DLNLP at ClinicalSkillQA: EvidenceFlow for Structured Zero-Shot Clinical Keyframe Ordering

Kexin Li, Zhekun Wang, Yiran Wang, Di Zhao*
Dalian Minzu University
zhaodi@dlnu.edu.cn

Abstract

The ClinSkill QA shared task requires models to recover the temporal order of scrambled clinical keyframes and generate explanations. We propose EvidenceFlow, a structured zero-shot framework based on Qwen2.5-VL that decomposes the task into global overview, local evidence modeling, and ordering decision, with two variants: model-led EvidenceFlow-M and rule-guided EvidenceFlow-R. On the official test set, EvidenceFlow-R achieves better ordering performance, while EvidenceFlow-M produces better explanation quality, revealing a trade-off between ordering stability and rationale generation. EvidenceFlow provides an interpretable zero-shot baseline for clinical keyframe ordering.

1 Introduction

Clinical skill assessment is crucial in medical education. The ClinSkill QA task requires models to reorder scrambled clinical keyframes and provide a rationale for the predicted sequence, offering a fine-grained testbed for multimodal understanding of clinical processes in realistic teaching scenarios.

Recent multimodal LLMs have advanced visual understanding and zero-shot reasoning, yet clinical keyframe ordering remains challenging for three reasons. First, the input consists of only a few scrambled static images, requiring the model to infer temporal progression from subtle local differences. Second, multiple visual cues may evolve in parallel, such as chest exposure and hand positioning, which makes global ordering more difficult. Third, the task requires readable explanations grounded in clear evidence.

To address these challenges, we propose a structured zero-shot framework based on Qwen2.5-VL (Bai et al., 2025) that decomposes clinical keyframe ordering into three stages: global overview, local evidence modeling, and ordering decision. The framework explores two ordering approaches: model-led reasoning and rule-guided correction with explicit evidence constraints. This design is motivated by a practical tension in the task: flexible reasoning may support better explanation generation, whereas stronger evidence constraints may improve ordering stability.

Our contributions are threefold:

1. We propose **EvidenceFlow**, a structured zero-shot framework for clinical keyframe ordering and rationale generation.

2. We design a three-stage pipeline with two ordering variants: model-led **EvidenceFlow-M** and rule-guided **EvidenceFlow-R**.

3. We show a trade-off between ordering stability and explanation quality, clarifying current MLLM limitations on this task.

2 Related Work

In recent years, multimodal research in medical and clinical scenarios has mainly focused on medical visual question answering, image dialogue, and multi-image understanding. Representative studies such as PMC-VQA (Zhang et al., 2024a), LLaVA-Med (Li et al., 2023), and Med-MIM (Yang et al., 2025) demonstrate the potential of multimodal models in medical image understanding, open-ended QA, and multi-image reasoning.

For structured multimodal reasoning, Multimodal-CoT (Zhang et al., 2024b) shows that making intermediate reasoning explicit can facilitate complex vision-language tasks. For temporal and multi-image understanding, MuirBench (Wang et al., 2025), Mementos (Wang et al., 2024), TempCompass (Liu et al., 2024), and TempVS (Song et al., 2025) indicate that robust cross-image or cross-frame temporal reasoning remains challenging for current MLLMs. In the

*Corresponding author.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 33–37

zero-shot setting, Socratic Models (Zeng et al., 2023) further suggest that organizing intermediate evidence through language can improve interpretability without task-specific training.

Our work differs from these studies by focusing on structured zero-shot ordering in clinical keyframe sequences, where both temporal recovery and evidence-grounded explanation are required.

3 Method

3.1 Task Definition and Overall Framework

Given a set of clinical keyframes $X = \{x_1, x_2, \ldots, x_n\}$ (typically $n \in [4, 6]$), the model must output the correct temporal order $Y = [y_1, y_2, \ldots, y_n]$ and a natural language explanation. Unlike classification, the core task is to recover the full procedural order from subtle cross-image differences.

We propose EvidenceFlow, a structured zero-shot framework that decomposes clinical keyframe ordering into three stages—global overview, local evidence modeling, and ordering decision—with two variants: model-led EvidenceFlow-M and rule-guided EvidenceFlow-R, as shown in Figure 1.

3.2 Global Overview and Local Evidence Modeling

In the global overview stage, all keyframes in the same sample are first concatenated into a grid image according to their labels, and then input into the multimodal model for holistic analysis. The model needs to extract main change cues, early or late candidate anchors, and corresponding uncertainty information at the group level, denoted as $G = \{a, s^{early}, s^{late}, u\}$, where a denotes the dominant axis of change; s^{early} and s^{late} denote the confidence that an image serves as an early or late anchor, respectively; and u denotes global uncertainty information. Subsequently, the axis of change with the highest confidence is extracted as the main reference axis for subsequent sorting, and candidate anchors are extracted for subsequent ordering decisions.

