# LEARNING BIDIRECTIONAL NORMATIVE LATENT DYNAMICS WITH DISCREPANCY REASONING FOR CLINICAL SKILL ASSESSMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision-Language Models (VLMs) have become increasingly capable of recognizing events in videos. However, reliable procedural reasoning requires more than identifying what happened: a model must track how interactions update the underlying world state and verify whether the resulting state evolution is valid. In this work, we study this problem through clinical skill assessment, where the correctness of an action depends on the procedural state accumulated through prior interactions among the operator, patient, instruments, and environment. We propose ATV, an Anticipate-Then-Verify framework that formulates procedural assessment as verification against normative latent dynamics. ATV first organizes video observations into temporally aligned object-centric states. It then learns bidirectional dynamics exclusively from correctly performed procedures, where a forward predictor anticipates how object states should evolve and a reverse-time reflector captures whether the resulting states are compatible with their expected antecedents. At inference time, ATV converts the discrepancies between observed and normative state transitions into rubric-conditioned evidence for procedural scoring. We further construct ClinSkill-V, a large-scale dataset of 15,775 step-aligned clinical skill video clips collected from real-world clinical skill examinations, with expert rubric-based ordinal annotations. Extensive experiments against open-source VLMs and latent predictive models demonstrate that ATV consistently improves procedural assessment, with particularly pronounced gains in fine-grained ordinal scoring.

## 1 INTRODUCTION

How can Vision-Language Models (VLMs) track interaction-driven world evolution in videos? As VLMs are increasingly viewed as world models (He et al., 2025; Ge et al., 2024), a key challenge along this direction is to model how visual interactions change the underlying state of the scene (Xue et al., 2024), rather than merely encode events along a temporal axis. Unlike conventional video understanding, which primarily interprets what has happened, state transition tracking requires a model to infer how each action updates the current world state (Zameni et al., 2025) and how the updated state determines the validity of subsequent actions. Clinical skill assessment (Martin et al., 1997; Whalen et al., 2022) makes this challenge particularly evident. Specifically, the correctness of a trainee's action depends not only on what is visually observed at the current moment, but also on the procedural state accumulated through prior interactions with instruments, the patient, and the surrounding environment (Malau-Aduli et al., 2021). Expert examiners inherently perform such interaction-grounded state tracking when assessing clinical skill performance. Given the scale of medical training and licensure, automating even a portion of this expert judgment could greatly alleviate the demand for manual assessment and make fine-grained, expert-level feedback more scalable for clinical education (Johnsson et al., 2024).

Existing video understanding models (Zhang et al., 2024; Li et al., 2025a; Zhang et al., 2025) remain confined to the recognition and prediction of interaction-relevant visual manifestations, and thus overlook a key requirement of procedural assessment, i.e., *normative state verification*. As shown in Fig 1 (c) and (d), these models recognize which interactions occur without modeling the underlying state changes they induce, and extrapolate future observations without determining whether those



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

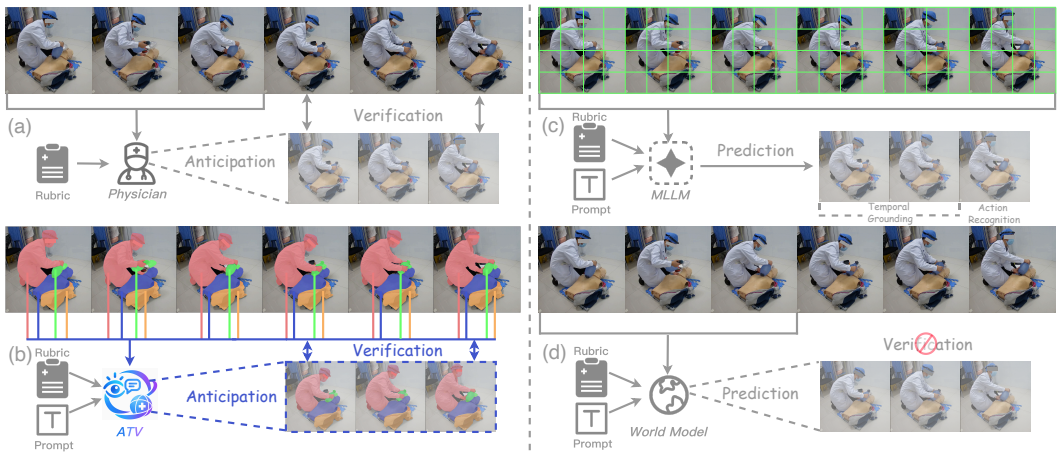



Figure 1: From prediction to verification. (a) Physicians anticipate expected procedural evolution and verify observed outcomes against it. (b) ATV mirrors this process in object-centric latent space by anticipating normative state evolution and verifying observed transitions. In contrast, (c) video MLLMs mainly recognize and ground visible interactions, while (d) conventional world models predict likely futures without verifying whether the resulting state transitions are procedurally correct.

state changes follow procedurally correct trajectories. For instance, Video-VLMs such as Qwen3-VL (Bai et al., 2025) and Molmo2 (Clark et al., 2026) emphasize spatiotemporal grounding of objects and events, but do not explicitly model continuous state transitions. Video world models such as V-JEPA 2 (Assran et al., 2025) and Cosmos (NVIDIA et al., 2025) predict future evolution in latent or observation space, yet fail to characterize the discrepancy between predicted and observed state transitions. In other words, conventional world models predict what is likely to happen. If the predicted future instead represents normative state evolution, its discrepancy from observation becomes evidence of whether the execution follows the expected procedural dynamics.

To this end, inspired by clinicians' assessment process (St-Onge et al., 2016) (Fig 1 (a)) and *hippocampal reverse replay*[1] (Foster & Wilson, 2006; Diba & Buzsáki, 2007), we propose ATV, an **A**nticipate-**T**hen-**V**erify framework for clinical skill assessment. Given the current procedural state, ATV first anticipates how the underlying object states should evolve under a correct execution, and then verifies the observed action by comparing this normative future with the subsequently observed state. As shown in Fig 1 (b), ATV consists of three progressive modules. (1) *Object-centric Procedural State Representation* (OPSR) organizes video observations into temporally aligned object-centric states through recurrent Slot Attention (Locatello et al., 2020; Kipf et al., 2022), so that interaction-driven changes of the operator, patient, and instruments can be tracked beyond isolated visual events. (2) *Bidirectional Normative Dynamics Modeling* (BNDM) learns normative dynamics exclusively from correctly performed procedures through JEPA-style latent world modeling (Bardes et al., 2024; Assran et al., 2025). Specifically, a forward predictor anticipates the object states that should follow the current procedural state, while a reverse-time reflector models whether the resulting states are compatible with their expected antecedents. (3) *Discrepancy Reasoning for Procedural Verification* (DRPV) compares the observed state evolution against these predictive and reflective normative references, and transforms their discrepancies into rubric-conditioned evidence of procedural correctness. In this way, ATV turns latent world modeling from predicting what is likely to happen into verifying whether what actually happens follows how the procedure should evolve.

To support research on interaction-driven state tracking for procedural correctness assessment, we further construct ClinSkill-V, an expert-annotated video dataset collected from real-world clinical skill examinations. ClinSkill-V covers three clinical skills, totaling about 110 hours of video with an average duration of 25.1 seconds. Each video corresponds to a rubric-defined procedural step and is paired with an expert-defined scoring criterion and an ordinal score. ATV is trained on ClinSkill-V in three stages to progressively learn **object-state representations, normative procedural dynamics, and rubric-conditioned assessment**. To empirically validate ATV, we compare it against 12 open-source Video-MLLMs, Omni-MLLMs, medical Video-MLLMs, and 3 latent predictive mod-

[1]Hippocampal reverse replay refers to the reverse-order reactivation of recently experienced neural sequences, effectively revisiting preceding states from a later outcome. Details are provided in Appendix B.

els. ATV achieves the best performance on all evaluation metrics. Compared with the strongest baseline, Molmo-2, ATV reduces $\mathrm{NMAE}^M$ by $21.9\%$ and improves BAcc@0.5 by 8.9 percentage points, demonstrating the effectiveness of explicitly modeling normative state evolution for procedural correctness assessment.

In summary, the contributions of this paper are three-fold:

- We formulate clinical skill assessment as a **state-transition tracking and procedural correctness verification** task, and introduce **ClinSkill-V**, a physician-annotated dataset consisting of clinical skill videos paired with rubric-defined criteria and expert-assigned ordinal scores.
- We propose **ATV**, an Anticipate-Then-Verify framework with three progressive modules, i.e., OPSR, BNDM, and DRPV. Beyond recognizing observed events or predicting likely futures, ATV learns **bidirectional normative object-state dynamics** and performs **latent discrepancy reasoning** to verify whether observed state transitions follow the expected procedural evolution.
- Extensive experiments on ClinSkill-V demonstrate the superiority of ATV over competing methods as well as the effectiveness of its constituent modules. Further analyses reveal that **generic predictive modeling is insufficient for fine-grained procedural judgment**, highlighting normative state modeling and discrepancy-based verification as a promising direction for evaluative video understanding.

## 2 METHODS

### 2.1 TASK FORMULATION

Clinical skill assessment aims to quantify the quality of a learner's performance on a specific procedural step from video evidence. Given a video $V_{1:T} = \{V_1, \ldots, V_T\}$ depicting one step of a clinical procedure, a textual step-specific scoring rubric $\mathcal{R}$ defining the assessment criteria and score allocation, and a generic scoring prompt $Q$, a model predicts a scalar score:

$$\hat{s} = f_\theta(V_{1:T}, \mathcal{R}, Q), \tag{1}$$

where $\hat{s} \in [0, s_{\max}]$ and $s_{\max}$ denotes the maximum attainable score for the corresponding step. The predicted score reflects the extent to which the performance shown in the video satisfies the criteria specified in $\mathcal{R}$. Following clinician-defined evaluation protocols, model performance is measured by the agreement between the predicted scores $\hat{s}$ and the corresponding clinician-annotated scores $s$ across clinical skill steps.

Table 1: Statistics of ClinSkill-V.

| Skill | Avg. Dur. (s) | # Clips |
|---|---|---|
| CPR | 21.8 | 9,247 |
| AED | 41.8 | 3,808 |
| BMV | 12.7 | 2,720 |
| Full-length videos | 544 | |
| Total duration | 110 h | |
| Rubric steps | 29 | |
| Total clips | 15,775 | |
| Avg. clip duration | 25.1 s | |
| Train / Val. / Test | 11,048 / 1,566 / 3,161 | |

### 2.2 DATASET CURATION

To support the study of interaction-driven state tracking for procedural correctness assessment, we construct a step-level clinical skill video dataset from real procedural assessment recordings collected at our collaborating hospitals[2]. The recordings capture medical students performing clinical procedures under standardized training and examination protocols, naturally exhibiting meaningful variations in execution quality.

In total, we collect 654 full-length recordings spanning three clinical skills: cardiopulmonary resuscitation (CPR), automated external defibrillator use (AED), and bag-mask ventilation (BMV). After quality control, 544 recordings remain and are manually segmented into 15,775 step-aligned clips totaling 110 hours (25.1 s on average), each paired with a rubric criterion and expert-assigned performance score. Detailed statistics for each clinical skill are provided in Table 1.

The resulting dataset[3] consists of step-aligned procedural video clips paired with physician-defined step-specific rubrics and scores, providing fine-grained supervision for modeling clinically relevant state transitions and evaluating step-level procedural quality.

[2]The corresponding IRB approval document is shown in Appendix H.

[3]Details of data collection, annotation, and quality control are provided in Appendix C.

