zzunlp at ClinSkill QA: Perceive-and-Plan with Decomposed In-Context Learning and Saliency-Guided Perception for Clinical Skill Keyframe Reordering

Bin Huang[†] **Yi Luo**[†] **Zhongtian Hua**[†] **Guanghui Zhao**[†]
Kaixuan Yuan[†] Kunli Zhang[†*]
[†]School of Information Engineering, Zhengzhou University
[*]Corresponding author
1084893712@qq.com, nancetide@outlook.com, hzt1113@gs.zzu.edu.cn
2285986836@qq.com, 2805476399@qq.com, ieklzhang@zzu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) show strong medical visual understanding. Their capability for *continuous perception* in procedural clinical workflows remains underexplored. We present **Perceive-and-Plan**, a decomposed in-context learning paradigm for clinical skill keyframe reordering. The method separates visual perception from temporal planning: (1) **structured visual perception** with saliency-guided Picture-in-Picture (PiP) composition that magnifies critical regions (head, chest) as color-coded insets, and (2) **temporal reasoning** with chain-style self-verification via fresh conversation reset and visual-evidence anchoring under BLS Rules R1–R11. With frozen backbone weights, our system achieves **71.43** overall (2nd, ClinSkill QA 2026), 0.86 pairwise accuracy, and 1.0 rationale coverage. Structured prompting with saliency-guided inputs improves procedural understanding in MLLMs. Code is available at https://github.com/NanceTide/clinskillqa-perceive-and-plan.

1 Introduction

Multimodal Large Language Models (MLLMs) perform well on visual understanding (Liu et al., 2023; OpenAI et al., 2023), yet applying them to *procedural clinical workflows*, where temporal causality and state transitions matter, remains difficult. Clinical skill assessment requires **continuous perception**: interpreting keyframe sequences, tracking subtle state changes (e.g., clothing exposure, automated external defibrillator pad adhesion), and reconstructing procedural timelines from visual evidence.

The **ClinSkill QA 2026** benchmark formalizes this as a keyframe reordering task: given 4–6 shuffled BLS frames, models must reconstruct the chronological sequence and generate clinically grounded explanations with **visual-evidence anchoring**.

Current *end-to-end prompting* approaches suffer from two limitations: (1) **perceptual ambiguity** (fine-grained details are lost in wide-angle views), and (2) **reasoning interference** (when description and ordering are conflated, models exhibit *temporal hallucination* and favor textbook templates over visual evidence).

We propose **Perceive-and-Plan**, a decomposed prompting paradigm: (1) **perception** (visual analysis with saliency-guided Picture-in-Picture composition), and (2) **planning** (temporal reconstruction in a fresh conversation with chain-style verification). Both stages rely on in-context learning (ICL) with frozen parameters.

Contributions: (1) **PiP composition** that magnifies head and chest regions as color-coded insets, (2) **decomposed ICL** that separates description from ordering via a Visual Anchors protocol, and (3) strong official results on ClinSkill QA 2026: **71.43** overall (2nd), 0.86 pairwise accuracy, and 1.0 rationale coverage.

2 Related Work

Prior work studies temporal multimodal reasoning on generic sequence QA and medical multi-image tasks (Wang et al., 2024; Yu et al., 2025; Huang et al., 2026), with surgical phase recognition (Twinanda et al., 2017) and laparoscopic skill assessment (Liao et al., 2025) as related clinical precedents. Clinical keyframe ordering remains comparatively underexplored.

In-context learning (ICL) supports multimodal adaptation without gradient updates (Dong et al., 2023; Baldassini et al., 2024), yet much of it targets discrete classification or VQA rather than constrained sequence generation. We focus on *clinical procedural ordering* with evidence-grounded rationales, using stage-wise decomposition and saliency-guided perception. Further benchmark-level discussion appears in Appendix A.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 24–32

3 Methodology

3.1 Overview

Our **Perceive-and-Plan** paradigm decomposes clinical keyframe reordering into two functionally isolated stages. Formally, we define a mapping

$$\mathcal{F} : \mathcal{V} \to \mathcal{S} \tag{1}$$

where $\mathcal{V} = \{I_1, \ldots, I_N\}$ denotes the unordered keyframe sequence and $\mathcal{S}$ denotes the chronologically sorted permutation. As illustrated in Figure 1, the framework factorizes this mapping as:

$$\mathcal{S} = \Psi\big(\mathcal{V}, \Phi(\mathcal{V}')\big) \tag{2}$$

where Φ denotes the perception module, Ψ denotes the reasoning module, and $\mathcal{V}'$ represents the set of saliency-enhanced frames.

