VerbaNexAI at ClinicalSkillQA: A Two-Stage Generative Vision-Language Framework for Procedural Frame Ordering

Andrea Menco-Tovar and Jairo E. Serrano and Edwin Puertas and Juan Carlos Martinez-Santos
Universidad Tecnologica de Bolivar
Cartagena, Colombia
amenco@utb.edu.co, jserrano@utb.edu.co, epuerta@utb.edu.co
jcmartinezs@utb.edu.co

Abstract

This work addresses the temporal ordering task of clinical frames in the Basic Life Support (BLS) subset of ClinSkillQA. A two-stage hybrid pipeline based on Qwen2-VL-2B-Instruct in a zero-shot configuration is proposed. In Stage 1, each image is processed independently to extract factual visual evidence, which is then transformed, using deterministic rules, into a structured representation. In Stage 2, ordering is formulated as an ordinal scoring task over procedural stages, with ties broken using PCA applied to multimodal embeddings. Evaluation followed the official benchmark protocol, considering Task Accuracy, Pairwise Accuracy, and BERTScore. In the test phase, the system achieved Task Accuracy = 0.17, Pairwise Micro Accuracy = 0.60, and BERT F1 = 0.71, with complete coverage in both predictions and rationales. The results demonstrate an interpretable and reproducible foundation, although challenges in fine-grained temporal discrimination remain.

1 Introduction

Clinical skills assessment is a central component of medical training because it evaluates not only declarative knowledge but also the accurate and sequential execution of procedures according to predefined standards. Instruments such as the Objective Structured Clinical Examination (OSCE) support structured assessment of clinical competence, but their large-scale implementation remains limited by time, infrastructure, and expert availability (Issenberg et al., 2005; Vermylen et al., 2025). In this context, evidence from simulation-based education and deliberate practice has shown that repeated procedural training improves skill acquisition and retention, particularly in critical scenarios such as resuscitation and basic life support (BLS), where both action quality and procedural order are essential (Khanghahi and Azar, 2018; Panchal et al., 2020; McGaghie et al., 2011).

The digitalization of training environments and the growing availability of image and video data have created opportunities for automated support in clinical skills assessment. Large multimodal models are especially relevant because they can integrate visual and textual information for clinical image interpretation, guided analysis, and feedback generation (Li et al., 2023; Liu et al., 2023; OpenAI, 2024). However, most biomedical multimodal benchmarks, including VQA-RAD, SLAKE, and PMC-VQA, focus mainly on visual question answering over static medical images (Lau et al., 2018; Liu et al., 2023). Although these resources have advanced medical multimodal AI, they do not explicitly evaluate temporal progression, procedural state transitions, or the reconstruction of ordered clinical actions. Similar limitations persist in video-based assessment of surgical and procedural skills, where temporal granularity, generalization, and reliable interpretation remain challenging (AlSaad et al., 2024; Hartsock and Rasool, 2024; Liu et al., 2021; Seenivasan et al., 2022; Zhang et al., 2024).

The ClinSkillQA 2026 challenge addresses this gap by formulating clinical-skill understanding as a frame-ordering and explanation task. Given shuffled keyframes extracted from videos of student-performed clinical procedures, the system must reconstruct the correct sequence of actions and provide rationales aligned with expert reasoning (Huang et al., 2026). This setting moves beyond isolated visual recognition, requiring models to infer precedence relationships between procedural states while generating explanations that can support formative feedback. Accordingly, the benchmark evaluates both ordering performance, using sequence-level and pairwise accuracy, and rationale quality, using BERTScore and an LLM-as-judge strategy based on G-Eval (Zhang et al., 2024; Li et al., 2023; Huang et al., 2026).

In this work, we present our participation in ClinSkillQA through a two-stage multimodal frame-

Proceedings of the BioNLP 2026 (Shared Tasks), pages 6–12
July 3-4, 2026 ©2026 Association for Computational Linguistics

work for ordering BLS clinical keyframes and generating verifiable rationales. The proposed approach combines structured visual evidence extraction, ordinal procedural scoring, and latent-space tie-breaking to handle ambiguities between visually similar states. Beyond reporting challenge performance, this study analyzes the potential and limitations of multimodal models for procedural assessment tasks requiring temporal reasoning, explanatory consistency, and alignment with clinical expert criteria.