In the local evidence stage, each image is independently analyzed along six dimensions (chest exposure, hand positioning, hand stability, airway, ventilation, AED), producing a vector $e_i = [chest_i, hand_i, stab_i, airway_i, vent_i, aed_i, c_i, z_i]$, where the first six encode local states, c_i denotes confidence, and z_i denotes a coarse stage label. This stage complements the global overview by providing fine-grained cues for ordering.

3.3 EvidenceFlow Ordering Decision

Given the global overview and local evidence, the framework proceeds to the ordering decision stage. This stage requires further integrating the overall change trends at the image-group level with the fine-grained clinical cues at the single-image level to generate a consistent global order.

Under this framework, we instantiate two ordering variants: EvidenceFlow-M emphasizes model-led reasoning, whereas EvidenceFlow-R relies more on explicit evidence constraints and rule-based correction. By comparing these two variants, we analyze how different control strategies affect ordering stability and explanation generation.

3.3.1 EvidenceFlow-M: Model-Led Ordering Implementation

This stage first determines whether stable start/end anchors exist to select an ordering mode. To do so, it computes a pairwise relation score for each image from pairwise comparisons:

$$R_{\text{pair}}(x) = \frac{1}{|X| - 1} \sum_{y \neq x} c_{x,y} \delta_{x,y} \tag{1}$$

The resulting score characterizes the relative earliness or lateness tendency of each image within the set, where $c_{x,y}$ denotes the pairwise comparison confidence between images x and y, and $\delta_{x,y}$ is defined as follows:

$$\delta_{x,y} = \begin{cases} 1, & \text{x is later than y} \\ -1, & \text{x is earlier than y} \\ 0, & \text{UNCERTAIN} \end{cases} \tag{2}$$

Based on the candidate anchor information, main change cues, and coarse-grained stage labels from the global overview, we compute an anchor score for each image:

$$S_{\text{early}}(x) = A_{\text{early}}(x) - P_{\text{axis}}(x) + B_{\text{early}}(x) \tag{3}$$

$$S_{\text{late}}(x) = A_{\text{late}}(x) + P_{\text{axis}}(x) + B_{\text{late}}(x) \tag{4}$$

Here, $A_{\text{early}}(x)$ and $A_{\text{late}}(x)$ denote the early and late candidate confidences given in the global overview stage, respectively; $P_{\text{axis}}(x)$ denotes the relative progression value of the main change cues for the image; and $B_{\text{early}}(x)$ and $B_{\text{late}}(x)$ respectively denote the stage reward terms derived from the coarse-grained stage labels.

Figure 1: Overview of the EvidenceFlow framework for clinical keyframe ordering.

Based on anchor scores and marginal differences, the model adaptively selects a mode $m \in \{\text{dual}, \text{single}, \text{none}\}$. In dual mode, early and late anchors fix the start and end, with intermediate images sorted by axis progression. In single mode, one anchor fixes one end, and the remaining images are sorted according to their progression estimates. In none mode, the model estimates relative progression directly. After obtaining the initial sequence $Y^{(0)} = [y_1, y_2, \ldots, y_n]$, local corrections using pairwise scores and cached comparisons, followed by lightweight global verification, yield the final sequence Y and its explanation.

3.3.2 EvidenceFlow-R: Rule-Guided Ordering Implementation

Unlike EvidenceFlow-M, EvidenceFlow-R generates ordering results through stage consistency correction, group-level position priors, continuous progression scoring, and directed pairwise comparisons. Specifically, we first refine the coarse-grained stage label z_i obtained from the local evidence stage by combining the local evidence e_i and the global overview information G, yielding a corrected stage value $\tilde{z}_i$. On this basis, the corrected stage information, the aggregated local evidence scores, and the group-level position priors are jointly mapped into a rule-flow ordering score:

$$S_R(x_i) = \alpha \tilde{z}_i + \beta E_i + \gamma P_i \tag{5}$$

Here, E_i denotes the aggregated local evidence score, and P_i denotes the group-level position prior. The coefficients α, β, and γ denote the corresponding weights. By sorting $S_R(x_i)$ in ascending order, an initial sequence $Y^{(0)} = [y_1, y_2, \ldots, y_n]$ can be obtained. For image pairs with close ordering scores, the framework further performs directed pairwise comparisons for local refinement. If necessary, lightweight global verification is then applied to further correct adjacent errors and produce the final sequence Y together with its natural language explanation.

4 Experimental Setup

Experiments are conducted on the ClinSkill QA dataset (200 samples, each containing 4–6 keyframes). The base model Qwen2.5-VL-7B-Instruct is used in a zero-shot setting with 4-bit quantization and a temperature of 0 on a single RTX4090. All experiments are conducted without task-specific training. Both variants share the same backbone and evidence extraction setting, differing only in the final ordering strategy. Following the official evaluation protocol, we report the overall Score, Task Acc (exact sequence match), Pair Micro (pairwise order accuracy), and BERT F1 (Zhang et al., 2020), which measures semantic similarity between the generated explanation and the reference.

