in the binary step judgment setting. Incorrect steps are treated as the positive class.

several models, including LLaVA-Video (72B), Qwen3-Omni, MedGemma, and Gemini-3 Pro, perform worse under the state-aware prompt.

More importantly, explicit reasoning does not close the step-level judgment gap. Even when state-aware reasoning improves overall-score correlation, Cohen's κ remains low for most models, and the best step-level agreement across both prompting variants is only 0.137. This suggests that explicitly asking models to reason about procedural conditions and state updates can sometimes help coarse score prediction, but it does not reliably enable fine-grained, expert-aligned step-level assessment. The persistent weakness at the step level suggests that the bottleneck is not merely a matter of eliciting longer reasoning traces. Rather, the results are consistent with a difficulty in grounding observed interactions in evolving procedural states and applying rubric-defined correctness criteria consistently.

G.3 Skill-wise Analysis

A further concern is that the observed failure may primarily reflect fine-grained domain perception, especially for CPR steps that require estimating compression depth, rhythm, or hand placement from imperfect video observations. Clinical skill assessment indeed requires fine-grained perception, and we do not treat perception and procedural judgment as fully separable. However, if CPR-specific perceptual difficulty were the dominant source of failure, substantially stronger and more consistent agreement would be expected on AED and BMV, where many steps involve device handling, safety checks, airway preparation, mask positioning, and action order rather than precise physical measurement.

To examine this, we report skill-wise results across CPR, AED, and BMV. For each skill, we report the mean, best, and top-3 average weighted Cohen's κ across evaluated models, together with binary error-detection MCC and F1 computed by treating incorrect steps as the positive class.

Table 9 shows weak agreement across all three skills. The mean Cohen's κ remains close to zero for CPR, AED, and BMV, and the binary MCC is also near zero across the three procedures. Although BMV reaches a higher best κ than CPR and AED, this improvement is not reflected in the mean score, and even the top-3 average remains low. The F1 scores also remain modest across skills, indicating limited error-detection performance when incorrect steps are treated as the positive class. This pattern suggests that the failure is not confined to CPR items requiring fine-grained physical estimation. Models also struggle on AED and BMV, where many judgments depend on whether prior interactions have established the required procedural conditions.

G.4 Label Distribution in Binary Step Judgment

We analyze the label distribution in the binary step judgment setting, where incorrect steps are treated as the positive class for precision, recall, and F1. This analysis clarifies how class imbalance affects the reported metrics.

As shown in Table 10, the dataset is moderately skewed toward correct steps, with an overall incorrect rate of 18.6%. This imbalance motivates the use of MCC and Cohen's κ, rather than accuracy alone. It also helps interpret model behavior: high recall can result from over-predicting errors, while low recall indicates a bias toward marking steps as correct. For example, Qwen3-Omni achieves perfect recall but near-zero MCC, reflecting an error-prediction bias, whereas several other models miss most incorrect steps. Thus, the binary results should be interpreted through balanced metrics. The low MCC values indicate that models still fail to reliably distinguish correct from incorrect procedural executions, even after ordinal rubric scoring is reduced to binary error detection.

Model	PLCC	95% CI	p-value
GPT-4o	0.158	[0.020, 0.290]	0.025
LLaVA-Video-7B	0.104	[-0.035, 0.239]	0.143
LLaVA-Video-72B	-0.047	[-0.185, 0.092]	0.509

Table 11: Statistical reliability of representative overall-score PLCC results. Confidence intervals are computed using Fisher's z transformation over 200 evaluation videos. The p-value tests whether each PLCC is significantly different from zero.

Method	κ	MCC	P	R	F1
Rubric-min	0.000	0.000	0.186	1.000	0.314
Rubric-mean	0.060	0.000	0.186	1.000	0.314
Rubric-max	0.000	0.000	0.000	0.000	0.000

Table 12: Performance of video-free rubric-prior controls. These controls predict fixed scores from rubric score ranges without observing the video. Incorrect steps are treated as the positive class in binary error detection.

G.5 Statistical Reliability of Overall-score Correlations

Because overall-score correlations are computed over a finite evaluation set, small PLCC differences should not be over-interpreted as reliable pairwise model rankings. We compute 95% confidence intervals using Fisher's z transformation over the 200 evaluation videos and test whether each PLCC differs from zero.

Table 11 shows that GPT-4o obtains a small but statistically non-zero correlation with expert overall scores, whereas the selected LLaVA-Video results are not significantly different from zero. However, the confidence intervals are wide and overlapping. These results should therefore be interpreted as evidence of weak overall-score agreement, rather than as statistically reliable pairwise superiority among models. The main conclusion is that all evaluated MLLMs remain far from expert-level agreement, particularly at the step level.