2 Background

Clinical skills assessment is a central component of medical education, since professional competence depends not only on declarative knowledge but also on the accurate, timely, and sequential execution of procedures. Evidence from simulation-based education shows that high-fidelity simulation, deliberate practice, competency-based training, and structured observational tools such as Direct Observation of Procedural Skills (DOPS) can support the acquisition and assessment of procedural competence. These findings highlight the need to evaluate not only whether a clinical action is performed, but also whether it follows the appropriate procedural order and fidelity (Issenberg et al., 2005; Khanghahi and Azar, 2018; McGaghie et al., 2011; Vermylen et al., 2025).

Recent advances in multimodal large language models (MLLMs) and vision-language models (VLMs) have expanded the possibilities for automating clinically relevant visual-textual reasoning tasks. In healthcare, these models have been increasingly applied to image understanding, clinical assistance, report generation, visual question answering, and multimodal decision support. Nevertheless, most medical VQA benchmarks and medical VLMs remain centered on single-image reasoning or localized diagnostic contexts, particularly in radiology. Even procedure-oriented datasets, such as those developed for surgical VQA, tend to emphasize scene understanding or question answering rather than reconstructing a coherent temporal sequence of actions from shuffled visual observations. Similarly, conversational biomedical assistants demonstrate the feasibility of open-ended multimodal interaction, but they do not directly address procedural ordering from fragmented visual evidence (AlSaad et al., 2024; Hartsock and Rasool, 2024; Lau et al., 2018; Zhang et al., 2024; Seenivasan et al., 2022; Li et al., 2023).

ClinSkillQA addresses this gap by formulating clinical-skill understanding as a frame-ordering task over shuffled keyframes. In this setting, models must arrange frames into a coherent sequence of clinical actions and generate explanations that justify the predicted order. The benchmark includes 200 sets of shuffled keyframes from three types of clinical-skills videos, each accompanied by reference orderings and expert-annotated rationales. Its evaluation is explicitly dual, combining ordering metrics, namely Task Accuracy and Pairwise Accuracy, with explanation-quality metrics based on BERTScore and an LLM-as-a-Judge scheme using G-Eval (Huang et al., 2026).

Against this background, a two-stage generative vision-language framework provides a suitable response to the task. By separating per-frame visual evidence extraction from cross-frame procedural reasoning, the proposed approach first captures observable and reusable cues from each image and then uses them to infer the most coherent clinical sequence and generate set-level rationales. This decomposition aligns with the evaluation structure of ClinSkillQA and connects recent progress in multimodal artificial intelligence with the need for scalable, structured, and procedure-faithful clinical skills assessment.

3 System Overview

The proposed system addresses the ClinSkillQA frame-ordering task as a two-stage hybrid pipeline for Basic Life Support (BLS) procedural sequences. Each input sample contains an unordered set of four or six clinical frames, and the objective is to infer the most plausible procedural order from visual evidence. As shown in Figure 1, the system combines frame-level semantic evidence extraction with sample-level ordinal ranking and tie resolution.

The pipeline as shown in Figure 1, begins by loading the benchmark input file, `BLS_input.json`, and resolving the relative image paths to their corresponding disk locations. This step creates a frame-level index containing the `sample_id`, `frame_id`, and image path, allowing frames to be processed independently in the first stage and later regrouped by sample for sequence prediction.

3.1 Stage 1

In the first stage, each frame is analyzed independently using Qwen2-VL-2B-Instruct (Wang et al., 2024). The model is prompted to produce a short factual description limited to visible content, avoiding speculation and explicit temporal reasoning. The prompt requests three to five bullet points describing observable cues such as hand placement, body posture, chest exposure, clothing condition, and visible equipment.

The generated text is converted into a structured intermediate representation through deterministic post-processing rules. This representation includes detected entities, detected actions, a coarse procedural state, key visual evidence, uncertainty flags, and a short caption summary. The procedural state is inferred from lexical cues in the description: for instance, references to CPR or active compression are mapped to `compressions`, hands placed on the chest without clear downward motion to `position_hands`, and clothing opening or thorax exposure to `expose_chest`. When the evidence is insufficient, the frame is labeled as `other_or_unclear`. All outputs are cached at the frame level to avoid redundant inference.

Rather than producing the final order directly, Stage 1 provides an interpretable semantic layer that summarizes the visual content of each frame. This evidence is later used as auxiliary guidance for ranking and as the basis for textual rationales.