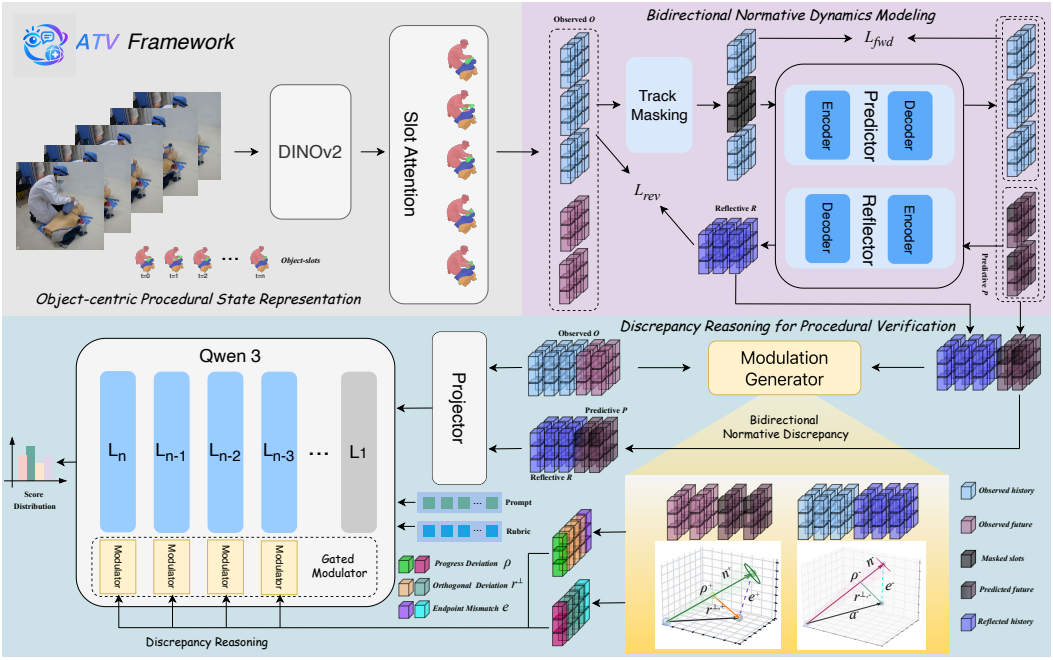



Figure 2: Overview of ATV framework. ATV constructs temporally aligned object-centric states, learns bidirectional normative dynamics through forward anticipation and reverse-time reflection, and reasons over discrepancies between observed and normative state evolution to produce rubric-conditioned procedural assessments. ATV architecture details are provided in Appendix E.

### 2.3 Overview

ATV consists of three stages: *Object-centric Procedural State Representation* (OPSR), *Bidirectional Normative Dynamics Modeling* (BNDM), and *Discrepancy Reasoning for Procedural Verification* (DRPV). As illustrated in Fig. 2, (1) OPSR constructs temporally aligned object states, (2) BNDM learns bidirectional normative dynamics, and (3) DRPV interprets the discrepancies between observed and normative state evolution under the clinical rubric to produce ordinal assessment scores.

### 2.4 Object-centric Procedural State Representation

Clinical skill videos are typically recorded in largely static environments, where procedurally relevant evidence is concentrated in a few interacting entities. Holistic representations are therefore prone to entangling meaningful state changes with redundant background appearance. To tackle this issue, we introduce *Object-centric Procedural State Representation* (OPSR) to transform each video clip from dense visual observations into a compact sequence of slot-based entity representations.

Given a video clip $\{x_t\}_{t=1}^{T}$, we first encode each frame with a frozen DINOv2 (Oquab et al., 2024) patch encoder:

$$F_t = E_{\text{DINO}}(x_t) = [f_{t,1}, \ldots, f_{t,N}], \tag{2}$$

where each $f_{t,n}$ is a semantic patch feature. Following prior work on Slot Attention (Locatello et al., 2020; Kipf et al., 2022) and temporally consistent video slots (Manasyan et al., 2025), we use recurrent slots to group these low-level frozen patch features into temporally aligned entity states:

$$(Z_t, M_t) = \text{SlotAttn}(F_t; Z_{t-1}). \tag{3}$$

Here, $Z_t = \{z_{t,k}\}_{k=1}^{K}$ denotes temporally aligned latent object states used for subsequent dynamics modeling, and $M_t = \{m_{t,k,n}\}$ the corresponding patch-to-slot spatial assignments.

### 2.5 Bidirectional Normative Dynamics Modeling

The design of *Bidirectional Normative Dynamics Modeling* (BNDM) is motivated by complementary prospective and retrospective reasoning. Specifically, while clinicians anticipate the expected

consequences of procedural actions (St-Onge et al., 2016), *hippocampal reverse replay* (Foster & Wilson, 2006; Diba & Buzsáki, 2007) provides a neurocognitive motif for relating resulting states back to their preceding experiences. Accordingly, BNDM models normative dynamics through *Forward Normative State Anticipation* and *Reverse-Time Antecedent Modeling*, which predict normative future object states and reconstruct plausible antecedent states, respectively. To emphasize procedurally informative state transitions during dynamics learning, we further introduce *Interaction-Aware Track Masking* and train the entire framework exclusively on normative procedure videos.

**Forward Normative State Anticipation.** Given the object-centric states extracted by OPSR, the forward branch learns how they should evolve under a correctly performed clinical procedure. Inspired by latent predictive learning (Bardes et al., 2024; Wu et al., 2022), we instantiate the JEPA-style predictor as a Temporal Object Transformer to model procedural dynamics directly in the object-state space. Let $H$ denote the number of observed history states and $F$ the prediction horizon. The transformer first encodes the masked object-state history $\widetilde{Z}^H$ [4]into contextualized history representations. For each future horizon $h$ and object $k$, we construct an object- and horizon-specific query, which attends to the encoded history through the temporal decoder:

$$\begin{gathered}\mathcal{H} = \mathrm{Enc}_{\mathcal{F}}\left(\widetilde{Z}^H; M_{\mathrm{hist}}\right), \qquad q_{h,k} = q_{h,k}^{\mathrm{learn}} + e_h^{\mathrm{time}} + e_k^{\mathrm{slot}}, \\ u_{h,k} = \mathrm{Dec}_{\mathcal{F}}\left(q_{h,k}, \mathcal{H}; M_{\mathrm{future}}\right).\end{gathered} \tag{4}$$

Here, $\mathcal{H}$ denotes the encoded object-state history, $e_h^{\mathrm{time}}$ and $e_k^{\mathrm{slot}}$ encode the future horizon and object-slot identity, respectively, and $u_{h,k}$ is the resulting object- and horizon-specific representation used for future-state prediction. $M_{\mathrm{hist}}$ and $M_{\mathrm{future}}$ denote the causal attention masks used by the history encoder and future decoder. Rather than predicting an absolute future state, we parameterize each prediction as a gated change from the last observed state $z_{H,k}$. Specifically,

$$\begin{aligned}\Delta\widehat{z}_{h,k} &= f_\Delta(u_{h,k}), \qquad g_{h,k} = \sigma\left(f_g(u_{h,k})\right), \\ \widehat{z}_{H+h,k} &= z_{H,k} + g_{h,k}\Delta\widehat{z}_{h,k}, \qquad h = 1, \ldots, F.\end{aligned} \tag{5}$$

Here, $f_\Delta$ and $f_g$ are lightweight prediction heads that map $u_{h,k}$ to the candidate state update and its corresponding update gate, respectively. By anchoring all future predictions to the same observed state $z_{H,k}$, the model predicts each horizon directly without recursive rollout, avoiding error accumulation while allowing horizon-dependent state transitions. During training, future object states from correctly performed procedures serve as normative targets.

**Reverse-Time Antecedent Modeling.** The reverse-time reflector $\mathcal{B}$ adopts the same Temporal Object Transformer architecture as the forward predictor $\mathcal{F}$, but uses independent parameters and directly reconstructs absolute antecedent states, as reverse-time reconstruction lacks the natural shared boundary anchor used for forward residual prediction.[5] Given a future segment $Z_{H+1:H+F}$, we reverse its temporal order and predict the immediately preceding object states:

$$\hat{A} = \mathcal{B}(\mathrm{rev}(Z_{H+1:H+F})), \qquad A = \mathrm{rev}(Z_{H-F+1:H}). \tag{6}$$

Here, $\mathrm{rev}(\cdot)$ denotes temporal reversal, $A$ is the ground-truth antecedent sequence preceding the future segment, and $\hat{A}$ is its reconstruction by the reflector $\mathcal{B}$. Trained on correctly performed procedures, the two branches provide complementary normative references: $\widehat{Z}_{H+1:H+F}$ specifies how object states should evolve forward from the observed boundary, whereas $\widehat{A}$ specifies the antecedent states expected from the resulting future states. In the verification stage, these bidirectional deviations are integrated through latent discrepancy reasoning for rubric-conditioned assessment.

**Interaction-Aware Track Masking.** Random masking (Tong et al., 2022; Wei et al., 2022) ignores persistent object identity and procedural relevance. We instead preferentially mask object tracks exhibiting interaction-associated state changes, encouraging BNDM to infer their evolution

[4] $\widetilde{Z}^H$ is obtained by Interaction-Aware Track Masking in this section, which masks the intermediate states of selected object tracks while retaining their boundary states as temporal anchors.

[5] Forward prediction is anchored at the last observed state $z_{H,k}$, whereas reverse reconstruction targets a sequence of antecedent states at different temporal offsets, for which no equivalent common boundary state is naturally available.

from temporal and cross-object context. For track $k$, we aggregate frame-level evidence from object change, inter-object contact, and coordinated change:

$$\omega_k = \frac{1}{H}\sum_{t=1}^{H} \mathcal{N}\Big(\mathcal{N}(\delta_{t,k}^{\mathrm{chg}}) + \mathcal{N}(\delta_{t,k}^{\mathrm{con}}) + \mathcal{N}(\delta_{t,k}^{\mathrm{co}})\Big). \tag{7}$$

where $\delta_{t,k}^{\mathrm{chg}}$ measures the state change of object $k$ between adjacent frames, $\delta_{t,k}^{\mathrm{con}}$ measures its spatial contact with other objects, $\delta_{t,k}^{\mathrm{co}}$ measures coordinated change with other objects, and $\mathcal{N}(\cdot)$ denotes min–max normalization. Tracks with larger $\omega_k$ are preferentially selected, together with randomly sampled tracks to avoid over-reliance on the interaction heuristic. Their intermediate states are replaced by a learned mask embedding while retaining the boundary states as temporal anchors.

### 2.6 Discrepancy Reasoning for Procedural Verification

BNDM learns how object states should evolve under normative procedures. *Discrepancy Reasoning for Procedural Verification* (DRPV) turns these learned dynamics into assessment evidence by comparing the observed execution against both predictive and reflective normative references and interpreting the resulting deviations under a clinical scoring rubric.

**Bidirectional Normative Discrepancy.** A pointwise prediction error indicates whether an observed state matches its normative reference, but not how the underlying transition deviates from expected procedural evolution. We therefore characterize discrepancy in transition space, capturing both progress along the normative direction and deviation from it. For horizon $h$ and object $k$, the predictive path compares the actual and normative transitions from their common history boundary:

$$a_{h,k}^{+} = z_{H+h,k} - z_{H,k}, \qquad n_{h,k}^{+} = \widehat{z}_{H+h,k} - z_{H,k}. \tag{8}$$

Rather than reducing their disagreement to a single prediction error, we decompose the actual transition into its progress along the normative direction, orthogonal deviation, and endpoint mismatch:

$$\rho_{h,k}^{+} = \frac{\langle a_{h,k}^{+}, n_{h,k}^{+}\rangle}{\|n_{h,k}^{+}\|_2^2 + \epsilon}, \qquad r_{h,k}^{\perp,+} = a_{h,k}^{+} - \rho_{h,k}^{+} n_{h,k}^{+}, \tag{9}$$

$$e_{h,k}^{+} = \|z_{H+h,k} - \widehat{z}_{H+h,k}\|_2^2. \tag{10}$$

where $\epsilon > 0$ is a small constant for numerical stability. The reflective path analogously compares reverse transitions from the observed outcome, with $a_{h,k}^{-} = z_{H+2-h,k} - z_{H+1-h,k}$ and $n_{h,k}^{-} = z_{H+2-h,k} - \widehat{z}_{H+1-h,k}^{-}$, where $\widehat{z}_{H+1-h,k}^{-} = \widehat{A}_{h,k}$ is the predicted antecedent. Applying the same decomposition yields $(\rho_{h,k}^{-}, r_{h,k}^{\perp,-}, e_{h,k}^{-})$.