1. **Stage I: Structured Visual Perception (Φ)**: Extracts a 4-dimensional clinical state vector $\mathbf{d}_i$ from each keyframe I_i, covering:
   - Operator posture
   - Hand movement dynamics
   - Equipment status
   - Patient surface/exposure state

   Mitigation of Perceptual Ambiguity: We apply a saliency enhancement function $\mathcal{E}_{\text{PiP}}$ defined as:

   $$\begin{aligned} I_i' &= \mathcal{E}_{\text{PiP}}(I_i) \\ &= \text{Composite}\big(I_i,\ \mathcal{C}_{\text{head}}(I_i),\ \mathcal{C}_{\text{chest}}(I_i)\big) \end{aligned} \tag{3}$$

   where $\mathcal{C}_{\text{head}}$ and $\mathcal{C}_{\text{chest}}$ denote magnified, color-coded insets of head-neck (red border) and chest-hands (blue border) regions.

2. **Stage II: Temporal Planning with Self-Verification (Ψ)**: Operates under **Conversation Isolation**:

   $$\text{Context}(\Psi) \cap \text{Context}(\Phi) = \emptyset \tag{4}$$

   The reasoning incorporates original frames $\mathcal{V}$, pre-computed descriptors $\{\mathbf{d}_i\}$, constraint rules $\mathcal{R}$, and a Visual Anchors Protocol $\mathcal{A}$ for evidentiary grounding.

Both stages call Qwen3.6-Plus under **ICL**, leveraging long-context multi-image reasoning. Parameters θ are frozen ($\Delta\theta = 0$), and behavior is shaped by prompts and input construction alone.

3.2 Stage I: Structured Visual Perception with PiP Enhancement

Stage I addresses missed fine-grained cues by increasing the effective resolution of salient clinical regions (ROIs). The input construction function $\mathcal{E}_{\text{PiP}}$ is:

$$\begin{aligned} I_i' = \text{Grid}\big(\ & I_i \oplus \\ & \text{Resize}(\text{Crop}(I_i, \text{bbox}_{\text{head}})) \oplus \\ & \text{Resize}(\text{Crop}(I_i, \text{bbox}_{\text{chest}}))\big) \end{aligned} \tag{5}$$

where $\oplus$ denotes spatial composition with color-coded borders. The perception module Φ projects this enhanced input into a structured semantic tuple:

$$\begin{aligned} \mathbf{d}_i = \Phi(I_i') = \langle &\text{Posture}, \text{Hands}, \\ &\text{Equipment}, \text{Surface}\rangle_i \end{aligned} \tag{6}$$

The prompt conditions Φ on BLS rule library $\mathcal{R}$ as procedural priors, enforcing concise representation for cross-frame comparability. Implementation details appear in Appendix C.3.

3.3 Stage II: Temporal Planning with Self-Verification

3.3.1 Fresh Conversation Reset

Conversation Isolation Protocol: Stage II initializes with context window $\mathcal{C}_\Psi$ strictly independent of Stage I:

$$\begin{aligned} \mathcal{C}_\Psi = \{&\text{System Prompt}, \mathcal{V}, \{\mathbf{d}_i\}, \mathcal{R}\} \\ &\text{s.t. } \nexists\ \text{turn } t \in \mathcal{C}_\Phi \end{aligned} \tag{7}$$

This isolation ensures that the ordering process Ψ is not biased by linguistic priors or confidence calibration of description stage Φ.

The planning module Ψ consumes original frames $\mathcal{V}$, pre-computed descriptors $\{\mathbf{d}_i\}$, and clinical constraints $\mathcal{R}$ to execute a chain-of-verification protocol:

1. **Anchoring**: Enumerate visual evidence $\mathcal{E}_v$ mapping each state transition to a spatial location in I_i
2. **Rationalization**: Generate clinical narrative $\mathcal{N}$ explaining temporal adjacency matrix $\mathbf{A}$
3. **Serialization**: Output machine-parseable ordered sequence $\mathcal{S}$

Temperature is reduced in this stage ($\tau = 0.2$) to minimize format drift, contrasting with higher variability in Stage I ($\tau = 0.6$).