G.6 Video-free Rubric-prior Controls

To contextualize the MLLM results, we add three video-free rubric-prior controls. These controls do not observe the input video and make fixed predictions only from the valid score range of each rubric item. Specifically, Rubric-min always predicts the minimum score allowed by each rubric item, Rubric-max always predicts the maximum score, and Rubric-mean predicts the midpoint of the allowed score range. The predicted scores are then evaluated using the same step-level score mapping and binary error-detection protocol as the MLLMs. In the binary setting, we treat an incorrect step, i.e., any step with score deduction, as the positive class. Since these fixed predictions have zero variance at the overall-score level, PLCC and SRCC are undefined. We therefore report weighted Cohen's κ for step-level agreement, together with MCC, precision, recall, and F1 for binary error detection.

Table 12 shows that rubric priors alone provide little agreement with physician annotations. Rubric-min and Rubric-max obtain zero weighted κ, while Rubric-mean yields only a small non-zero κ of 0.060. This is far below the physician-physician step-level agreement of 0.732, indicating that SiMing-Bench is not inherently unmeasurable, but cannot be solved by fixed rubric-range priors.

The binary metrics further reveal the effect of label imbalance. Rubric-min and Rubric-mean mark every step as incorrect because their predicted scores are below the maximum score for each rubric item. This gives perfect recall by construction, but precision is only 0.186, matching the incorrect-step rate, and MCC remains 0.000. Rubric-max always predicts full credit and therefore never detects an incorrect step. These results show that high recall or non-trivial F1 can arise from degenerate prediction priors, whereas MCC and κ better reflect genuine step-level discrimination.

Taken together, these controls separate two issues that would otherwise be conflated: physician-physician agreement shows that experts can apply the rubric consistently, yet rubric and label priors alone provide almost no discriminative signal. The remaining gap therefore points to a substantive limitation in current MLLMs' interaction-grounded procedural judgment in videos.

H Extended Ethical Considerations

This work uses video data collected from real clinical skill examinations involving medical students. The data collection and research use protocols were reviewed and approved by the Institutional Review Board (IRB) of the participating institution. All data were collected under institutional oversight in standard clinical training and examination settings and were used solely for education, training, and research purposes. All data were handled under approved ethical procedures, with appropriate

measures for privacy protection, controlled access, and research-only use. A screenshot of the IRB approval document is provided in Figure 13.

I Ongoing Expansion of SiMing-Bench

SiMing-Bench is currently instantiated with three emergency-care skills, namely cardiopulmonary resuscitation (CPR), automated external defibrillator (AED) operation, and bag-mask ventilation (BMV). This design provides a controlled setting for evaluating rubric-grounded procedural judgment under standardized clinical examination conditions. However, the current instantiation does not exhaust the broader space of clinical skill assessment. To improve procedural coverage while preserving the core evaluation principle of SiMing-Bench, we are expanding the benchmark to additional clinical skill examination videos.

The ongoing expansion includes suture removal, lumbar puncture, abdominal paracentesis, and thoracentesis. As illustrated in Figure 18, these procedures extend the current emergency-care workflow to a broader set of wound-care and invasive puncture-based skills. They introduce different anatomical sites, instruments, sterile-field requirements, localization strategies, tissue- or cavity-related interactions, and post-procedure handling. At the same time, they retain the same assessment structure as the current benchmark: each video records a complete student-performed clinical skill examination, each procedure is governed by an expert-defined step-wise rubric, and procedural correctness must be judged from how earlier interactions establish, maintain, or violate the conditions required by later steps.

Suture removal emphasizes wound inspection, disinfection, instrument selection, suture cutting, removal, and final wound handling. Its correctness depends on whether the local wound state and instrument-tissue interactions are handled appropriately throughout the procedure, rather than on isolated recognition of a single manipulation. Lumbar puncture introduces patient positioning, sterile preparation, anatomical landmark identification, puncture execution, sample-related handling, and post-procedure care. This procedure requires tracking whether positioning, aseptic operations, and landmark-dependent actions are performed in the correct order and under the required procedural conditions. Abdominal paracentesis involves pre-procedure preparation, puncture-site localization, sterile-field maintenance, needle or catheter manipulation, fluid-related handling, and post-puncture management. Correctness therefore depends on whether earlier preparation and localization steps establish the conditions for safe and valid puncture-related actions. Thoracentesis similarly requires patient positioning, puncture-site identification, sterile preparation, controlled instrument insertion, drainage-related operations, and post-procedure handling, where local actions must be interpreted with respect to the evolving anatomical, aseptic, and instrument-use state.