3.2 Stage 2

In the second stage, the same vision-language model is used under a constrained ordinal scoring formulation. Each frame receives a single score from 0 to 5, corresponding to the following procedural states: 0=`approach`, 1=`prepare`, 2=`expose_chest`, 3=`position_hands`, 4=`compressions`, and 5=`check_or_adjust`. The model receives the image and a prompt requiring a single-digit answer, with the Stage 1 caption optionally appended as a hint. This converts frame ordering into ordinal state estimation, which is more stable than direct pairwise comparison.

The predicted score is automatically parsed from the model output. If no valid digit is obtained, the system performs a stricter retry; if parsing still fails, a conservative fallback score is assigned. Frames are then grouped by sample and sorted in ascending order according to their predicted procedural score.

Because multiple frames may receive the same score, the pipeline applies an embedding-based tie-breaking step. For tied frames only, hidden-state representations are extracted from the multimodal model, averaged into one vector per image, and projected onto a one-dimensional axis using Principal Component Analysis (PCA). Frames within the tied group are then ordered according to their PCA projection values. The final sequence is obtained by concatenating the ordered groups from lower to higher procedural stages.

For each sample, the system exports the predicted frame order together with an auxiliary rationale derived from the Stage 1 evidence. In this way, the framework integrates frame-level visual grounding, ordinal procedural ranking, and interpretable evidence traces for each prediction.

4 Experimental Setup

Experiments were conducted on the Basic Life Support (BLS) subset of ClinSkillQA, using the temporal ordering task over shuffled clinical frames. The benchmark was loaded from `BLS_input.json`, and a frame-level index was created with `sample_id`, `frame_id`, and the corresponding absolute image path. The final processed set included 200 samples: 150 samples with four frames and 50 samples with six frames, with no missing images after path verification.

The implementation was executed on Kaggle using PyTorch and Transformers with CUDA support on a Tesla T4 GPU. The base model was `Qwen/Qwen2-VL-2B-Instruct`, loaded in `float16` when GPU acceleration was available. All experiments were performed in a zero-shot setting, without task-specific fine-tuning, and deterministic decoding was used by setting `do_sample=False`.

In Stage 1, each image was processed independently to generate a factual description of visible clinical evidence. These outputs were then converted through lexical rules into a structured representation containing detected entities, actions, procedural state, key evidence, a short summary, and uncertainty markers. In Stage 2, the same model assigned each frame an ordinal score from 0 to 5, representing the expected progression of the BLS procedure. Frames within each sample were ordered according to this score, while ties were resolved using PCA over multimodal embeddings extracted from the model's last hidden layer.

For development and debugging, a DEV subset

of 10 samples was used, whereas the final run was performed on the complete BLS set. Intermediate and final outputs were stored in JSONL format, including the predicted sequence for each sample, denoted as `pred_order`, and a textual rationale derived from the Stage 1 evidence.

Following the official ClinSkillQA protocol, evaluation considers both ordering and explanation quality. Ordering performance is assessed using Task Accuracy, which measures exact sequence prediction, and Pairwise Accuracy, which evaluates correctly ordered adjacent pairs. Explanation quality is evaluated using BERTScore and an LLM-as-a-judge scheme based on G-Eval. Thus, the experimental setup produces both sequence predictions and textual rationales aligned with the official challenge format.

5 Results

Table 1 summarizes the visible test-phase leaderboard results for the ClinSkillQA benchmark, including the position achieved by VerbaNexAI Lab. Our system ranked fourth among the eight visible teams, obtaining an overall score of 37.96. At the ordering level, the method achieved a Task Accuracy of 0.17 and a Pairwise Micro Accuracy of 0.60. This indicates that the system was able to capture some local temporal relationships between frames, but still struggled to recover the complete procedural sequence exactly.

System	Metric	Rank
Score		
Highest visible system	71.43	1/8
VerbaNexAI Lab	*37.96*	*4/8*
Task Accuracy		
Highest visible system	0.63	1/8
VerbaNexAI Lab	*0.17*	*4/8*
Pairwise Micro Accuracy		
Highest visible system	0.86	1/8
VerbaNexAI Lab	*0.60*	*4/8*
BERT F1		
Highest visible system	0.79	1/8
VerbaNexAI Lab	*0.71*	*4/8*

Table 1: Compact summary of VerbaNexAI Lab performance in the visible test-phase leaderboard.