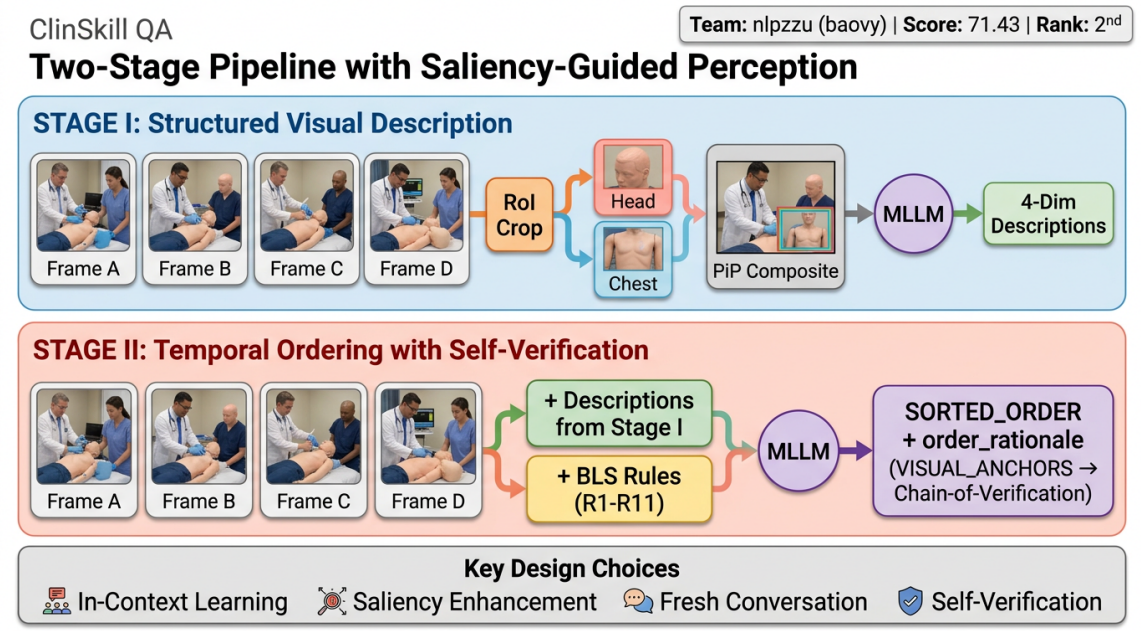

Figure 1: Two-stage pipeline with saliency-guided perception. **Stage I** extracts structured four-dimensional descriptions from ROI-enhanced PiP composites. **Stage II** orders frames with self-verification in a fresh conversation to limit carry-over bias from Stage I. Design elements include ICL with frozen weights, head/chest saliency zoom, a stage-wise conversation reset, and chain-of-verification with visual anchors.

3.4 Implementation Details

We use the Qwen3.6-Plus API with $\tau_1 = 0.6$ (Stage I) and $\tau_2 = 0.2$ (Stage II), without fine-tuning, LoRA, or PEFT. Improvements come from prompts and PiP-style input augmentation ($\mathcal{E}_{\text{PiP}}$). Full prompts, retry logic, and parsing schemas appear in Appendices C–C.5.

4 Experiments

4.1 Dataset and Evaluation

ClinSkill QA 2026 comprises 200 sets of shuffled keyframes from Basic Life Support (BLS) training videos at Zhongnan Hospital, Wuhan University. Each set has 4–6 frames from a continuous procedure (CPR, automated external defibrillator deployment, airway management). Ground truth is expert-annotated chronological order and clinical rationales.

Evaluation metrics:

- **Task Accuracy**: Exact match of predicted order to ground truth
- **Pairwise Accuracy**: Fraction of adjacent frame pairs correctly ordered
- **BERTScore F1**: Similarity between generated and reference rationales
- **Prediction Coverage / Rationale Coverage**: Completeness metrics

4.2 Main Results

Table 1 lists detailed metrics on the official CodaBench test set.