Importantly, these additional procedures are intended to expand the diversity of SiMing-Bench rather than change its evaluation target. Across both the current and expanded settings, SiMing-Bench is not designed to test broad medical knowledge in isolation. Instead, it focuses on whether a model can use the full interaction history to determine whether each rubric-defined action is procedurally valid under the evolving examination state. The expanded procedures therefore preserve the same core properties as the current benchmark: full-workflow videos, standardized clinical examination settings, physician-defined step-wise rubrics, and expert annotation of procedural correctness.

The expanded data are not included in the experiments reported in the current version. We describe them here to clarify the intended scope and development trajectory of SiMing-Bench. Accordingly, the current results should be interpreted as evidence about interaction-grounded procedural judgment within structured emergency-care skill assessment, while the ongoing expansion is expected to support broader evaluation across a more diverse set of clinical procedures in future releases.



Figure 6: Standardized CPR rubric defined by physicians and used for step-level annotation in our benchmark.



Figure 7: Standardized AED rubric defined by physicians and used for step-level annotation in our benchmark.

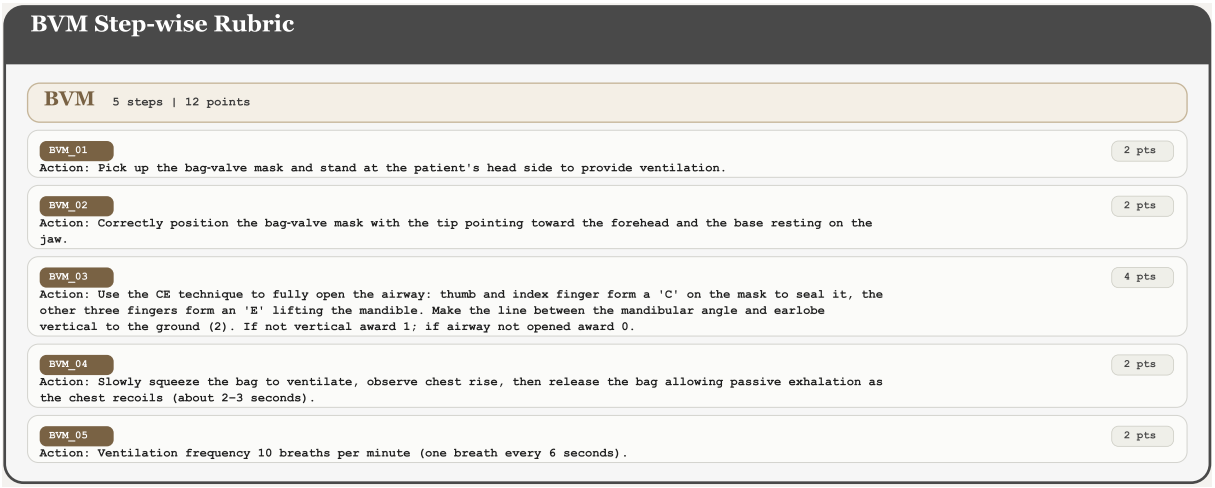


Figure 8: Standardized BMV rubric defined by physicians and used for step-level annotation in our benchmark.

Step-wise BLS Assessment Prompt

```
You are an expert OSCE examiner for Basic Life Support (BLS) / CPR / AED / BVM.

## Your Job

Given uniformly sampled video frames with timestamps and a step-wise rubric, assign a score for EACH rubric step.

## Critical Scoring Rules

Visual evidence only

use ONLY the provided frames and timestamps. Do not guess or assume actions happened off-camera.

Per-step scoring

if the step is clearly and correctly performed, give full points (max_points). If partially performed or unclear, give
an intermediate integer score. If missing, incorrect, unsafe, out of order, or not verifiable, give 0.

Integer range

the score must be an integer in [0, max_points]. If max_points is not an integer, round DOWN the allowed maximum to the
nearest integer.

Sub-requirements

if a step includes multiple required parts, award full points only when ALL required sub-parts are satisfied and visible
or confirmable.

Order and mapping

output exactly one item per rubric step. step_id must match the rubric exactly, and the original rubric order must be
preserved.

Unverifiable steps

if a step cannot be confirmed from the frames because of occlusion, camera angle, or low confidence, score it as 0, not
unknown.

## Output Constraints

Output ONLY one JSON object. No markdown. No explanations. No extra keys.
Do not include comments, rationale, confidence, or any text outside the JSON.

## Prompt Assembly

The final prompt includes the schema instruction, the step-task instruction, the segment duration, sampled frame timestamps,
video metadata, and the rubric step to score.

Dynamic fields:

- Video ID: <video_id>
- View angle: <view_angle>
- Segment duration: <video_time> seconds
- Sampled frames at: <frame_time_str>

## Rubric Step Input

{
"step_id": "<exact rubric step id>",
"step_name": "<step name>",
"max_points": <number>,
"criteria": "<step description or requirements>"
}

## Final Reminder

Remember: output ONLY one JSON object.
```

Figure 9: Prompt template for full-video step-wise scoring.