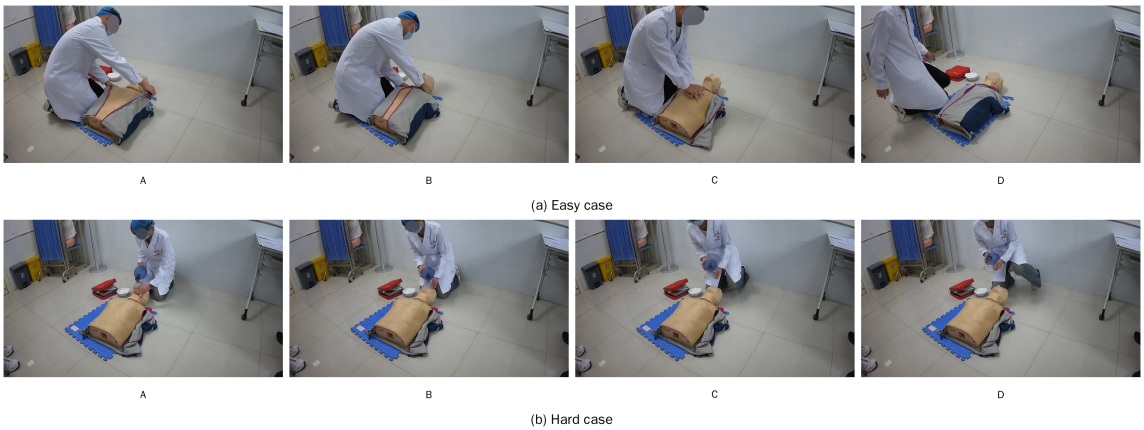

Figure 2: Case study of easy and hard scenarios in CPR emergency response.

5 Experimental Results

5.1 Main Experimental Results

We compare EvidenceFlow-R and EvidenceFlow-M under the same base model and official evaluation protocol. The results are shown in Table 1.

Variant	Score	Task Acc	Pair Micro	BERT F1
EvidenceFlow-R	30.44	0.10	0.55	0.57
EvidenceFlow-M	28.71	0.03	0.51	0.74

Table 1: Main results on the ClinSkill QA test set.

As shown in Table 1, EvidenceFlow-R outperforms EvidenceFlow-M on Score, Task Acc, and Pair Micro, indicating more stable sequence recovery, while EvidenceFlow-M achieves higher BERT F1, reflecting better semantic similarity in explanation generation.

5.2 Comparison with Public Submission Results

Table 2 compares our method with several public submissions from the same test phase. Compared

Method	Score	Task Acc	Pair Micro	BERT F1
baovy	71.43	0.63	0.86	0.79
qppprun	56.73	0.47	0.79	0.55
VerbaNexAI Lab	37.96	0.17	0.60	0.71
EvidenceFlow-R	30.44	0.10	0.55	0.57
EvidenceFlow-M	28.71	0.03	0.51	0.74

Table 2: Comparison of public submissions from the same test phase.

with public submissions, our frame- work appears more competitive in explanation qual- ity than in ordering accuracy. This suggests that the model can extract useful local cues, but still struggles to organize them into a consistent global temporal structure. Together with the internal com- parison in Table 1, these results indicate that stable cross-image cue integration remains the main bot- tleneck of the task.

5.3 Error Analysis

Figure 2 shows that (a) is an easy case correctly predicted by Qwen2.5-VL-7B-Instruct; (b) is a hard case with ground truth C→A→B→D, but the model predicts C→B→A→D, swapping frames A and B. This error arises from high visual similarity during bag-valve mask ventilation, where subtle hand and mask changes are hard to capture, revealing insufficient fine-grained temporal discrimination as the main bottleneck.

6 Conclusion

We propose EvidenceFlow, a structured zero-shot framework for clinical keyframe ordering. The results show that the rule-guided variant yields better ordering performance, whereas the model-led variant produces better explanations, revealing a trade-off between ordering stability and rationale quality. This work provides an interpretable zero-shot baseline for multi-image temporal reasoning in clinical scenarios.

Limitations

The 7B model is limited in fine-grained cross-image temporal integration, especially on hard samples. Our analysis lacks ablations and large-scale verification, and the framework depends on prompt design and handcrafted constraints.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei 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, pages 28541–28564. Curran Associates, Inc. Datasets and Benchmarks Track.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand. Association for Computational Linguistics.

Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. 2025. Burn after reading: Do multimodal large language models truly capture order of events in image sequences? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24316–24342, Vienna, Austria. Association for Computational Linguistics.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442, Bangkok, Thailand. Association for Computational Linguistics.

Xikai Yang, Juzheng Miao, Yuchen Yuan, Jiaze Wang, Qi Dou, Jinpeng Li, and Pheng-Ann Heng. 2025. Medical large vision language models with multi-image visual ability. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964 of *Lecture Notes in Computer Science*, pages 402–412. Springer Nature Switzerland.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024a. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4:277.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alexander Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.