Metric	Value	Interpretation
Overall Score	71.43	Primary metric
Task Accuracy	0.63	126/200 exactly correct
Pairwise Micro	0.86	Strong local ordering
BERT Precision	0.79	Semantic precision
BERT Recall	0.79	Semantic recall
BERT F1	0.79	Balanced quality
Prediction Coverage	0.99	198/200 valid predictions
Rationale Coverage	1.0	200/200 with rationale

Table 1: Performance metrics on ClinSkill QA 2026 test set (Team: baovy/zzunlp).

Strengths: Pairwise accuracy 0.86 indicates strong relative ordering even when exact permutations fail. Rationale coverage 1.0 shows that the chain-of-verification protocol reliably yields structured explanations.

4.3 Result Analysis

The gap between task accuracy (0.63) and pairwise accuracy (0.86) reflects correct local relations with residual long-range errors. BERTScore F1 (0.79) indicates solid semantic alignment with references.

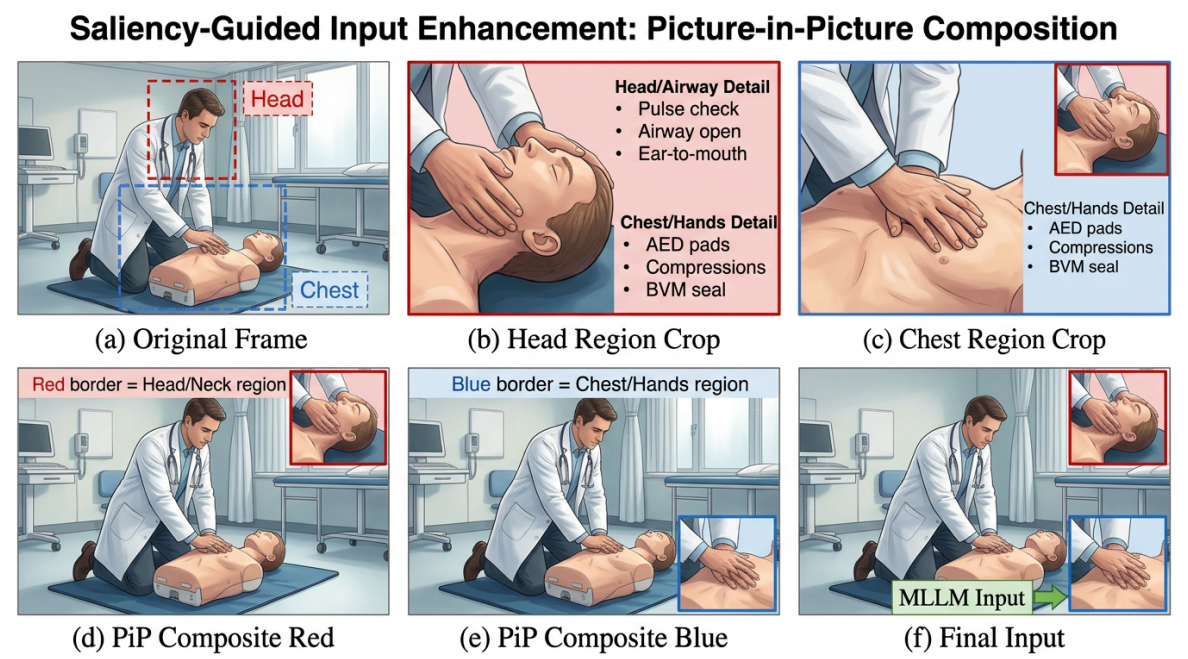

Figure 2: Saliency-guided input enhancement via PiP composition. (a) Original frame with indicated Head (red) and Chest (blue) regions. (b-c) Zoomed crops for head/airway and chest/hands details. (d-e) Individual PiP composites with color-coded borders. (f) Final input to MLLM combining full scene with both magnified insets.

Additional error patterns and qualitative cases appear in Appendix B.

5 Conclusion

We present **Perceive-and-Plan**, a decomposed ICL approach that isolates saliency-guided PiP perception from chain-style temporal planning with self-verification. On ClinSkill QA 2026 it ranks second overall (**71.43**) with frozen model weights. Together, saliency-guided inputs, lightweight domain rules, and explicit verification steps provide a practical approach to procedural reasoning with MLLMs, suited for data-limited and privacy-sensitive educational settings.

Limitations: The system depends on proprietary API access and rate limits. PiP crops are hand-designed for typical BLS manikin layouts and may require retuning for other skills or camera viewpoints. We do not study ensembling or test-time adaptation.