Step-wise Error Detection Prompt

```
You are an expert clinical skills examiner for Basic Life Support (BLS), CPR, AED, and BVM assessment.

## JSON Schema
{
"video_id": "<string>",
"view_angle": "<string>",
"model_score": {
"steps": [
{"step_id": "<string>", "score": <integer>}

 }
## Task Objective
You will review uniformly sampled video frames and a step-wise rubric. Your task is not to assign points. Instead, determine
whether each rubric step contains an error.

## Binary Label Meaning
score = 1 means no error: the step is correctly performed and the key evidence is clearly visible.
score = 0 means has error: the step is missing, incorrect, unsafe, out of order, or cannot be verified.

## Evidence Rules
Use only the visible evidence in the provided frames and timestamps.
Do not infer that an action occurred off-camera.
If the visual evidence is insufficient to confirm correct completion, output 0.
Treat unverifiable steps as errors.

## Output Requirements
score must be either 0 or 1.
score is an error flag, not a point score.
You must output exactly one item per rubric step.
step_id must match the rubric exactly, and the original rubric order must be preserved.
Do not output explanations, comments, confidence, markdown, or extra keys.

## Prompt Assembly
The final prompt includes the schema instruction, task instruction, segment duration, sampled frame timestamps, rubric steps,
video ID, and view angle.

Dynamic fields:
- Video ID: <video_id>
- View angle: <view_angle>
- Segment duration: <video_time> seconds
- Sampled frames at: <frame_time_str>

## Rubric Input
{
"steps": [
{"step_id": "<string>", "step_name": "<string>", "max_points": <number>}

}
## Final Reminder
Judge only whether each step has an error. Output one JSON object only.
```

Figure 10: Prompt template for full-video step-wise error detection.

Video Clips Scoring Prompt

You are an expert OSCE examiner for Basic Life Support (BLS), CPR, AED, and BVM.

JSON Schema

```
{
"step_id": "<string>",
"score": <integer>
}
```

Schema Rules

step_id must match the given step exactly.
score must be an integer in [0, floor(max_points)].
Do not add any extra keys.

Task Objective

Given uniformly sampled video frames with timestamps and a step-wise rubric, assign a score for the target rubric step.

Critical Scoring Rules

Visual evidence only

use ONLY the provided frames and timestamps. Do not guess or assume actions happened off-camera.

Per-step scoring

if the step is clearly and correctly performed, give full points (max_points). If partially performed or unclear, give an intermediate integer score. If missing, incorrect, unsafe, out of order, or not verifiable, give 0.

Integer range

the score must be an integer in [0, max_points]. If max_points is not an integer, round DOWN the allowed maximum to the nearest integer.

Sub-requirements

if a step includes multiple required parts, award full points only when ALL required sub-parts are satisfied and visible or confirmable.

Unverifiable steps

if a step cannot be confirmed from the frames because of occlusion, camera angle, or low confidence, score it as 0, not unknown.

Output Constraints

Output ONLY one JSON object. No markdown. No explanations. No extra keys.
Do not include comments, rationale, confidence, or any text outside the JSON.

Prompt Assembly

The final prompt includes the schema instruction, the task instruction, the segment duration, sampled frame timestamps, video metadata, and the rubric step to score.

Dynamic fields:

- Video ID: <video_id>
- View angle: <view_angle>
- Segment duration: <video_time> seconds
- Sampled frames at: <frame_time_str>

Rubric Step Input

```
{
"step_id": "<exact rubric step id>",
"step_name": "<step name>",
"max_points": <number>,
"criteria": "<step description or requirements>"
}
```

Final Reminder

Rubric step to score: ONLY this step. Output one JSON object only.

Figure 11: Prompt template for step-aligned clip scoring.