We combine both directions as

$$d_{h,k} = \Big[\rho_{h,k}^{+} - 1;\ r_{h,k}^{\perp,+};\ e_{h,k}^{+};\ \rho_{h,k}^{-} - 1;\ r_{h,k}^{\perp,-};\ e_{h,k}^{-}\Big]. \tag{11}$$

Thus, $d_{h,k}$ characterizes not only whether an observed state differs from its normative reference, but also whether the corresponding object transition follows the expected procedural direction.

**Rubric-Conditioned Discrepancy Reasoning.** Procedural deviation is meaningful only in the context of what a rubric requires: the same object-state change may be critical for one criterion but irrelevant to another. We therefore use a frozen language backbone to interpret the observed dynamics and their normative references under the target rubric. After projection into the language space, the scorer receives

$$[\mathcal{R}; O; P; R; q_{\mathrm{ans}}], \tag{12}$$

where $\mathcal{R}$ denotes the rubric, $O$ denotes the observed object states, $P$ and $R$ provide the predictive and reflective normative references, and $q_{\mathrm{ans}}$ is a learned query used to produce the final assessment.

Unlike $O$, $P$, and $R$, the bidirectional discrepancy is not an independent state observation, but a relation between an observed transition and its aligned normative references. Let $\mathbf{h}_{\ell,h,k}$ denote the

Table 2: Comparison with representative models for clinical procedural assessment. Open-source MLLMs are evaluated with supervised fine-tuning (SFT) and latent predictive models with supervised task adaptation. **Bold** indicates the best result, and underline indicates the second-best result.

| Model | $\mathrm{NMAE}^M \downarrow$ | BAcc↑ | BAcc@0.5↑ | BAcc@1↑ | MCC↑ |
|---|---|---|---|---|---|
| *Open-source Video-MLLMs* | | | | | |
| LLaVA-OneVision (Li et al., 2025a) | 0.431 | 0.332 | 0.401 | 0.645 | 0.166 |
| LLaVA-Video (Zhang et al., 2024) | 0.431 | 0.333 | 0.409 | 0.649 | 0.179 |
| VideoLLaMA3 (Zhang et al., 2025) | 0.428 | 0.338 | 0.409 | 0.650 | 0.242 |
| Qwen3-VL(8B) (Bai et al., 2025) | 0.426 | 0.336 | 0.412 | 0.652 | 0.205 |
| Qwen3-VL(30B) (Bai et al., 2025) | 0.416 | 0.340 | 0.418 | 0.659 | 0.281 |
| Molmo-2 (Clark et al., 2026) | 0.415 | 0.347 | 0.423 | 0.659 | 0.309 |
| *Open-source Omni-MLLMs* | | | | | |
| Qwen2.5-Omni (Xu et al., 2025a) | 0.422 | 0.326 | 0.400 | 0.653 | 0.237 |
| Qwen3-Omni (Xu et al., 2025b) | 0.433 | 0.331 | 0.407 | 0.647 | 0.153 |
| Baichuan-Omni-1.5 (Li et al., 2025b) | 0.443 | 0.260 | 0.359 | 0.588 | 0.036 |
| *Medical Video-MLLMs* | | | | | |
| uAI-NEXUS-MedVLM (Su et al., 2026) | 0.433 | 0.327 | 0.396 | 0.640 | 0.117 |
| UniMed-VL (Ning et al., 2026) | 0.437 | 0.326 | 0.395 | 0.638 | 0.136 |
| SurgVidLM (Wang et al., 2026) | 0.425 | 0.335 | 0.416 | 0.656 | 0.217 |
| *Latent Predictive Models* | | | | | |
| V-JEPA 2 (Assran et al., 2025) | 0.431 | 0.332 | 0.401 | 0.644 | 0.159 |
| V-JEPA 2.1 (Mur-Labadia et al., 2026) | 0.429 | 0.335 | 0.404 | 0.647 | 0.183 |
| VL-JEPA (Chen et al., 2026) | 0.439 | 0.325 | 0.394 | 0.638 | -0.015 |
| **ATV (Ours)** | **0.324** | **0.410** | **0.512** | **0.714** | **0.312** |

hidden state at language-model layer $\ell$ aligned with the observation $z_{H+h,k}$. We therefore use $d_{h,k}$ to condition the observation representation rather than introducing it as a separate token:

$$\mathbf{h}'_{\ell,h,k} = \mathbf{h}_{\ell,h,k} + \mathcal{M}_\ell\left(\mathbf{h}_{\ell,h,k}; d_{h,k}\right), \tag{13}$$

where $\mathcal{M}$ is a lightweight gated modulator. This design keeps the observed and normative states explicitly available to the language model, while allowing their directional disagreement to directly shape how each observed state is interpreted under the rubric. The answer query then aggregates this discrepancy-aware evidence into a rubric-specific assessment.

# 3 Experiments

## 3.1 Experiment setup

We compare ATV against representative models on ClinSkill-V spanning general-purpose Video-MLLMs, omni-modal MLLMs, medical Video-MLLMs, and latent predictive video models. We evaluate procedural assessment using five metrics: $\mathrm{NMAE}^M$, balanced accuracy (BAcc) (Brodersen et al., 2010), tolerance-based balanced accuracy (BAcc@0.5 and BAcc@1), and Matthews correlation coefficient (MCC) (Matthews, 1975). Together, these metrics measure score estimation error, exact and tolerance-based ordinal agreement, and binary correctness discrimination. Detailed experimental settings and evaluation metrics are provided in Appendices G and D, respectively.

## 3.2 Main Results

**ATV consistently improves procedural assessment across all evaluation criteria.** As shown in Table 2, among all baselines with reported results, ATV achieves the best performance on every metric. Compared with the strongest overall baseline, Molmo-2, ATV reduces $\mathrm{NMAE}^M$ from 0.415 to 0.324, corresponding to a relative reduction of 21.9%, while improving BAcc, BAcc@0.5, and BAcc@1 by 6.3, 8.9, and 5.5 percentage points, respectively. Notably, the improvement remains consistent from exact-score agreement to increasingly tolerant ordinal evaluation, with the largest absolute gain observed at the stricter $\pm 0.5$ tolerance. Consequently, ATV improves not only exact score prediction, but also the overall agreement between predicted assessments and expert scores.

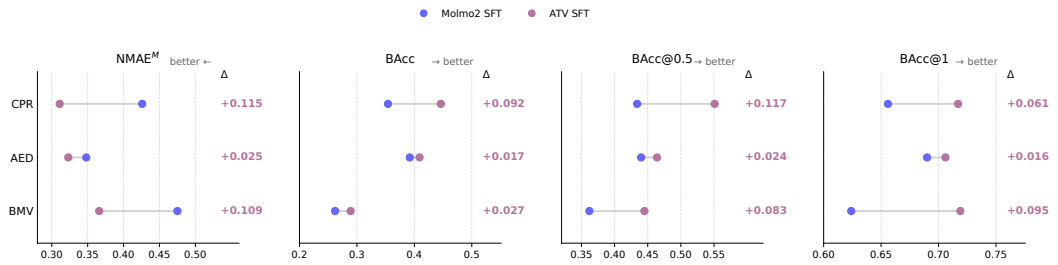



Figure 3: Performance across clinical skills. ATV is compared with the strongest SFT baseline, Molmo2, on CPR, AED, and BMV. $\Delta$ denotes the absolute improvement of ATV over Molmo2, with the direction reversed for $\mathrm{NMAE}^M$ where lower is better.

**The main advantage of ATV lies in fine-grained procedural judgment rather than coarse error detection.** An informative contrast is provided by Molmo-2, which achieves an MCC of 0.309, nearly matching ATV's 0.312, but remains substantially behind on the ordinal assessment metrics. In particular, its BAcc and BAcc@0.5 are only 0.347 and 0.423, compared with 0.410 and 0.512 for ATV. This gap suggests that distinguishing broadly correct from incorrect executions is not sufficient for reliable clinical skill assessment. To assign an appropriate rubric-level score, a model must additionally characterize the extent to which an execution deviates from the expected procedure. ATV is designed precisely for this setting, using observed–normative state discrepancies as structured evidence for rubric-conditioned assessment. Therefore, its gains primarily reflect improved fine-grained procedural judgment rather than merely stronger binary error recognition.

**Generic predictive modeling alone is insufficient for normative procedural assessment.** The latent predictive baselines further support this distinction. V-JEPA 2 and V-JEPA 2.1 explicitly learn predictive video representations, yet obtain substantially lower performance than ATV across all metrics. Relative to V-JEPA 2.1, ATV reduces $\mathrm{NMAE}^M$ from 0.429 to 0.324 and improves BAcc from 0.335 to 0.410, BAcc@0.5 from 0.404 to 0.512, and MCC from 0.183 to 0.312. This suggests that predictive representation learning by itself does not provide sufficient evidence to determine whether an observed state transition is procedurally correct. In contrast, ATV learns dynamics specifically from correctly performed procedures and interprets deviations from these normative trajectories during assessment, turning latent prediction into a reference for procedural verification rather than an end in itself.

**Performance Across Clinical Skills.** We further compare ATV with the strongest SFT baseline, Molmo2, across CPR, AED, and BMV. As shown in Fig. 3, ATV improves all four score-balanced metrics for all three skills, yielding gains in all 12 skill–metric comparisons. The improvements are strongest for CPR and remain substantial for BMV, while AED shows smaller but consistent gains. Overall, the results demonstrate cross-skill consistency without uniform gain magnitude.

### 3.3 Ablation Study

We conduct ablations at two levels: component-level ablations assess the contributions of normative dynamics and discrepancy reasoning, while design-level examine how discrepancy evidence should be represented and integrated. Additional ablation analyses are provided in Appendix F.4.

**Normative Dynamics and Discrepancy Reasoning.** As shown in Fig. 4, removing normative dynamics or discrepancy conditioning reduces BAcc by 6.7 and 7.8 points, respectively, showing that modeling correct procedural evolution alone is insufficient. Assessment also requires reasoning over deviations from normative dynamics. Removing the reverse branch further degrades all ordinal and tolerance-based metrics while slightly improving MCC, suggesting that reverse-time modeling mainly helps characterize the magnitude and ordinal structure of procedural deviations rather than binary error detection. This distinction indicates that fine-grained procedural grading and binary error discrimination capture distinct aspects of assessment quality.

**Discrepancy Representation and Integration.** As shown in Table 3, explicit discrepancy tokens and additive LM injection both underperform the full model, indicating that discrepancy is most

Table 3: Ablation of discrepancy representation and integration in DRPV. We compare structured discrepancy conditioning with explicit discrepancy tokens, additive LM injection, and raw signed prediction errors. All variants use the same data, seed, optimization protocol, frozen language model, and checkpoint selection.

| Variant | $\text{NMAE}^M \downarrow$ | BAcc↑ | BAcc@0.5↑ | BAcc@1↑ | MCC↑ |
|---|---|---|---|---|---|
| **ATV** | **0.324** | **0.410** | **0.512** | **0.714** | 0.312 |
| w/o Discrepancy Conditioning | 0.365 | 0.332 | 0.417 | 0.673 | 0.122 |
| Explicit Discrepancy Tokens | 0.353 | 0.282 | 0.404 | 0.650 | 0.054 |
| Additive LM Injection | 0.366 | 0.383 | 0.461 | 0.679 | 0.276 |
| Raw Signed Prediction Errors | 0.350 | 0.400 | 0.478 | 0.689 | **0.372** |

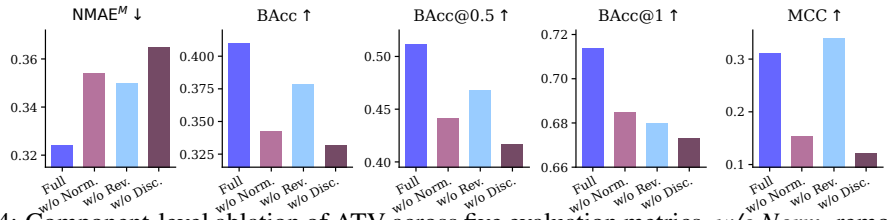



Figure 4: Component-level ablation of ATV across five evaluation metrics. *w/o Norm.* removes both forward and reverse normative dynamics, *w/o Rev.* removes only the reverse-time branch, and *w/o Disc.* removes discrepancy conditioning.

effective when preserved as structured relational evidence between observed transitions and their normative references. Replacing this representation with raw signed prediction errors yields a different performance profile: ordinal and tolerance-based metrics deteriorate, while MCC increases from 0.312 to 0.372. Intuitively, raw residuals provide a strong signal for detecting whether an execution deviates from expectation, whereas structured discrepancies better preserve the degree and structure of deviation required for fine-grained grading. More broadly, procedural assessment should not be reduced to binary anomaly detection, as identifying that an error occurred and estimating how severely an execution departs from the normative procedure require different forms of evidence.