Outlook: We plan to extend the framework to full-video temporal assessment, to explore complementary audio cues (e.g., ventilation sounds, instructor prompts), and to study learned or attention-based region proposals so that saliency cues can transfer beyond fixed BLS-style crops.

Ethics Statement

This study uses the official ClinSkill QA 2026 benchmark, collected under Institutional Review Board (IRB) approval at Zhongnan Hospital, Wuhan University. All footage depicts simulated scenarios with manikins. No real patient data are included. The system is intended for educational assessment only and must not be used for clinical diagnosis or patient care without rigorous validation and regulatory clearance.

Acknowledgments

This work was supported by the Natural Science Foundation of Henan Province (Grant No. 252300421877). We thank the ClinSkill QA organizers for the benchmark and evaluation platform, and the ModelScope team for API access to the Qwen model family.

References

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1539–1550.

Mohamad Ballout, Okajevo Wilfred, Seyedalireza Yaghoubi, Nohayr Muhammad Abdelmoneim, Julius

Mayer, and Elia Bruni. 2025. Can you SPLICE it together? A human curated benchmark for probing visual reasoning in VLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11288–11309, Suzhou, China. Association for Computational Linguistics.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics. ArXiv:2402.18060.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Sivan Doveh, Shaked Perek, M. Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. 2025. Towards multimodal in-context learning for vision & language models. In *Computer Vision – ECCV 2024 Workshops*, volume 15641 of *Lecture Notes in Computer Science*, pages 250–267. Springer.

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

Chuanhao Li, Chenchen Jing, Zhen Li, Mingliang Zhai, Yuwei Wu, and Yunde Jia. 2024. In-context compositional generalization for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17954–17966, Miami, Florida, USA. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2306.00890*.

Wenqiang Liao, Ying Zhu, Hanwei Zhang, Dan Wang, Lijun Zhang, Tianxiang Chen, Ru Zhou, and Zi Ye. 2025. Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: A systematic review. *Frontiers in Surgery*, 12.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kavya Dasaramoole Prakash, Kiseong Kim, and Youngmahn Han. 2025. Enhancing clinical reasoning in medical vision-language model through structured prompts. *medRxiv*.

Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. 2017. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97.

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.

Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, Yannan Yu, and Yuyin Zhou. 2025. MedframeQA: A multi-image medical VQA benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.

Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. 2025. Guiding medical vision-language models with diverse visual prompts: Framework design and comprehensive exploration of prompt variations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11726–11739, Albuquerque, New Mexico. Association for Computational Linguistics.

A Extended Related Work

A.1 Sequence Reasoning Benchmarks

Recent benchmarks (Mementos (Wang et al., 2024), MedFrameQA (Yu et al., 2025), SiMing-Bench (Huang et al., 2026), SPLICE (Ballout et al., 2025)) show that temporal consistency remains difficult for MLLMs across general, medical, and instructional domains. ClinSkill QA additionally requires clinically grounded ordering and rationale generation in one setting.

A.2 Medical and ICL Context

Prior medical vision-language systems often rely on supervised adaptation (Li et al., 2023; Chen et al., 2025) or structured visual prompting (Zhu et al., 2025; Prakash et al., 2025). We instead use decomposed constrained prompting under ICL (Dong et al., 2023; Baldassini et al., 2024; Doveh et al., 2025; Li et al., 2024) to obtain robust structured outputs with frozen backbone weights.

B Extended Error Analysis

We observe three dominant failure modes: (1) **long-range permutation slips**, where most local relations are correct but a single middle transition is misplaced, (2) **partial visibility confusion**, especially for automated external defibrillator pad adhesion and bag-valve-mask sealing under occlusion, and (3) **state granularity mismatch**, where frames with similar global posture but subtle hand-state differences are swapped.

C In-Context Learning Implementation Details

C.1 Full Prompt Templates

C.1.1 Stage I: Structured Visual Description Prompt

Role: Emergency medical training examiner. **Clinical Rules:** Assessment before intervention. Chest exposure progresses from covered to partial to fully exposed. AED phases: package, peel pads, apply to chest, connect to device. BVM sequence: prepare equipment, seal mask, squeeze bag. CPR mandates pulse check before compressions. **BLS Rules (R1–R11):** R1: hand shape and contact site. R2: object and clothing state. R3: space and posture. R4: exposure direction. R5: guideline variants. R6: dependencies. R7: priorities. R8: all people in scene. R9: metadata. R10: visual evidence over theory. R11: continuity constraints. **Output:** Four-dimensional description per image: posture, hands, equipment, surface. Cap at 20 words per dimension and 120 words total.