Video Clips Error Detection Prompt

```
You are an expert OSCE examiner. This is an error detection task for a single rubric step.

## JSON Schema

{
"model_score": {
"steps": [

"step_id": "<STEP_ID>",
"has_error": <0_or_1>

}

## Schema Rules

step_id must match the given step exactly.
has_error must be an integer and must be either 0 or 1.
Do not output score, points, max_points, or similar fields.
Do not add any extra keys.

## Task Objective

Judge whether this one rubric step contains an error based ONLY on the visual evidence in the frames.

## Label Definition

has_error = 1: the step is missing, incorrect, unsafe, performed in the wrong order, or not supported by the visual evidence.
has_error = 0: the step is clearly and correctly performed, and the evidence is visible in the frames.

## Decision Rules

Base the decision strictly on visible evidence in the frames.
Do not assume actions that are not visible.
If the step is clearly performed correctly, output has_error = 0.
If the step is clearly wrong or missing, output has_error = 1.

## Output Constraints

This is NOT a scoring task. Do not output score, points, or max_points.
Output ONLY one JSON object.
No markdown.
No explanations.
No extra text.

## Prompt Assembly

The final prompt includes the schema instruction, the task instruction, segment metadata, sampled frame timestamps, and the rubric step to
classify.

Dynamic fields:
- Video ID: <video_id>
- View angle: <view_angle>
- Segment duration: <video_time> seconds
- Sampled frames at: <frame_time_str>

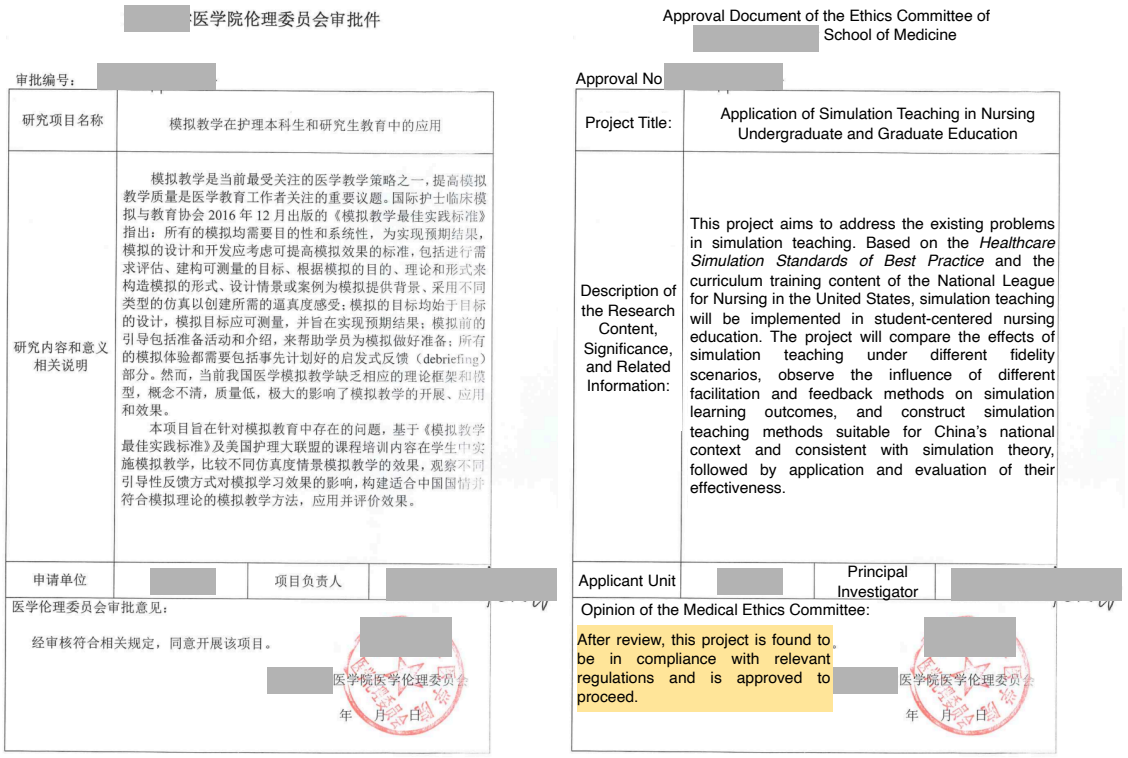
## Rubric Step Input

{
"step_id": "<exact rubric step id>",
"step_name": "<step name>",
"criteria": "<step description or requirements>"
}

## Final Reminder

Classify ONLY this step. Output one JSON object only.
```

Figure 12: Prompt template for step-aligned clip error detection.



[redacted]医学院伦理委员会审批件

审批编号：[redacted]