## 4 Related Work

Recent video MLLMs have advanced temporal and long-context reasoning (Li et al., 2025a; Wang et al., 2025; Bai et al., 2025), while video world models learn how visual states are likely to evolve through latent prediction or generation (Bardes et al., 2024; Assran et al., 2025; Wang et al., 2024b). However, these approaches primarily recognize which interactions occur without modeling the underlying state changes they induce, and extrapolate future observations without determining whether those state changes follow procedurally correct trajectories[6].

## 5 Conclusion

In this work, we formulate clinical skill assessment as verifying observed state evolution against normative procedural dynamics, rather than merely recognizing events or predicting likely futures. To this end, we propose ATV, which learns bidirectional normative object-state dynamics from correctly performed procedures and converts the discrepancies between observed and normative state evolution into rubric-conditioned evidence for fine-grained assessment. Extensive experiments on ClinSkill-V show that generic predictive modeling alone is insufficient for procedural judgment, whereas normative dynamics and explicit discrepancy reasoning are both critical for reliable ordinal assessment. Ablation studies further reveal that detecting whether a deviation occurs and estimating its degree require different forms of evidence, with structured discrepancies better preserving the information needed for fine-grained grading. We believe these findings point to world models that serve not only as predictors of future states, but also as normative references for judging whether observed behavior evolves as it should.

[6] A more comprehensive discussion of video MLLMs and video world models is provided in Appendix A.

## AI USE STATEMENT

In this work, Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript, including grammar, wording, fluency, and readability. We did not use generative AI tools for any research tasks requiring mandatory disclosure under the ICLR 2027 AI Policy, including the formulation of research ideas or hypotheses, methodology or experimental design, method implementation, data processing or analysis, interpretation of results, or development of scientific conclusions.

All AI-assisted edits were manually reviewed by the authors to ensure that they preserved the intended scientific meaning and did not introduce or alter technical claims. All scientific content and decisions were made and verified by the authors. We take responsibility for the final content of this work, including text produced or edited with the aid of LLMs.

## ETHICS STATEMENT

This work studies automated clinical skill assessment using videos collected from real-world clinical skill examinations. The dataset involves medical students performing standardized clinical procedures and therefore requires careful consideration of participant privacy and responsible data use. The data collection and research-use protocols were reviewed and approved by the relevant Institutional Review Board (IRB) (Details are provided in Appendix H). Data collection, storage, annotation, and any subsequent release follow the applicable institutional requirements for consent, privacy protection, and de-identification. Any released data will exclude personally identifying information and retain only information necessary for research on procedural assessment.

ATV and ClinSkill-V are intended for research on interaction-driven state tracking for procedural correctness assessment, rather than for autonomous educational or clinical decision making. The proposed system has not been validated for high-stakes uses such as competency certification, licensure decisions, or patient-facing clinical decision support. Model predictions should therefore not be interpreted as authoritative judgments of an individual student's overall clinical competence, which extends beyond the specific procedures, rubrics, and examination settings represented in this study.

More broadly, automated assessment systems may inherit biases from annotation practices, institutional scoring conventions, or the demographic and procedural composition of the collected data. Downstream deployment could consequently produce unreliable or inequitable assessments when applied to populations, institutions, or procedures outside the conditions studied here. We therefore recommend that any future use preserve human expert oversight, make assessment criteria transparent, and treat model outputs as assistive evidence rather than replacements for qualified clinical educators or examiners.

## REPRODUCIBILITY STATEMENT

We provide detailed specifications for reproducing our method and experiments throughout the paper and appendices. Section 2.1 describes the task formulation and the complete ATV framework, and Table 1 reports the dataset statistics and train/validation/test split. Appendix C documents the construction, annotation, and quality control of ClinSkill-V, while Appendix D defines all evaluation metrics. Appendix E provides the architecture, objectives, optimization, and implementation details of OPSR, BNDM, and DRPV. Appendix G specifies the training and inference configurations for all compared baselines, and Appendix F provides additional controlled experiments and ablations. Upon acceptance, the source code and implementation will be released on GitHub. The complete ClinSkill-V dataset, including the video clips, physician-defined rubrics, and expert-assigned scores, will be made available on Hugging Face through a controlled-access procedure, subject to the applicable data-use protocol and institutional requirements, for scientific research purposes only.

## REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL https://arxiv.org/abs/2404.08471.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pp. 3121–3124. IEEE, 2010.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 89791–89807, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/9144aded4e536bc5fb7bfc660a2d7a3d-Paper-Conference.pdf.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28652–28668, June 2026.

Kamran Diba and György Buzsáki. Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, 10(10):1241–1242, 2007.

David J Foster and Matthew A Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683, 2006.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 7346–7355, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681488. URL https://doi.org/10.1145/3664647.3681488.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Wang, Lijuan Wang, and Xin Wang. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 26931–26964, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4364fef031fdf7bfd9d1c9c56b287084-Paper-Conference.pdf.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving, 2023. URL https://arxiv.org/abs/2309.17080.

Vilma Johnsson, Morten Bo Søndergaard, Kulamakan Kulasegaram, Karin Sundberg, Eleonor Tiblad, Lotta Herling, Olav Bjørn Petersen, and Martin G Tolsgaard. Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. *Medical Education*, 58(1):105–117, 2024.

Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *ICLR*, 2022. URL https://openreview.net/forum?id=aD7uesX1GF_.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025, 2025a. ISSN 2835-8856. Publisher Copyright: © 2025, Transactions on Machine Learning Research. All rights reserved.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8511df98c02ab60aea1b2356c013bc0f-Paper.pdf.

Weiheng Lu, Jian Li, An Yu, and Ming-Ching Chang. Llava-mr: Large language-and-vision assistant for video moment retrieval. In *International Conference on Pattern Recognition*, pp. 678–693. Springer, 2026.

Bunmi S Malau-Aduli, Richard B Hays, Karen D'Souza, Amy M Smith, Karina Jones, Richard Turner, Lizzi Shires, Jane Smith, Shannon Saad, Cassandra Richmond, et al. Examiners' decision-making processes in observation-based clinical examinations. *Medical Education*, 55(3):344–353, 2021.

Anna Manasyan, Maximilian Seitzer, Filip Radovic, Georg Martius, and Andrii Zadaianchuk. Temporally consistent object-centric learning by contrasting slots. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5401–5411, 2025. doi: 10.1109/CVPR52734.2025.00508.

JA Martin, Glenn Regehr, Richard Reznick, Helen Macrae, John Murnaghan, Carol Hutchison, and M Brown. Objective structured assessment of technical skill (osats) for surgical residents. *British journal of surgery*, 84(2):273–278, 1997.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-jepa 2.1: Unlocking dense features in video self-supervised learning. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 671–689, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37369-4.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Yuandong Pu, Huihui Xu, Chenhui Gou, Ziyan Huang, Yi Xin, Qi Qin, Diping Song, Bin Fu, Guang Yang, Yuanfeng Ji, Tianbin Li, Yanzhou Su, Jin Ye, Shixiang Tang, Zhongying Deng, Lihao Liu, Ming Hu, and Junjun He. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis, 2026. URL https://arxiv.org/abs/2510.15710.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Christina St-Onge, Martine Chamberland, Annie Lévesque, and Lara Varpio. Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in Health Sciences Education*, 21(3):627–642, 2016.

Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, and Ziyan Wu. Medgrpo: Multi-task reinforcement learning for heterogeneous medical video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2788–2798, June 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Guankun Wang, Junyi Wang, Wenjin Mo, Long Bai, Kun Yuan, Ming Hu, Jinlin Wu, Junjun He, Yiming Huang, Nicolas Padoy, Zhen Lei, Hongbin Liu, Nassir Navab, and Hongliang Ren. Surgvidlm: Towards multi-grained surgical video understanding with large language model, 2026. URL https://arxiv.org/abs/2506.17873.

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

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024a.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens, 2024b. URL https://arxiv.org/abs/2401.09985.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14648–14658, 2022. doi: 10.1109/CVPR52688.2022.01426.

Allison M Whalen, Matthew H Merves, Priyanka Kharayat, James S Barry, Kristen M Glass, Robert A Berg, Taylor Sawyer, Vinay Nadkarni, Donald L Boyer, and Akira Nishisaki. Validity evidence for a novel, comprehensive bag–mask ventilation assessment tool. *The Journal of Pediatrics*, 245:165–171, 2022.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Pandora: Towards general world model with natural language actions and video states, 2024. URL https://arxiv.org/abs/2406.09455.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18493–18503, 2024. doi: 10.1109/CVPR52733.2024.01750.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114v3*, 2023.

Daichi Yashima, Shuhei Kurita, Yusuke Oda, and Komei Sugiura. Remora: Multimodal large language model based on refined motion representation for long-video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31845–31855, June 2026.

Parnian Zameni, Yuhan Shen, and Ehsan Elhamifar. Moscato: Predicting multiple object state change through actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11600–11611, October 2025.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving mllms for long video understanding via grounded tuning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5e8309c9ca683e11672e3dbcd4b87776-Paper-Conference.pdf.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025. URL https://arxiv.org/abs/2501.13106.

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

## APPENDIX ROADMAP

This appendix provides additional background, dataset and evaluation details, implementation specifications, and complementary analyses supporting the main paper. Appendix A reviews related work on video MLLMs and video world models, while Appendix B discusses the computational motivation for bidirectional normative dynamics from hippocampal reverse replay. Appendix C details the collection, annotation, and quality control of ClinSkill-V, and Appendix D defines the evaluation metrics. Appendix E provides the implementation details of the three ATV components: OPSR, BNDM, and DRPV. Appendix F presents complementary experiments on task-specific adaptation, rubric conditioning of normative dynamics, OPSR representations, and discrepancy evidence. Finally, Appendix G reports baseline implementation details, and Appendix H discusses data governance and potential downstream risks.

### Appendix Contents


**A Related Work** 16
- A.1 MLLMs for Video Understanding 16
- A.2 Video World Models 16

**B Motivation from Hippocampal Reverse Replay** 16

**C Data Collection, Annotation, and Quality Control** 17

**D Evaluation Metrics** 18

**E ATV Architecture Details** 19
- E.1 OPSR Details 19
- E.2 BNDM Details 20
- E.3 DRPV Details 22

**F Additional Experiments** 24
- F.1 From Zero-Shot Inference to Task-Specific Adaptation 24
- F.2 Does Normative Dynamics Require Rubric Conditioning? 25
- F.3 Qualitative Analysis of OPSR 27
- F.4 Further Analysis of Discrepancy Evidence 27

**G Experimental Details** 28

**H Extended Ethical Considerations** 30

# A Related work

## A.1 MLLMs for Video Understanding

Recent video MLLMs have substantially advanced open-ended video understanding through stronger long-context modeling, temporal grounding, and spatiotemporal reasoning. Representative models such as LLaVA-Video (Zhang et al., 2024), LLaVA-OneVision (Li et al., 2025a), InternVL3.5 (Wang et al., 2025), Qwen3-VL (Bai et al., 2025), and Molmo2 (Clark et al., 2026) extend general-purpose multimodal models to video, while methods such as TimeSuite (Zeng et al., 2025), LLaVA-MR (Lu et al., 2026), and ReMoRa (Yashima et al., 2026) further improve temporal localization and long-video reasoning. Despite these advances, existing video MLLMs primarily characterize observable video content, including what happens, when it occurs, and how events unfold over time, rather than explicitly modeling how interactions update the underlying states of procedure-relevant entities. Procedural assessment requires a stronger form of temporal reasoning that maintains interaction-induced state transitions and determines whether the resulting evolution is consistent with normative procedural dynamics. Our work addresses this gap by modeling object-level state evolution and using deviations from learned normative trajectories as evidence for procedural correctness.