C.1.2 Stage II: Temporal Ordering Prompt

Role: BLS expert, **Input:** Shuffled frames {A, B, C, D} + descriptions from Stage I, **State Checklist:** (1) EXPOSURE: less-exposed precedes more-exposed, (2) AED PADS: adhered count never decreases, (3) AIRWAY: maneuver precedes sealed BVM, **Chain-of-Verification:** Visual Anchors first (Figure A: [irreversible marker]), then narrative (overview + per-figure paragraphs in chronological order), **Output:** SORTED_ORDER: [“A”, “B”, ...]

C.2 API Configuration and Retry Logic

- **Model:** Qwen3.6-Plus, accessed via ModelScope API.
- **Stage I:** temperature 0.6, max output 4096 tokens.
- **Stage II:** temperature 0.2, max output 2048 tokens.
- **Timeout:** 300 seconds per request.
- **Retries:** up to 5 attempts with exponential backoff (15, 30, 45, 60, 75 seconds).
- **API Keys:** 20 keys used for rate-limit rotation.

C.3 Picture-in-Picture (PiP) Implementation

ROI Crop Ratios. Head region: (0.05, 0.05, 0.45, 0.55). Chest region: (0.25, 0.40, 0.75, 0.85). **Inset Constraints.** Width: 0.30 of frame width. Height: 0.38 of frame height. **Colors and Borders.** Head border: red (255, 60, 60). Chest border: blue (40, 120, 255). Border thickness: 3 px. **Pipeline.** (1) Load RGB image. (2) Crop head and chest regions. (3) Resize crops with LANCZOS interpolation. (4) Draw color-coded borders. (5) Paste head inset at top-right and chest inset at bottom-right of the original frame. (6) Save composite as JPEG (quality 95).

C.4 Hyperparameter Tuning Log

Hyperparameter sweep on the development set:

Stage I Temperature. Swept 0.0 (too rigid), 0.3 (misses details), 0.6 (selected), 0.9 (too verbose), 1.0 (unreliable). We chose 0.6 as the best balance of detail and conciseness. **Stage II Temperature.** Swept 0.0 (rigid format), 0.2 (selected), 0.5 (occasional errors), 0.7 (inconsistent). We selected 0.2 for reliable format adherence. **Max Tokens.** Stage I: 4096 tokens for four-dimensional descriptions. Stage II: 2048 tokens for rationale and ordering. **Retry Strategy.** Exponential backoff at 15, 30, 45, 60, and 75 seconds, up to 5 retries. Success rate across the full test set was 99.2%.

C.5 Prompt Engineering Iterations

We refined prompts over more than 15 iterations across three weeks:

- **Iterations 1–3 (Structure).** Initial prompting produced irrelevant details. We introduced the four-dimensional description format to focus on clinically relevant content.
- **Iterations 4–6 (Rules).** Injected BLS rules, but the model ignored them. Added explicit "STRICT" markers to enforce compliance.
- **Iterations 7–9 (Format).** Output contained frequent JSON parsing errors. Switched to a SORTED_ORDER line for reliable extraction.
- **Iterations 10–12 (Evidence).** Model engaged in post-hoc rationalization. Introduced Visual Anchors to ground reasoning in observable frame content.
- **Iterations 13–15 (Brevity).** Descriptions were either too sparse or too verbose. Added explicit 20-word-per-dimension caps.

D Sample Predictions and Analysis

D.1 Successful Case: CPR Sequence

Input: [C, A, D, B] (shuffled)
Ground Truth: [D, B, A, C]
Prediction: [D, B, A, C] (Correct)
Rationale: This CPR sequence shows the progression from initial approach to chest compression. Figure D shows the operator approaching the manikin with clothing fully fastened: consistent with the earliest pre-compression stage. Figure B shows partial clothing exposure via unzipping: must follow Figure D and precede full exposure. Figure A shows fully opened clothing with complete chest exposure: immediately before compressions. Figure C shows hands stacked on sternum performing compressions: confirming the final position.
Analysis: The model tracked the monotonic clothing exposure sequence and identified the irreversible compression initiation step.