研究项目名称	模拟教学在护理本科生和研究生教育中的应用
研究内容和意义相关说明	模拟教学是当前最受关注的医学教学策略之一，提高模拟教学质量是医学教育工作者关注的重要议题。国际护士临床模拟与教育协会 2016 年 12 月出版的《模拟教学最佳实践标准》指出：所有的模拟均需要目的性和系统性，为实现预期结果，模拟的设计和开发应考虑可提高模拟效果的标准，包括进行需求评估、建构可测量的目标、根据模拟的目的、理论和形式来构造模拟的形式、设计情景或案例为模拟提供背景、采用不同类型的仿真以创建所需的逼真度感受；模拟的目标均始于目标的设计，模拟目标应可测量，并旨在实现预期结果；模拟前的引导包括准备活动和介绍，来帮助学员为模拟做好准备；所有的模拟体验都需要包括事先计划好的启发式反馈（debriefing）部分。然而，当前我国医学模拟教学缺乏相应的理论框架和模型，概念不清，质量低，极大的影响了模拟教学的开展、应用和效果。 本项目旨在针对模拟教育中存在的问题，基于《模拟教学最佳实践标准》及美国护理大联盟的课程培训内容在学生中实施模拟教学，比较不同仿真度情景模拟教学的效果，观察不同引导性反馈方式对模拟学习效果的影响，构建适合中国国情并符合模拟理论的模拟教学方法，应用并评价效果。
申请单位	[redacted] 项目负责人 [redacted]

医学伦理委员会审批意见：

经审核符合相关规定，同意开展该项目。

[redacted]医学院医学伦理委员会

年 月 日

Approval Document of the Ethics Committee of [redacted] School of Medicine

Approval No [redacted]

Project Title:	Application of Simulation Teaching in Nursing Undergraduate and Graduate Education
Description of the Research Content, Significance, and Related Information:	This project aims to address the existing problems in simulation teaching. Based on the *Healthcare Simulation Standards of Best Practice* and the curriculum training content of the National League for Nursing in the United States, simulation teaching will be implemented in student-centered nursing education. The project will compare the effects of simulation teaching under different fidelity scenarios, observe the influence of different facilitation and feedback methods on simulation learning outcomes, and construct simulation teaching methods suitable for China's national context and consistent with simulation theory, followed by application and evaluation of their effectiveness.
Applicant Unit	[redacted] Principal Investigator [redacted]

Opinion of the Medical Ethics Committee:

After review, this project is found to be in compliance with relevant regulations and is approved to proceed.

[redacted]医学院医学伦理委员会

年 月 日

Figure 13: Screenshot of the IRB approval document.

Medical Knowledge Enhanced Prompt

```
You are a strict and professional medical OSCE examiner. Score all given BLS procedure steps one by one according to the
video content.

## Dynamic fields

The current assessment video_id is <video_id>.
The current view_angle is <view_angle>.
The video duration is approximately <video_time> seconds.
The input is represented by uniformly sampled video frames, with timestamps: <frame_time_str>.
This request requires scoring <N> steps in one pass.

## Medical background block

MEDICAL_BACKGROUND = """
Medical background description:
This video is a BLS OSCE skills examination. It mainly includes three parts: cardiopulmonary resuscitation, AED use, and
bag-valve-mask ventilation.

1. Cardiopulmonary resuscitation:
Focus on scene safety assessment, patient responsiveness assessment, calling for help and obtaining an AED, assessing
breathing and carotid pulse within 5-10 seconds, chest compression location, rate, depth, recoil, interruption time, airway
opening, and rescue breathing. High-quality CPR usually requires correct hand position, straight arms, vertical downward
compression, a rate of about 100-120/min, a depth of about 5-6 cm, and minimal interruptions.

2. AED:
Focus on whether the AED is used promptly after it is obtained, whether it is powered on, whether pad placement is correct,
whether the connector is attached, whether everyone is warned and confirmed clear during analysis and shock delivery, and
whether CPR is resumed immediately after shock. No one should touch the patient during AED analysis or shock delivery.

3. Bag-valve-mask ventilation:
Focus on whether the rescuer is positioned at the patient's head side, whether the mask orientation and placement are
correct, whether the CE technique is used to secure the mask and lift the jaw, whether the airway is adequately opened,
whether visible chest rise and natural recoil occur after squeezing the bag, and whether the ventilation rate is
appropriate.

Note:
The medical background is only used to help interpret the rubric. Final scoring must strictly follow the given rubric and
visible evidence in the video, and must not add new scoring requirements.
""".strip()

## Prompt setting

Below is the original physician-defined rubric. Please score strictly according to these steps:

## Rubric input

{
  "total_steps": <N>,
  "steps": [
    {
      "step_id": "CPR_01",
      "task": "CPR",
      "step_name": "<full step description from the original physician rubric>",
      "max_points": <number>
    }
  ]
}

## General scoring rules

1. Score every step listed in the rubric.
2. Score each step independently; do not count performance from other steps toward the current step.
3. Each step can receive partial credit. If only part of the rubric requirement is completed, assign a score between 0 and
   max_points according to the visible completion level.
4. If the rubric explicitly specifies sub-item points, add the sub-item points as much as possible. If no sub-item points
   are specified, estimate the step score by the overall completion quality.
5. The score must not exceed max_points. When evidence is ambiguous, choose the lower score consistent with the visible
   evidence.
6. If the video only shows preparation but not the key completion evidence, do not award full points.
7. If a step is completely invisible in the sampled frames, assign 0 or a very low score.

## Strict output requirements

Output exactly one valid JSON object. Do not output Markdown, do not explain, and do not output reasoning.
score must be a numeric value; each step_id must appear exactly once.
Score only actions that are clearly visible in the video frames and supported by continuous visual evidence.
If the action is not visible, occluded, insufficiently supported, or cannot be confirmed, treat it as incomplete or
low-scoring. Do not infer missing actions from procedural common sense.

## Output JSON

{
  "model_score": {
    "steps": [
      {"step_id": "CPR_01", "score": 0.0},
      {"step_id": "CPR_02", "score": 0.0}
    ]
  }
}

## Final instruction

Output scores for all steps in one pass. The output must include <step_id_list>, and each step_id may appear only once.
Each score must be between 0 and the corresponding max_points, and may be an integer or a half-point increment.
Reminder: output JSON only. Do not explain. Do not output reasoning.
```