## A.2 Video World Models

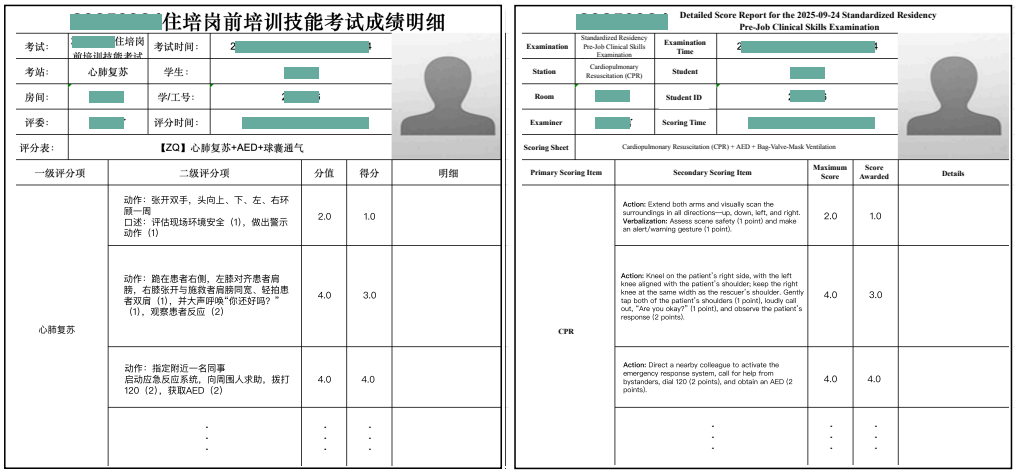
Video world models aim to capture how visual environments evolve over time, either by predicting future representations or by generating future observations. Latent predictive approaches, including V-JEPA (Bardes et al., 2024), V-JEPA 2 (Assran et al., 2025), and V-JEPA 2.1 (Mur-Labadia et al., 2026), learn temporal structure by predicting masked or future representations in feature space, with recent extensions further supporting action-conditioned prediction and planning. Generative world models instead simulate future visual states directly. Representative examples include GAIA-1 (Hu et al., 2023) and DriveDreamer (Wang et al., 2024a) for action-conditioned driving simulation, UniSim (Yang et al., 2023) for interactive real-world simulation, WorldDreamer (Wang et al., 2024b) and Pandora (Xiang et al., 2024) for more general controllable world generation, and Cosmos (NVIDIA et al., 2025) for large-scale world simulation in physical AI. Despite differences in prediction space and conditioning signals, these approaches primarily learn descriptive dynamics that capture how an environment is likely to evolve given its current state and possible actions. Procedural assessment requires a different objective: modeling how procedure-relevant states should evolve under correct execution and determining whether the observed state transitions conform to this normative dynamics. ATV addresses this distinction by learning normative object-state evolution from correctly performed procedures and using observed-to-normative discrepancies as evidence for procedural correctness.

# B Motivation from Hippocampal Reverse Replay

The *hippocampus*, a brain structure closely involved in spatial navigation and memory for temporally organized experiences, exhibits a phenomenon known as *replay*, in which neural activity patterns associated with previous experience are reactivated after or around that experience. In spatial navigation, these patterns can be expressed as sequences of *place cells*, hippocampal neurons whose activity is associated with particular locations in the environment; the temporal ordering of their activity therefore reflects the animal's experienced trajectory.

Foster and Wilson (Foster & Wilson, 2006) observed that hippocampal place-cell sequences associated with recent spatial experience were reactivated during awake periods immediately following behavior, but in the reverse temporal order relative to the original trajectory. This phenomenon, termed *reverse replay*, provides evidence that an experienced temporal sequence can be internally re-expressed from its resulting states back toward earlier states. The authors further suggested that such reverse replay may support the evaluation of preceding event sequences in a manner related to reinforcement learning.

Diba and Buzsáki (Diba & Buzsáki, 2007) subsequently reported *bidirectional replay*: place-cell sequences recurred in reverse order at the end of a run, but in forward order in anticipation of a run. These sequences coincided with *sharp-wave events*, brief highly coordinated population events



[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、右环颜一周<br>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: The physician-defined step-wise rubrics used in real clinical skill assessment.

in the hippocampus that are commonly associated with replay. Their observations demonstrate that previously experienced sequential structure can be reactivated in complementary temporal directions rather than being restricted to its original direction of execution.

We draw only a computational analogy from these findings. In BNDM, the forward branch models how preceding object states should evolve into their expected consequences under normative procedural dynamics, whereas the reverse branch models the complementary relation from resulting states back to plausible antecedent states. Accordingly, hippocampal reverse replay serves as a neurocognitive motif for complementing prospective prediction with retrospective consistency modeling. We do not assume that clinicians employ hippocampal replay during procedural assessment, nor do we claim that BNDM models the biological mechanism of hippocampal replay.

## C Data Collection, Annotation, and Quality Control

**Data Collection.** ClinSkill-V is curated from real clinical skill assessment videos collected from real-world clinical skill examinations. Medical students performed complete procedures rather than expert demonstrations, naturally introducing clinically meaningful variation in execution quality. We collected 654 full-length videos covering cardiopulmonary resuscitation (CPR), automated external defibrillator (AED) operation, and bag-mask ventilation (BMV). After quality control, 544 videos were retained. Each recording preserves a complete procedural attempt with clear visibility of rubric-relevant interactions.

**Annotation.** ClinSkill-V provides dense expert annotations. The collected raw full-length videos were annotated by medical experts following the physician-defined step-wise rubrics(as shown in Fig 5) used in real clinical skill assessment. Each rubric specifies predefined procedural steps and explicit scoring criteria. Two clinically trained physicians independently reviewed each full-length video and assigned rubric-based scores to every predefined step. Thus, each sample contains dual-expert, rubric-aligned step annotations.

**Quality Control.** Quality control was conducted at both the video curation and annotation levels. At the video curation level, we retained only recordings with complete procedural coverage and clear visibility of rubric-relevant interactions, excluding videos that were unsuitable for rubric-based assessment due to interruption, severe occlusion, or incomplete capture. This process filtered out 110 videos. After manually segmenting the retained full-length recordings into rubric-aligned step clips, we further removed 2176 video clips corresponding to four rubric steps involving speech-dependent criteria (as shown in Fig 6), whose correctness cannot be reliably determined from visual evidence alone. At the annotation level, two physicians independently reviewed each full-length video and assigned step-level scores according to the official rubric. Discrepant annotations were jointly re-examined and resolved through step-wise adjudication. Overall, 323 step instances across 102 videos (18.75%) required adjudication, resulting in a single consensus score for each video clip.

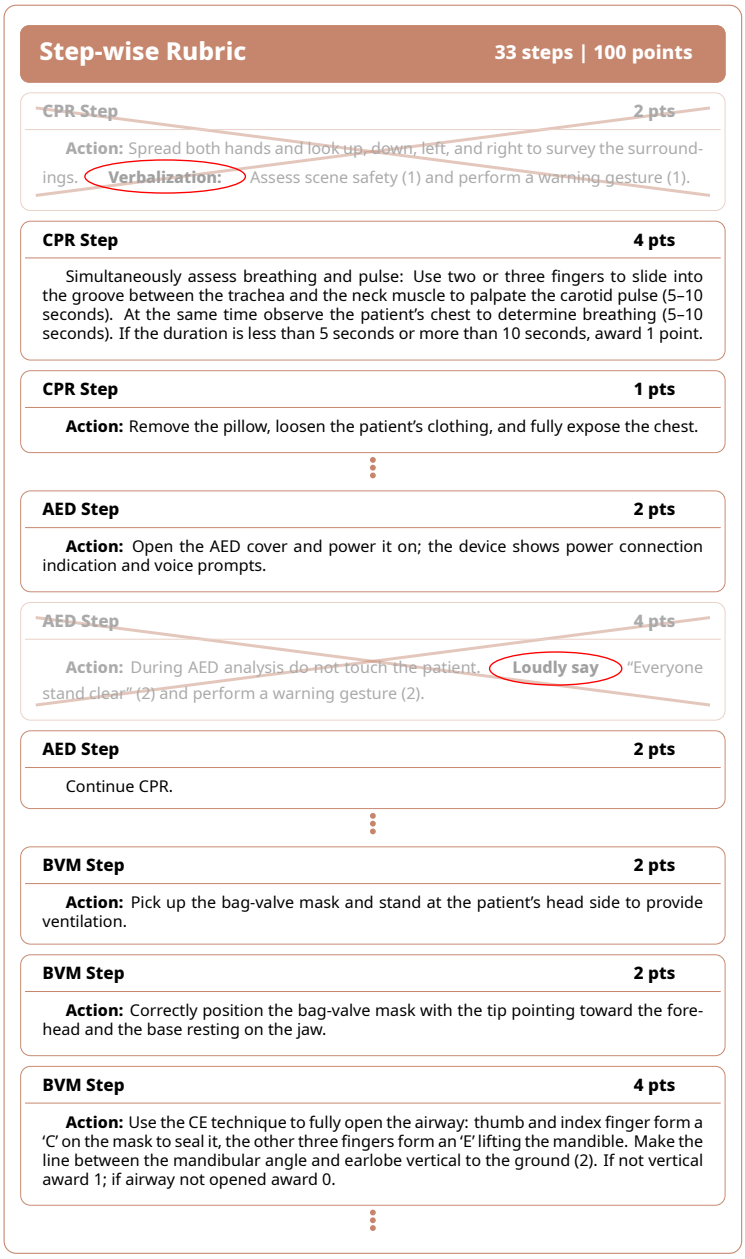



Figure 6: Examples of physician-defined step-wise rubrics used in clinical skill assessment. Speech-dependent criteria (highlighted in red) are excluded because their correctness cannot be reliably determined from visual evidence alone.

After all quality-control procedures, the final dataset contains 15,775 aligned triplets of video clips, step-specific rubrics, and physician-assigned scores. The clips have an average duration of 25.1,s and a total duration of approximately 110 hours, with scores defined within the step-specific admissible range of $[0, 6]$. We split ClinSkill-V at the full-length video level with a 7:1:2 train/validation/test ratio while approximately preserving the data distributions across the three splits.

## D Evaluation Metrics

Clinical assessment scores are highly imbalanced, with some score levels occurring substantially more frequently than others, particularly the maximum score. Direct sample-wise averaging can therefore be dominated by frequent score levels. To account for this imbalance, we use score-

balanced metrics that assign equal importance to the observed ground-truth score levels within each rubric and then average equally across rubrics. We additionally report Matthews correlation coefficient (MCC) to evaluate discrimination between full-score and non-full-score performance.

Let $r \in \{1, \ldots, R\}$ index the rubrics, and let $i$ index samples belonging to rubric $r$. For sample $i$, $y_i$ and $\hat{y}_i$ denote the ground-truth and model-predicted scores, respectively, and $[a_r, b_r]$ denotes the valid score range of rubric $r$. Let $\mathcal{C}_r$ denote the set of ground-truth score levels observed for rubric $r$ in the evaluation split, and let $N_{r,c}$ denote the number of samples with ground-truth score $c \in \mathcal{C}_r$.

**Score balancing.** Within each rubric, all observed ground-truth score levels are assigned equal total weight. Equivalently, each sample $i$ with $y_i = c$ receives weight

$$w_i = \frac{1}{|\mathcal{C}_r| N_{r,c}}. \tag{14}$$

Score levels that do not occur in the evaluation split are excluded. Each metric is first computed with this score balancing within individual rubrics and then macro-averaged across rubrics.

**Score-balanced macro NMAE.** For rubric $r$, we define

$$\mathrm{NMAE}_r^{\mathrm{SB}} = \frac{1}{|\mathcal{C}_r|} \sum_{c \in \mathcal{C}_r} \frac{1}{N_{r,c}} \sum_{i: y_i = c} \frac{|\hat{y}_i - y_i|}{b_r - a_r}. \tag{15}$$

The final score-balanced macro NMAE is

$$\mathrm{NMAE}^M = \frac{1}{R} \sum_{r=1}^{R} \mathrm{NMAE}_r^{\mathrm{SB}}. \tag{16}$$

Lower values indicate better performance.