For explanation quality, the system obtained BERT Precision = 0.74, BERT Recall = 0.68, and BERT F1 = 0.71. In addition, both Predicted Coverage and Rationale Coverage reached 1.0, showing that the pipeline generated complete predictions and rationales for all evaluated samples. Compared with the highest visible BERT F1 value of 0.79, the semantic quality of the generated explanations was relatively competitive.

The main performance gap was therefore concentrated in temporal ordering rather than in rationale generation. While the best visible systems reached substantially higher Task Accuracy and Pairwise Micro Accuracy, our method showed a larger drop in exact sequence reconstruction. This suggests that the proposed zero-shot ordinal scoring strategy can identify partial procedural progression, but remains limited when distinguishing visually similar or contiguous BLS states.

Overall, the results show that the proposed framework can generate interpretable evidence traces and complete rationales, but accurate frame ordering remains the main bottleneck. Future improvements should therefore focus on strengthening temporal discrimination, reducing ambiguity between adjacent procedural states, and improving the tie-resolution mechanism used when frames receive similar ordinal scores.

6 Conclusion

This work presented a two-stage hybrid architecture for ordering clinical frames in the BLS subset of ClinSkillQA. The framework combines visual evidence extraction, ordinal procedural scoring, and embedding-based tie resolution to generate both sequence predictions and interpretable rationales. The results show that the system produced complete predictions and explanations for all evaluated samples. However, its main limitation was the exact reconstruction of procedural sequences, especially when adjacent clinical states were visually similar. This suggests that isolated visual recognition is not sufficient for robust procedural ordering. Overall, the study highlights the potential of evidence-driven vision-language reasoning for interpretable clinical skills assessment, while also showing the need for stronger temporal modeling to better distinguish fine-grained procedural transitions.

References

Rawan AlSaad, Ala Abd-Alrazaq, Sabah Boughorbel, Arfan Ahmed, Marie A. Renault, Rami Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: Applications, challenges, and

future outlook. *Journal of Medical Internet Research*, 26:e59505.

Ian Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, 7:1430984.

Xiyang Huang, Jiawei Lin, Keying Wu, Jiaxin Huang, Kailai Yang, Renxiong Wei, Jiayi Xiang, Ziyan Kuang, Min Peng, Qianqian Xie, and 1 others. 2026. Siming-bench: Evaluating procedural correctness from continuous interactions in clinical skill videos. *arXiv preprint arXiv:2604.09037*.

S. Barry Issenberg, William C. McGaghie, E. Ronald Petrusa, David L. Gordon, and Rosemary J. Scalese. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher*, 27(1):10–28.

Mahdieh Ebrahimi Khanghahi and Farahnaz Esmaeili Fomani Azar. 2018. Direct Observation of Procedural Skills (DOPS) evaluation method: Systematic review of evidence. *Medical Journal of the Islamic Republic of Iran*, 32(1):45.

Jessica J. Lau, Shibashis Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Hao Liu, Jian Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36.

Bo Liu, Li-Ming Zhan, Lin Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

William C. McGaghie, S. Barry Issenberg, Elaine R. Cohen, Jeffrey H. Barsuk, and Diane B. Wayne. 2011. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? a meta-analytic comparative review of the evidence. *Academic Medicine*, 86(6):706–711.

OpenAI. 2024. GPT-4 Technical Report.

Ashish R. Panchal, Jason A. Bartos, José G. Cabañas, Michael W. Donnino, Ian R. Drennan, Karen G. Hirsch, Peter J. Kudenchuk, Michael C. Kurz, Eric J. Lavonas, Peter T. Morley, Brian J. O'Neil, Mary Ann Peberdy, Jon C. Rittenberger, Alexis J. Rodriguez, Katherine N. Sawyer, and Katherine M. Berg. 2020. Part 3: Adult basic and advanced life support: 2020 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 142(16_suppl_2):S366–S468.

Lakshmi Seenivasan, Mobarakol Islam, A. K. Krishna, and Hongliang Ren. 2022. Surgical-VQA: Visual question answering in surgical scenes using transformer. In *Lecture Notes in Computer Science*, volume 13437, pages 33–43.

Julie H. Vermylen, Elaine R. Cohen, David A. Cook, William C. McGaghie, S. Barry Issenberg, Jennifer Arnold, Heather Ballard, Mohamed Bayoumi, Michael Beestrum, Ruth Bremner, Sarah Crawford, Naomi Einstein, Christina Mannarino, Anjali Misra, T. M. Tomita, Hannah Waldron, Francesca Yanko, and Deborah O. Kessler. 2025. Competency-based simulation training for procedural skills: A systematic review and meta-analysis. *Simulation in Healthcare*.