D.2 Challenging Case: Ambiguous AED Sequence

Input: [A, B, C, D]
Ground Truth: [B, A, D, C]
Prediction: [B, A, C, D] (Incorrect)
Error: Frame C showed one pad partially adhered with unclear visibility of the second pad. The model treated partial adhesion as full completion and placed C before D, illustrating sensitivity to occlusion and ambiguous pad states.

E System Architecture Pseudocode

The PerceiveAndPlan function accepts a sample identifier and a dictionary of image paths. It executes two stages in sequence, each calling the MLLM API with frozen parameters.
Stage I: Structured Visual Perception. For each image, if the saliency-PiP mode is active, the function builds a composite frame with cropped and color-bordered head and chest insets. Each composite is base64-encoded and sent to the MLLM with the Stage I prompt (temperature 0.6). The response is parsed to extract a four-dimensional description (posture, hands, equipment, surface).

Stage II: Temporal Planning. In a fresh conversation, the original frames are base64-encoded and sent alongside the pre-computed Stage I descriptions and the Stage II prompt (temperature 0.2). The response is parsed to extract the sorted ordering and the clinical rationale.
Return Value. The function returns a dictionary mapping the sample ID to the predicted frame order and the corresponding rationale text.

F Detailed BLS Rules Explanation

The BLS Rules Library (R1–R11) encodes emergency medicine practice guidelines, refined through iterative error analysis:

- **R1: Hand and contact disambiguation.** Identify hand shape and contact site. Neck contact indicates pulse check. Forehead or chin contact indicates airway management.
- **R2: Progressive state tracking.** Clothing follows a monotonic sequence: zipper closed, then half open, then fully open. Chest coverage: full, then partial, then exposed. BVM: hovering, then sealed on face, then squeezing. AED pads: 0, then 1, then 2 adhered pads, and the count never decreases.
- **R3: Spatial and posture cues.** Standing back indicates early-stage observation. Kneeling or bent-over posture indicates active intervention. Mask floating near the face suggests preparation, while mask sealed suggests execution.
- **R4: Clinical direction.** Exposure precedes recovery. End-of-scene indicators include examiner writing notes and supplies being stowed.
- **R5: Guideline variants.** Both C-A-B and A-B-C sequences may appear. Assessment precedes intervention unless the scenario involves drowning.
- **R6: Action dependencies.** Airway maneuvers precede rescue breaths. Peeling backing precedes pad adhesion. Pads must be placed before rhythm analysis and shock delivery.
- **R7: Priority rules.** Patient assessment takes priority over hygiene steps and equipment assembly.
- **R8: Multi-person analysis.** Examiner crouching and writing indicates a late-stage frame. Assistant standing back indicates an early-stage frame.
- **R9: Metadata as soft prior.** Sample IDs provide weak ordering hints but must never override direct visual evidence.
- **R10: Evidence over theory.** Prefer the visible timeline in the frames over textbook procedure defaults when they conflict.
- **R11: Continuity constraints.** Adhered pads never decrease in count. Torn packages remain open. Exposed chest stays exposed until the recovery phase.

G Leaderboard Submission Details

Team **baovy** (zzunlp) submitted on 2026-04-07 at 09:07 (submission ID 667353) to the CodaBench platform (https://www.codabench.org/competitions/14884/).

G.1 Detailed Performance Metrics

Submission ID 667353, submitted on 2026-04-07 at 09:07 to the CodaBench platform.
Detailed metrics:

- Overall Score: 71.43 (primary metric).
- Task Accuracy: 0.63 (126 out of 200 samples exactly correct).
- Pairwise Micro: 0.86 (strong local ordering).
- BERTScore Precision, Recall, F1: 0.79 (balanced semantic quality).
- Prediction Coverage: 0.99 (198 of 200 predictions valid).
- Rationale Coverage: 1.0 (all 200 samples include an explanation).

H Computational Resources

- Model: Qwen3.6-Plus (via ModelScope API)
- API Calls: 400 requests (200 samples × 2 stages)
- API Keys: 20 keys for rate limit rotation
- Average Latency: 15–45 seconds per request
- Total Processing Time: 3 hours for full test set
- Development Experiments: 50 hours (hyperparameter tuning)
- Prompt Iterations: 15+ versions over 3 weeks
- Cost: API credits via ModelScope platform