Figure 14: Prompt template for the medical-knowledge-enhanced setting. The prompt prepends concise clinical background to the original scoring instruction while keeping the video input, step-wise rubric, and scoring protocol unchanged.

Simplified Rubric All-Steps Prompt

```
You are a strict and professional medical OSCE examiner. Score all given BLS procedure steps one by one according to the
video content.

## Dynamic fields

The current assessment video_id is <video_id>.
The current view_angle is <view_angle>.
The video duration is approximately <video_time> seconds.
The input is represented by uniformly sampled video frames, with timestamps: <frame_time_str>.
This request requires scoring <N> steps in one pass.

## Prompt setting

Below is the simplified operational rubric. It preserves the original scoring meaning and max_points, but the wording is
shorter and clearer.
Please score strictly according to these steps:

## Rubric input

{
  "total_steps": <N>,
  "steps": [
    {
      "step_id": "CPR_01",
      "task": "CPR",
      "step_name": "<simplified operational requirement>",
      "max_points": <number>
    }
  ]
}

## General scoring rules

1. Score every step listed in the rubric.
2. Score each step independently; do not count performance from other steps toward the current step.
3. Each step can receive partial credit. If only part of the rubric requirement is completed, assign a score between 0 and
   max_points according to the visible completion level.
4. If the rubric explicitly specifies sub-item points, add the sub-item points as much as possible. If no sub-item points
   are specified, estimate the step score by the overall completion quality.
5. The score must not exceed max_points. When evidence is ambiguous, choose the lower score consistent with the visible
   evidence.
6. If the video only shows preparation but not the key completion evidence, do not award full points.
7. If a step is completely invisible in the sampled frames, assign 0 or a very low score.

## Strict output requirements

1. Output exactly one valid JSON object. Do not output Markdown, do not use ```json fences, and do not explain.
2. This task is scoring, not binary correctness judgment; score must be a numeric value, not a 0/1 class label.
3. For every step_id in the rubric, output exactly one corresponding score.
4. Each score must be between 0 and the corresponding max_points, and may be an integer or a half-point increment.
5. Score only from actions that are clearly visible in the video frames and supported by continuous visual evidence.
6. If the action is not visible, occluded, insufficiently supported, or cannot be confirmed, treat it as incomplete or
   low-scoring. Do not infer missing actions from procedural common sense.
7. Do not output reasoning and do not add extra fields.

## Output JSON

{
  "model_score": {
    "steps": [
      {"step_id": "CPR_01", "score": 0.0},
      {"step_id": "CPR_02", "score": 0.0}
    ]
  }
}

## Final instruction

Output scores for all steps in one pass.
The output must include the following step IDs, and each step_id may appear only once: <step_id_list>
Each score must be between 0 and the corresponding max_points, and may be an integer or a half-point increment.
Reminder: output JSON only. Do not explain. Do not output reasoning.
```

Figure 15: Prompt template for the rubric-simplified setting. The prompt rewrites the physician-defined rubric into shorter, more operational criteria while preserving the original scoring semantics.