**Score-balanced accuracy.** For a tolerance $\tau$, the score-balanced accuracy for rubric $r$ is

$$A_r(\tau) = \frac{1}{|\mathcal{C}_r|} \sum_{c \in \mathcal{C}_r} \frac{1}{N_{r,c}} \sum_{i: y_i = c} \mathbb{1}\left[|\hat{y}_i - y_i| \le \tau\right]. \tag{17}$$

We then macro-average across rubrics:

$$\mathrm{BAcc@}\tau = \frac{1}{R} \sum_{r=1}^{R} A_r(\tau). \tag{18}$$

We report BAcc ($\tau = 0$), BAcc@0.5 ($\tau = 0.5$), and BAcc@1 ($\tau = 1$), corresponding to exact-match accuracy and accuracy within 0.5 and 1 score points, respectively. Here, *BAcc* refers to score-balanced accuracy rather than conventional binary balanced accuracy.

**Matthews correlation coefficient.** MCC evaluates whether the model distinguishes non-full-score from full-score performance. We convert the scores into binary indicators

$$t_i = \mathbb{1}[y_i < b_r], \qquad \hat{t}_i = \mathbb{1}[\hat{y}_i < b_r]. \tag{19}$$

Pooling all samples in the evaluation subset, MCC is computed as

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{20}$$

Unlike the score-balanced metrics above, MCC is computed from the pooled full-score/non-full-score confusion matrix and is not macro-averaged over score levels or rubrics.

# E ATV Architecture Details

## E.1 OPSR Details

For OPSR, videos are sampled at 1 FPS and resized to $224 \times 224$. Each frame is encoded by a frozen DINOv2 ViT-S/14 backbone into a $16 \times 16$ grid of 384-dimensional patch features, excluding the class token. We use $K = 6$ slots with dimension $d_s = 256$ and perform three Slot Attention

iterations per frame. The first frame is initialized with six distinct learnable slots, while each subsequent frame is initialized from the slots of the preceding frame, thereby maintaining slot identity over time. Slot updates are implemented using a GRUCell followed by a residual MLP with hidden dimension 512. Each slot is independently decoded back to the 384-dimensional DINOv2 feature space, and the reconstructed patch features are obtained by the mask-weighted combination of the slot-wise predictions.

**Training Objective.** OPSR is trained to reconstruct frozen patch features while encouraging temporal consistency and non-degenerate slot decomposition. The overall objective is

$$\begin{aligned}\mathcal{L}_{\text{OPSR}} = \mathcal{L}_{\text{rec}} + 0.5\mathcal{L}_{\text{temp}} \\ + 0.1\mathcal{L}_{\text{slot-div}} + 0.02\mathcal{L}_{\text{mask-div}}.\end{aligned} \tag{21}$$

$\mathcal{L}_{\text{rec}}$ is the mean-squared reconstruction error between the reconstructed and stop-gradient DINOv2 patch features. To maintain slot identity over time, $\mathcal{L}_{\text{temp}}$ applies bidirectional InfoNCE between corresponding slots in adjacent frames. Slots with the same sample and slot index are treated as positives, while the remaining slot instances in the batch serve as negatives, with temperature $\tau = 0.1$. Slots with the same sample and slot index form positive pairs. The forward-direction loss is

$$\mathcal{L}_{t\to t+1} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(S_{ii}^{(t)})}{\sum_{j=1}^{M}\exp(S_{ij}^{(t)})}, \tag{22}$$

and $\mathcal{L}_{t+1\to t}$ is defined analogously with the two frames exchanged. We then average both directions over adjacent frames:

$$\begin{aligned}\mathcal{L}_{\text{temp}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{1}{2}\big(\mathcal{L}_{t\to t+1} \\ + \mathcal{L}_{t+1\to t}\big).\end{aligned} \tag{23}$$

To discourage redundant slot representations, we minimize the squared pairwise cosine similarity between distinct slots within each frame:

$$\mathcal{L}_{\text{slot-div}} = \frac{1}{BTK(K-1)}\sum_{b,t}\sum_{k\neq j}\cos^2\left(z_{btk}, z_{btj}\right). \tag{24}$$

We further regularize the spatial assignments using an entropy term $\mathcal{L}_{\text{ent}}$ that encourages confident patch-to-slot assignments and a balance term $\mathcal{L}_{\text{bal}}$ that encourages uniform slot utilization. The complete mask regularization term is

$$\mathcal{L}_{\text{mask-div}} = \mathcal{L}_{\text{ent}} + 2\mathcal{L}_{\text{bal}}. \tag{25}$$

We train OPSR for 50 epochs using AdamW with a constant learning rate of $10^{-4}$ and weight decay of $0.05$. Training uses BF16 precision and gradient clipping with a maximum norm of $1.0$. We use 8 GPUs(NVIDIA A800) with a per-GPU batch size of 4 and gradient accumulation over 8 steps, resulting in an effective global batch size of 256. All experiments use random seed 42. The checkpoint achieving the lowest overall validation objective is used to initialize the subsequent dynamics modeling stage.

### E.2 BNDM Details

**Temporal Object Transformer.** BNDM operates on object-centric state sequences produced by OPSR. We use a history length of $H = 8$ and predict up to $F = 4$ future states. The forward predictor $\mathcal{F}$ and reverse predictor $\mathcal{B}$ use independent Temporal Object Transformers with an input slot dimension of 256 and a hidden dimension of 384. Each predictor consists of a 6-layer pre-norm Transformer encoder and a 3-layer pre-norm decoder with 6 attention heads. For $\mathcal{F}$, each future horizon and object slot further has an independent learned query. The two predictors do not share parameters.

**Interaction-Aware Track Masking.** During BNDM training, object tracks are selected for masking according to interaction evidence derived from object change, spatial contact, and coordinated change. Let $c_{t,k}$ and $m_{t,k,n}$ denote the centroid and spatial assignment of slot $k$, respectively, and let $\delta^{\mathrm{chg}}_{t,k}$ denote its frame-to-frame change score. The contact score is defined as

$$\delta^{\mathrm{con}}_{t,k} = \max_{j \neq k} \left[ 0.7 \exp\left( -\frac{\|c_{t,k} - c_{t,j}\|_2}{0.25} \right) + 0.3 \frac{\sum_n m_{t,k,n} m_{t,j,n}}{\sum_n m_{t,k,n}} \right]. \tag{26}$$

The coordinated-change score is defined as

$$\delta^{\mathrm{co}}_{t,k} = u_{t,k} \frac{1}{K} \sum_{j \neq k} \exp\left( - |u_{t,k} - u_{t,j}| \right). \tag{27}$$

The change, contact, and coordinated-change scores are independently min–max normalized within each sample, summed with equal weights, and normalized again to obtain the interaction score used for track ranking. Each training sample masks up to two complete slot tracks. Track selection combines interaction-ranked and random sampling with probabilities $0.7$ and $0.3$, respectively. The first and last history states are always retained as temporal anchors.

**Training Objective** The forward branch is optimized to predict normative future object states while preserving temporally coherent dynamics. Its objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{fwd}} = \mathcal{L}_{\mathrm{future}} + 0.5\mathcal{L}_{\mathrm{history}} + 0.1\mathcal{L}_{\cos} \\ + 0.1\mathcal{L}_{\mathrm{NCE}} + 0.25\mathcal{L}_{\mathrm{vel}} + 0.2\mathcal{L}_{\mathrm{static}}. \end{aligned} \tag{28}$$

$\mathcal{L}_{\mathrm{future}}$ is the mean-squared error between predicted future states and their stop-gradient targets. $\mathcal{L}_{\mathrm{history}}$ reconstructs the masked intermediate object states using mean-squared error. $\mathcal{L}_{\cos}$ further aligns predicted and target future states using cosine distance. $\mathcal{L}_{\mathrm{NCE}}$ applies contrastive alignment between predicted and target future slots using in-batch negatives with temperature $\tau = 0.1$. To preserve the temporal evolution of object states, we match successive state changes. Defining $\widehat{z}_{b,H,k} = z_{b,H,k}$, the velocity loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{vel}} = \frac{1}{BFKd_s} \sum_{b,h,k} \Big\| \widehat{z}_{b,H+h,k} - \widehat{z}_{b,H+h-1,k} \\ - \left( z_{b,H+h,k} - z_{b,H+h-1,k} \right) \Big\|_2^2. \end{aligned} \tag{29}$$

We further discourage spurious changes for nearly static objects. Using the last observed state $a_{bk} = z_{b,H,k}$ as an anchor, we define the target displacement:

$$\Delta^{\mathrm{target}}_{bhk} = \frac{1}{d_s} \|z_{b,H+h,k} - a_{bk}\|_2^2 \tag{30}$$

and weight each slot by $w_{bhk} = \mathrm{sg}[\exp(-\Delta^{\mathrm{target}}_{bhk}/0.005)]$. The static consistency loss is

$$\mathcal{L}_{\mathrm{static}} = \frac{\sum_{b,h,k} w_{bhk} \|\widehat{z}_{b,H+h,k} - a_{bk}\|_2^2 / d_s}{\max\left( \sum_{b,h,k} w_{bhk}, 10^{-6} \right)}. \tag{31}$$

The reverse branch is trained from both observed and forward-predicted future states. Let

$$A = \mathrm{rev}\left( Z_{H-F+1:H} \right) \tag{32}$$

denote the target antecedent sequence. Given the observed future,

$$\widehat{A}^{\mathrm{obs}} = \mathcal{B}(\mathrm{rev}\left( Z_{H+1:H+F} \right)), \tag{33}$$

whereas the predicted-future pathway is

$$\widehat{A}^{\mathrm{pred}} = \mathcal{B}\Big( \mathrm{rev}\Big( \mathrm{sg}\left( \widehat{Z}_{H+1:H+F} \right) \Big) \Big). \tag{34}$$

Thus, supervision through the latter pathway updates $\mathcal{B}$ but does not backpropagate into the forward predictor $\mathcal{F}$.

For the observed-future pathway, we supervise antecedent reconstruction with MSE, cosine-distance, and InfoNCE losses. InfoNCE is applied to $\ell_2$-normalized antecedent-slot representations using in-batch negatives with temperature $0.1$.

The predicted-future pathway is supervised by

$$\mathcal{L}_{\text{pred-MSE}} = \frac{1}{BFKd_s} \left\| \widehat{A}^{\text{pred}} - \text{sg}(A) \right\|_2^2, \tag{35}$$

together with cross-path alignment

$$\mathcal{L}_{\text{align}} = \frac{1}{BFKd_s} \left\| \widehat{A}^{\text{pred}} - \text{sg}\left(\widehat{A}^{\text{obs}}\right) \right\|_2^2. \tag{36}$$

The complete reverse objective is

$$\begin{aligned} \mathcal{L}_{\text{rev}} = 0.5\mathcal{L}_{\text{obs-MSE}} + 0.05\mathcal{L}_{\text{obs-cos}} + 0.05\mathcal{L}_{\text{obs-NCE}} \\ + 0.1\mathcal{L}_{\text{pred-MSE}} + 0.05\mathcal{L}_{\text{align}}. \end{aligned} \tag{37}$$

Finally, BNDM is optimized with

$$\mathcal{L}_{\text{BNDM}} = \mathcal{L}_{\text{fwd}} + \mathcal{L}_{\text{rev}}. \tag{38}$$

We first train the forward predictor $\mathcal{F}$ for 50 epochs. We then randomly initialize the reverse predictor $\mathcal{B}$ and jointly refine the bidirectional model. During the first five epochs of the second stage, $\mathcal{F}$ is frozen, while $\mathcal{B}$ and the learned mask embedding are optimized with a learning rate of $1 \times 10^{-4}$. From epoch 5 onward, $\mathcal{F}$ is optimized with a reduced learning rate of $2.5 \times 10^{-5}$, while $\mathcal{B}$ and the learned mask embedding retain a learning rate of $1 \times 10^{-4}$.Training uses AdamW with a weight decay of $0.05$, BF16 precision, gradient clipping at $1.0$, and a random seed of 42.