Peng Wang, Shuai Bai, Shidong Tan, Shunqiang Wang, Zhihao Fan, Jing Bai, Kai Chen, Xian Liu, Jie Wang, Wei Ge, Yu Fan, Kai Dang, Meng Du, Xin Ren, Rui Men, Dong Liu, Chen Zhou, Jingren Zhou, and Ji Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.

.1 Fine-grained Error Analysis

A qualitative inspection of the ordering behavior suggests that the main difficulty was associated with visually adjacent BLS states rather than with completely unrelated procedural phases. This interpretation is consistent with the difference between Pairwise Micro Accuracy and Task Accuracy: the model captured some local frame relationships, but local ambiguities accumulated and affected exact sequence recovery.

The most challenging transitions in BLS are those in which adjacent procedural states share similar visual cues. Early frames may show the rescuer approaching or preparing the scene without clear evidence of a specific procedural action. Similarly, chest exposure and hand positioning may overlap visually, since the chest can become visible before, during, or immediately after hand placement. The transition between hand positioning and active compressions is also difficult in static keyframes because the presence of hands on the chest does not necessarily indicate whether compression has already started. Likewise, compression and adjustment phases may share similar body posture when the rescuer pauses, corrects hand position, or resumes the maneuver.

These ambiguities help explain why the system could generate complete rationales while still fail-

ing to recover the exact sequence. The extracted visual evidence was often sufficient to describe relevant elements in the scene, but not always sufficient to discriminate fine-grained temporal transitions between adjacent BLS states. Therefore, the main limitation of the current framework lies in converting local visual evidence into a globally correct procedural order. This finding also supports the role of the PCA-based tie-breaking mechanism as a local ambiguity-resolution step for frames assigned to similar or tied procedural stages, rather than as a complete solution to temporal reasoning.

.2 Limitations of the Approach

The proposed approach has several limitations. First, the system was evaluated in a zero-shot setting, without task-specific fine-tuning, which likely constrained its ability to distinguish visually similar frames belonging to nearby procedural states. Second, the Stage 1 representation depends on rule-based extraction from short generated descriptions, so relevant cues may be missed when the model expresses the same evidence using unexpected wording. Third, the Stage 2 formulation reduces temporal reasoning to a discrete ordinal scale, thereby improving stability but potentially oversimplifying subtle transitions between adjacent clinical actions. In addition, the PCA-based tie-breaking strategy is only an approximate solution, since a one-dimensional projection does not explicitly model temporal dependencies among frames. Finally, the current experiments focused only on the BLS subset, so the generalizability of the proposed pipeline to other clinical procedures remains to be validated.

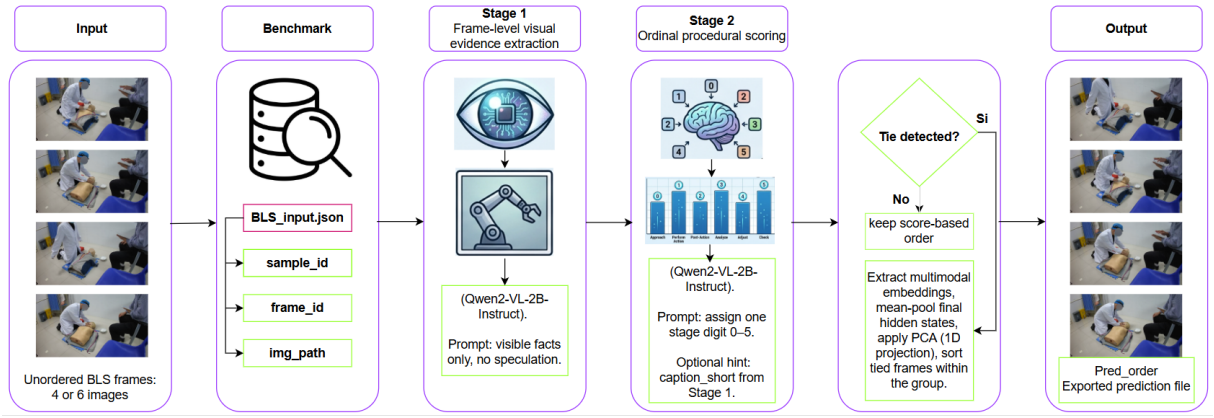

Figure 1: Outline of the proposed model.