CoT Generic BLS Scoring Prompt

```
You are a professional medical OSCE examiner. Your task is to score the BLS procedure in the video according to the given
rubric.

## Dynamic fields

The current assessment video_id is <video_id>.
The view_angle is <view_angle>.
The video lasts for <video_time> seconds. Several frames are uniformly sampled from the video.
The sampled frame timestamps are: <frame_time_str>.

## Rubric input

{
  "total_steps": <number>,
  "steps": [
    {
      "step_id": "CPR_01",
      "step_name": "<step name>",
      "max_points": <number>,
      "allowed_scores": [0.0, 0.5, 1.0, "..."]
    }
  ]
}

## General scoring rules

1. Score each rubric step independently.
2. The score for each step must be between 0 and max_points.
3. Use only valid scores listed in allowed_scores for each step.
4. Base the score only on clearly visible evidence in the video frames.
5. If the action is not visible, unclear, blocked, or cannot be confidently verified, assign 0 for that step.
6. Do not give credit for actions that are only implied but not visible.
7. Do not assign a score greater than max_points.
8. Before finalizing, check every step and correct any score that exceeds max_points.

## Important output rules

- All reasoning fields must be written in English.
- The final output must be a single valid JSON object.
- Do not output Markdown or wrap the JSON in code fences.
- Do not output any text before or after the JSON object.

## Generic CoT instruction

For each rubric step, reason briefly before assigning the score.
For every step, output:
- step_id
- reasoning: one or two short sentences explaining the visible evidence and why it supports the assigned score
- score: the final score

## Output JSON format

{
  "video_id": "<string>",
  "view_angle": "<string>",
  "model_score": {
    "steps": [
      {
        "step_id": "CPR_01",
        "reasoning": "Brief evidence-based reasoning for this score.",
        "score": 1.0
      }
    ]
  }
}

## Final check

You must output exactly <N> step results, one for each rubric step.
The step order should follow the rubric order.
The output JSON must contain video_id = "<video_id>" and view_angle = "<view_angle>".
```

Figure 16: Prompt template for the generic reasoning setting, where the model is instructed to reason step by step before assigning rubric scores.

State-Aware CoT BLS Scoring Prompt

```
You are a professional medical OSCE examiner. Your task is to score the BLS procedure in the video according to the given
rubric.

## Dynamic fields

The current assessment video_id is <video_id>.
The view_angle is <view_angle>.
The video lasts for <video_time> seconds. Several frames are uniformly sampled from the video.
The sampled frame timestamps are: <frame_time_str>.

## Rubric input

{
  "total_steps": <number>,
  "steps": [
    {
      "step_id": "CPR_01",
      "step_name": "<step name>",
      "max_points": <number>,
      "allowed_scores": [0.0, 0.5, 1.0, "..."]
    }
  ]
}

## General scoring rules

1. Score each rubric step independently.
2. The score for each step must be between 0 and max_points.
3. Use only valid scores listed in allowed_scores for each step.
4. Base the score only on clearly visible evidence in the video frames.
5. If the action is not visible, unclear, blocked, or cannot be confidently verified, assign 0 for that step.
6. Do not give credit for actions that are only implied but not visible.
7. Do not assign a score greater than max_points.
8. Before finalizing, check every step and correct any score that exceeds max_points.

## State-aware CoT instruction

For each rubric-defined step, do not score immediately. First analyze the procedural state in a structured way.
For every step, output:
- step_id
- required_condition: what condition must be established before this step can be judged correct
- observed_interaction: what action or interaction is clearly visible in the video frames
- state_update: how the observed interaction changes the procedural state
- rubric_alignment: whether the updated state satisfies the rubric criterion
- score: the final score

## Output JSON format

{
  "video_id": "<string>",
  "view_angle": "<string>",
  "model_score": {
    "steps": [
      {
        "step_id": "CPR_01",
        "required_condition": "The required procedural condition for this step.",
        "observed_interaction": "The visible action or interaction in the video frames.",
        "state_update": "How the procedural state changes after this interaction.",
        "rubric_alignment": "Whether the state satisfies the rubric criterion.",
        "score": 1.0
      }
    ]
  }
}

## Rules

- Each text field must be concise, preferably one short sentence.
- Do not write long chain-of-thought paragraphs.
- For each rubric step, output exactly one corresponding item in model_score.steps.
- The step_id values must exactly match the rubric step_id values.
- Do not add any step that is not in the rubric.

## Final check

You must output exactly <N> step results, one for each rubric step.
The step order should follow the rubric order.
The output JSON must contain video_id = "<video_id>" and view_angle = "<view_angle>".
```

Figure 17: Prompt template for the state-aware reasoning setting, where the model is guided to organize its prediction around rubric conditions, observed interactions, interaction-induced state updates, and state-conditioned correctness.



Figure 18: Multi-frame examples of the additional clinical skills under ongoing expansion, including suture removal, lumbar puncture, abdominal paracentesis, and thoracentesis. These procedures broaden the procedural coverage of SiMing-Bench from emergency resuscitation skills to wound-care and invasive puncture-based examinations, introducing more diverse instruments, anatomical regions, sterile operations, and interaction-dependent error modes while preserving the benchmark's focus on full-workflow, rubric-grounded procedural correctness from continuous interactions.