### E.3 DRPV Details

DRPV operates on frozen object-centric evidence extracted by the pretrained OPSR and BNDM modules. The scoring model receives three complementary state sequences: the observed object states $O \in \mathbb{R}^{12\times K\times 256}$, the prospective states $P \in \mathbb{R}^{4\times K\times 256}$ predicted by the forward normative model, and the reflective states $R \in \mathbb{R}^{4\times K\times 256}$ reconstructed by the reverse model from the observed future. Each window therefore contributes 72, 24, and 24 tokens from $O$, $P$, and $R$, respectively.

**State Projection and Sequence Construction.** The three state pathways are independently mapped into the language-model hidden space using pathway-specific two-layer MLP projectors ($256 \to 512 \to 4096$) with LayerNorm, GELU, and dropout $0.1$. Each projected token is further augmented with learned temporal, slot-identity, and pathway embeddings:

$$x_{t,k}^{p} = \text{Proj}_p(z_{t,k}^{p}) + e_t^{\text{time}} + e_k^{\text{slot}} + e_p^{\text{dir}}. \tag{39}$$

The complete causal input sequence is

$$[\text{rubric}; O; P; R; q_{\text{ans}}], \tag{40}$$

where $q_{\text{ans}}$ is a learned answer query placed at the end of the sequence to aggregate all preceding rubric and procedural evidence.

**Bidirectional Discrepancy Conditioning.** Concatenating the forward and reflective discrepancy components yields $d_{h,k} \in \mathbb{R}^{516}$, which is projected into the language-model hidden space using $\text{LN}(516) \to 512 \to \text{GELU} \to \text{Dropout}(0.1) \to 4096$. The encoded condition is further augmented with its aligned temporal and slot-identity embeddings.

Discrepancy conditioning is applied only to observation tokens corresponding to the future-aligned states $S9{:}S12$. The remaining observation states, prospective and reflective tokens, rubric tokens,

and answer query are left unmodified. Each of the last four frozen language-model blocks uses an independent lightweight generator $G_\ell : 4096 \to 64 \to 3 \times 4096$ with SiLU to predict $(\Delta\gamma, \beta, g)$. The aligned hidden state is modulated as

$$h'_{h,k} = h_{h,k} + \sigma(g)\left[\Delta\gamma \odot \mathrm{LN}(h_{h,k}) + \beta\right]. \tag{41}$$

The final linear layer of every generator is initialized to zero, making the modulation an exact identity transformation at initialization. We do not apply dataset-level normalization to the directional residuals, preserving their latent direction and magnitude as discrepancy signals.

**Training Objective** Clinical skill scoring is neither standard classification nor unconstrained regression. Its label space is ceiling-heavy, with many correctly performed samples receiving the maximum score, while non-full scores encode graded levels of procedural deficiency. Moreover, these score levels have an intrinsic ordinal geometry: confusing two adjacent scores should not be treated equivalently to a large scoring error. We therefore formulate assessment with a ceiling-aware hierarchical ordinal objective that separates full-score detection from non-full severity estimation while preserving ordinal distance.

We train DRPV with a ceiling-aware hierarchical ordinal objective that separates full-score detection from non-full severity estimation while preserving ordinal structure. For sample $i$, let $\{\ell_{iq}\}_{q=0}^{L_i-1}$ denote the logits over the $L_i$ admissible score levels, with corresponding scores

$$v_{iq} = a_i + 0.5q, \qquad q = 0, \ldots, L_i - 1, \tag{42}$$

and let $p_{iq} = \mathrm{softmax}(\ell_i)_q$. The ground-truth class index and the full-score class are

$$t_i = \mathrm{round}\left(\frac{y_i - a_i}{0.5}\right), \qquad f_i = L_i - 1, \qquad d_i = \mathbb{1}[t_i < f_i]. \tag{43}$$

We first distinguish non-full from full-score performance by aggregating all non-full logits:

$$u_i = \log \sum_{q=0}^{f_i-1} \exp(\ell_{iq}) - \ell_{i,f_i}. \tag{44}$$

The defect-detection loss is a class-weighted binary cross-entropy,

$$\begin{aligned}\mathcal{L}_{\mathrm{defect}} = -\frac{1}{B}\sum_{i=1}^{B}\Big[&\alpha d_i \log\sigma(u_i) \\ &+ (1-d_i)\log(1-\sigma(u_i))\Big],\end{aligned} \tag{45}$$

where $\alpha = 2.1$ weights non-full samples.

For samples whose ground-truth score is non-full, we further predict the severity level conditioned on the non-full subset. Let $\mathcal{D} = \{i : d_i = 1\}$. The conditional severity loss is

$$\mathcal{L}_{\mathrm{severity}} = -\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \log \frac{\exp(\ell_{i,t_i})}{\sum_{q=0}^{f_i-1}\exp(\ell_{iq})}, \tag{46}$$

To preserve the ordinal geometry of the score space, we additionally penalize probability mass according to its normalized distance from the ground-truth score:

$$\mathcal{L}_{\mathrm{ord}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{q=0}^{L_i-1} p_{iq}\frac{|v_{iq} - y_i|}{b_i - a_i}. \tag{47}$$

The overall objective is

$$\mathcal{L}_{\mathrm{C}} = \mathcal{L}_{\mathrm{defect}} + \mathcal{L}_{\mathrm{severity}} + 0.25\mathcal{L}_{\mathrm{ord}}. \tag{48}$$

All OPSR and BNDM parameters are frozen during DRPV training. The Qwen3 backbone, its input embeddings, and its native LM head are also frozen. We optimize only the three slot projectors, temporal and slot-identity embeddings, pathway embeddings, discrepancy encoder, answer query, and four discrepancy modulators, totaling $14{,}138{,}120$ trainable parameters. DRPV is trained with AdamW using a constant learning rate of $2\times10^{-4}$, weight decay $0.01$, BF16 precision, and gradient clipping at $1.0$. Training uses $24$ NVIDIA A800 GPUs with a per-GPU microbatch size of $1$ and gradient accumulation over $11$ steps, yielding an effective global batch size of $264$.

Table 4: Comparison between zero-shot inference and supervised fine-tuning (SFT) for clinical procedural assessment. Gray entries denote zero-shot results. For SFT results, the value in parentheses indicates the change from the corresponding zero-shot result; red denotes improved performance and green denotes degraded performance.

| Model | Setting | $NMAE^M \downarrow$ | BAcc↑ | BAcc@0.5↑ | BAcc@1↑ | MCC↑ |
|---|---|---|---|---|---|---|
| *Open-source Video-MLLMs* | | | | | | |
| LLaVA-OneVision | Zero-shot | 0.400 | 0.190 | 0.281 | 0.737 | 0.129 |
| | SFT | 0.431 (+0.031) | 0.332 (+0.143) | 0.401 (+0.120) | 0.645 (-0.092) | 0.166 (+0.037) |
| LLaVA-Video | Zero-shot | 0.395 | 0.190 | 0.280 | 0.748 | 0.129 |
| | SFT | 0.431 (+0.036) | 0.333 (+0.143) | 0.409 (+0.129) | 0.649 (-0.099) | 0.179 (+0.050) |
| VideoLLaMA3 | Zero-shot | 0.417 | 0.148 | 0.317 | 0.705 | 0.085 |
| | SFT | 0.428 (+0.011) | 0.338 (+0.190) | 0.409 (+0.092) | 0.650 (-0.055) | 0.242 (+0.157) |
| Qwen3-VL(8B) | Zero-shot | 0.432 | 0.160 | 0.350 | 0.667 | 0.047 |
| | SFT | 0.426 (-0.006) | 0.336 (+0.177) | 0.412 (+0.063) | 0.652 (-0.015) | 0.205 (+0.158) |
| Qwen3-VL(30B) | Zero-shot | 0.474 | 0.196 | 0.307 | 0.595 | 0.051 |
| | SFT | 0.416 (-0.058) | 0.340 (+0.144) | 0.418 (+0.111) | 0.659 (+0.064) | 0.281 (+0.230) |
| Molmo-2 | Zero-shot | 0.429 | 0.217 | 0.282 | 0.675 | 0.084 |
| | SFT | 0.415 (-0.014) | 0.347 (+0.130) | 0.423 (+0.141) | 0.659 (-0.016) | 0.309 (+0.225) |
| *Open-source Omni-MLLMs* | | | | | | |
| Qwen2.5-Omni | Zero-shot | 0.418 | 0.139 | 0.374 | 0.678 | 0.008 |
| | SFT | 0.422 (+0.004) | 0.326 (+0.187) | 0.400 (+0.026) | 0.653 (-0.025) | 0.237 (+0.229) |
| Qwen3-Omni | Zero-shot | 0.421 | 0.241 | 0.308 | 0.689 | 0.079 |
| | SFT | 0.433 (+0.012) | 0.331 (+0.091) | 0.407 (+0.100) | 0.647 (-0.042) | 0.153 (+0.074) |
| *Medical Video-MLLMs* | | | | | | |
| uAI-NEXUS-MedVLM | Zero-shot | 0.440 | 0.190 | 0.337 | 0.611 | 0.047 |
| | SFT | 0.433 (-0.007) | 0.327 (+0.137) | 0.396 (+0.059) | 0.640 (+0.029) | 0.117 (+0.070) |
| UniMed-VL | Zero-shot | 0.524 | 0.213 | 0.286 | 0.480 | 0.033 |
| | SFT | 0.437 (-0.087) | 0.326 (+0.113) | 0.395 (+0.109) | 0.638 (+0.158) | 0.136 (+0.103) |
| SurgVidLM | Zero-shot | 0.378 | 0.222 | 0.309 | 0.747 | 0.123 |
| | SFT | 0.425 (+0.047) | 0.335 (+0.113) | 0.416 (+0.107) | 0.656 (-0.091) | 0.217 (+0.094) |
| *Proprietary MLLMs* | | | | | | |
| MiniMax-M3 | Zero-shot | 0.479 | 0.178 | 0.347 | 0.569 | 0.007 |
| Gemini-3.8-Flash | Zero-shot | 0.393 | 0.093 | 0.466 | 0.661 | 0.008 |
| GPT-5.6-Sol | Zero-shot | 0.411 | 0.174 | 0.434 | 0.672 | 0.091 |
| **ATV** | Ours | **0.324** | **0.410** | **0.512** | **0.714** | **0.312** |

# F Additional Experiments

## F.1 From Zero-Shot Inference to Task-Specific Adaptation

We further compare zero-shot inference with supervised fine-tuning (SFT) to examine how much clinical procedural assessment can be supported by pretrained video-language capabilities alone, and how task-specific supervision changes model behavior. This comparison complements the main SFT evaluation by separating generic multimodal competence from task adaptation, while revealing where direct video-to-score learning remains limited.

**SFT reshapes the error structure rather than uniformly improving scoring.** As shown in Tab 4, task-specific supervision consistently improves BAcc, BAcc@0.5, and MCC across all 11 adapted models, indicating substantially stronger discrimination of procedural performance. However, these gains do not translate uniformly to other metrics: NMAE degrades for 6 of 11 models, while BAcc@1 decreases for 8 of 11. Thus, SFT does not simply move predictions closer to the ground truth under every criterion; instead, it changes the error profile toward finer score discrimination. This distinction highlights the importance of evaluating procedural assessment with complementary metrics rather than a single notion of score agreement.

**Directly fine-tuned MLLMs converge to a narrow performance regime.** Despite substantial differences in architecture, scale, and pretraining domain, directly fine-tuned MLLMs exhibit remarkably similar performance. Their SFT results concentrate within narrow ranges of 0.415–0.437 NMAE, 0.326–0.347 BAcc, 0.395–0.423 BAcc@0.5, and 0.638–0.659 BAcc@1. This limited variation suggests an apparent shared bottleneck in direct video-to-score adaptation, rather than performance being primarily determined by the choice of video-language backbone. In contrast, ATV

Table 5: Direct evaluation of normative dynamics under different rubric-conditioning strategies. All metrics are computed on normative test samples. **Bold** and underline denote the best and second-best results when distinguishable at the reported precision; ties share the same formatting.

| Skill | Variant | Forward MSE ↓ | Forward Cosine ↑ | Reverse MSE ↓ |
|---|---|---|---|---|
| Overall | No Condition | 0.023 | 0.920 | 0.024 |
| | Rubric-Text | 0.023 | **0.921** | 0.024 |
| | Shuffled Rubric | 0.023 | 0.920 | 0.024 |
| CPR | No Condition | 0.023 | 0.923 | 0.023 |
| | Rubric-Text | **0.022** | **0.925** | 0.023 |
| | Shuffled Rubric | **0.022** | 0.923 | 0.023 |
| AED | No Condition | 0.025 | **0.917** | 0.025 |
| | Rubric-Text | **0.024** | 0.916 | 0.025 |
| | Shuffled Rubric | 0.025 | 0.916 | 0.025 |
| BMV | No Condition | 0.024 | **0.917** | 0.024 |
| | Rubric-Text | 0.024 | 0.916 | 0.024 |
| | Shuffled Rubric | 0.024 | 0.915 | 0.024 |

operates outside this regime on the fine-grained assessment metrics, supporting the use of explicit normative dynamics and discrepancy-based reasoning.

**Coarse tolerance can overstate fine-grained assessment capability.** High tolerance-based agreement does not necessarily imply reliable procedural judgment. For example, zero-shot LLaVA-Video and SurgVidLM achieve BAcc@1 scores of 0.748 and 0.747, respectively, both exceeding ATV's 0.714, while their exact BAcc remains substantially lower at 0.190 and 0.222 compared with 0.410 for ATV. Similar discrepancies are observed across other zero-shot models. These results show that coarse tolerance metrics can mask weak fine-grained discrimination, motivating joint evaluation of exact-score accuracy, tolerance-based agreement, and score error.

**Zero-shot competence is only weakly indicative of post-adaptation performance.** The relative ordering of models changes substantially after SFT. Qwen3-Omni achieves the highest zero-shot BAcc among the evaluated open-source models but is no longer among the strongest models after adaptation, whereas Molmo-2 shows only moderate zero-shot performance yet becomes the strongest direct-SFT baseline. Similarly, SurgVidLM performs favorably in zero-shot NMAE but does not retain this advantage after SFT. These results indicate that zero-shot capability and downstream adaptation performance capture distinct properties of video MLLMs, and that zero-shot ranking alone is insufficient for selecting models for procedural assessment.

**Medical specialization does not yield a consistent advantage.** Medical Video-MLLMs do not exhibit a systematic advantage over general-purpose models in either zero-shot or SFT settings. Their zero-shot performance varies substantially, and after adaptation their BAcc values fall within the same narrow range as general-purpose Video-MLLMs and Omni-MLLMs. This suggests that medical-domain pretraining alone does not directly provide the fine-grained normative judgment required for procedural assessment. Recognizing medical entities and actions is therefore distinct from determining whether their state evolution conforms to a procedure-specific scoring criterion.

### F.2 Does Normative Dynamics Require Rubric Conditioning?

ATV deliberately separates normative dynamics modeling from rubric-conditioned assessment. BNDM learns how object states should evolve under correct execution from visual state history alone, whereas the rubric is introduced only in DRPV to determine how deviations from the normative trajectory should be interpreted for scoring. This design implicitly assumes that *what constitutes normative state evolution* is largely recoverable from procedural history, while the rubric primarily specifies *which deviations are relevant to the current assessment criterion*. However, because each video clip corresponds to a rubric-defined procedural step, an alternative possibility is that normative evolution itself depends on the target rubric. We therefore ask whether BNDM requires explicit

Table 6: Procedural assessment performance under different BNDM rubric-conditioning strategies. DRPV receives the correct rubric in all variants; only the conditioning of BNDM is changed. **Bold** and underline denote the best and second-best results when distinguishable at the reported precision.

| Skill | Variant | $\mathrm{NMAE}^M \downarrow$ | BAcc ↑ | BAcc@0.5 ↑ | BAcc@1 ↑ | MCC ↑ |
|---|---|---|---|---|---|---|
| Overall | No Condition | **0.331** | 0.401 | **0.489** | 0.695 | 0.297 |
| | Rubric-Text | 0.351 | 0.401 | 0.477 | **0.697** | 0.310 |
| | Shuffled Rubric | 0.352 | 0.401 | 0.477 | 0.696 | **0.316** |
| CPR | No Condition | **0.332** | **0.432** | **0.511** | 0.696 | **0.344** |
| | Rubric-Text | 0.352 | 0.411 | 0.501 | 0.714 | 0.331 |
| | Shuffled Rubric | 0.353 | 0.410 | 0.503 | **0.715** | 0.339 |
| AED | No Condition | **0.320** | **0.429** | **0.483** | 0.667 | 0.323 |
| | Rubric-Text | 0.348 | 0.400 | 0.447 | **0.694** | 0.336 |
| | Shuffled Rubric | 0.350 | 0.400 | 0.448 | **0.694** | **0.338** |
| BMV | No Condition | **0.341** | 0.258 | 0.424 | **0.728** | 0.117 |
| | Rubric-Text | 0.351 | **0.372** | **0.435** | 0.640 | 0.205 |
| | Shuffled Rubric | 0.352 | 0.369 | 0.432 | 0.637 | **0.211** |

rubric semantics to model normative state evolution, or whether such dynamics can be sufficiently inferred from visual history alone.

To isolate this question, we compare three controlled variants. **No Condition** uses the same BNDM architecture as the main model but receives no rubric input. **Rubric-Text** conditions both the forward and reverse branches on the embedding of the corresponding textual rubric. **Shuffled Rubric** uses the same trained rubric-conditioned model but replaces the correct rubric at evaluation with a wrong rubric from another step of the same skill. This same-skill intervention controls for coarse skill identity while testing whether the model actually relies on step-specific rubric semantics. Importantly, DRPV receives the correct rubric in all three settings, so the comparison isolates the role of rubric conditioning in constructing the normative reference rather than in defining the final scoring criterion. All variants are independently retrained under the same protocol for this controlled study.

**Rubric semantics provide limited additional information for normative dynamics.** As shown in Table 5, explicit rubric conditioning produces only marginal changes in normative dynamics. At the overall level, Rubric-Text increases forward cosine similarity only slightly, from 0.920 to 0.921, while both forward and reverse MSE remain effectively unchanged at the reported precision. More importantly, replacing the correct rubric with a same-skill shuffled rubric yields nearly identical dynamics quality. The same pattern largely persists across CPR, AED, and BMV. These results suggest that most of the information required to model normative state evolution is already contained in the procedural state history, with only limited additional contribution from step-specific rubric semantics.

**Marginal changes in predictive fidelity do not translate into better assessment.** As shown in Table 6, the small changes in normative prediction do not yield consistent improvements in downstream procedural assessment. No Condition achieves an overall $\mathrm{NMAE}^M$ of 0.331 and BAcc@0.5 of 0.489, whereas Rubric-Text obtains 0.351 and 0.477, respectively; exact BAcc remains unchanged at the reported precision (0.401). Shuffled Rubric is again nearly indistinguishable from correct rubric conditioning, with $\mathrm{NMAE}^M = 0.352$, $\mathrm{BAcc} = 0.401$, and $\mathrm{BAcc@0.5} = 0.477$. This correct-versus-shuffled similarity also persists across individual skills. Although some skill-specific metrics change under rubric conditioning, similar changes are often observed with shuffled rubrics, indicating that such variation should not be attributed to the use of rubric-specific semantics. Overall, these results show that marginal changes in latent predictive fidelity do not necessarily make the resulting normative reference more useful for procedural assessment.

**Normative dynamics and evaluative semantics are largely separable.** Taken together, these results support ATV's separation between modeling how a correctly executed procedure should evolve

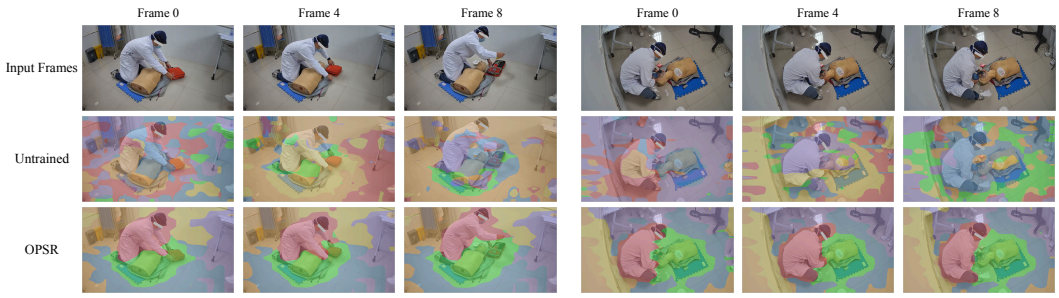



Figure 7: Qualitative visualization of object-centric representations learned by OPSR. We compare slot assignments before and after OPSR training across multiple frames from two random clinical procedure videos. The trained OPSR produces more coherent and temporally consistent object-centric decomposition than the untrained initialization.

and interpreting the significance of deviations from that evolution. For step-aligned procedural assessment, visual state history appears to provide most of the information needed to construct normative dynamics, whereas rubric semantics are more useful downstream for determining which discrepancies matter for the target criterion. More broadly, this analysis suggests that predictive fidelity, semantic conditioning, and evaluative utility should be treated as distinct properties of normative world models rather than assumed to improve together.

### F.3 Qualitative Analysis of OPSR

To examine whether OPSR learns the structured state representation required by subsequent normative dynamics modeling, we qualitatively visualize its patch-to-slot assignments in Fig. 7. We select temporally separated frames from two random procedural clips and compare the assignments produced by an untrained OPSR with those obtained after training. Each color denotes one slot identity and is kept consistent across frames. These visualizations correspond directly to the spatial assignment $M_t$ in Eq. (3), rather than predictions from an additional semantic segmentation model.

As shown in Fig. 7, the untrained model produces relatively fragmented assignments, with individual slots distributed across multiple foreground and background regions. After training, the assignments become substantially more spatially coherent: procedure-relevant foreground regions are grouped into more concentrated slot supports, while redundant mixing across unrelated image regions is reduced. Importantly, this organization emerges without object-level segmentation supervision, since OPSR is trained only to reconstruct frozen visual features while regularizing the resulting slot representation.

The trained slots also exhibit greater temporal persistence. Across Frames 0, 4, and 8, corresponding slot identities remain comparatively stable even as the operator's pose, object locations, and local interactions change over time. The spatial support of a slot can therefore evolve with the ongoing procedure while preserving its identity across temporally separated observations. This behavior is consistent with the recurrent initialization of each frame from the preceding slots and the temporal consistency objective used during OPSR training.

These properties are important for the subsequent BNDM stage, which models procedural dynamics over slot-indexed object trajectories. Stable slot identities allow changes in $z_{t,k}$ across time to be interpreted as the evolution of a persistent underlying entity rather than arbitrary reassignment between slots. The visualization therefore provides qualitative evidence that OPSR learns a spatially coherent and temporally aligned decomposition suitable for forward anticipation, reverse-time modeling, and object-wise discrepancy reasoning. We emphasize that this analysis evaluates the structure and temporal consistency of the learned representation rather than semantic segmentation accuracy, as OPSR is not trained with semantic mask annotations.

### F.4 Further Analysis of Ablation Study

We further analyze what discrepancy evidence contributes to procedural assessment and how its representation affects different assessment objectives. Figure 8 reorganizes the ablation results to expose two patterns that are less apparent from the aggregate results alone.

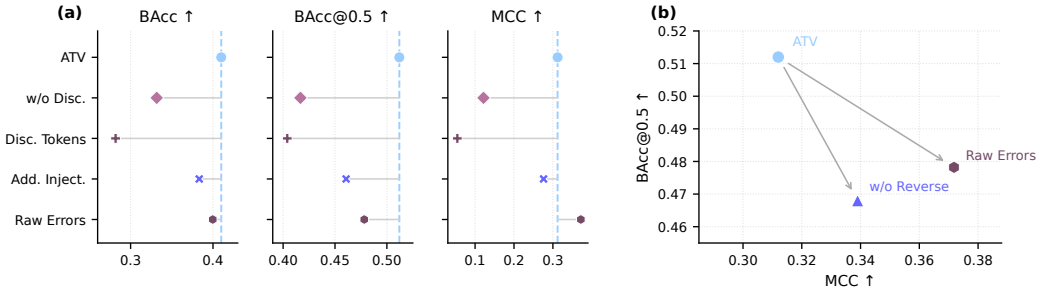



Figure 8: Further analysis of discrepancy evidence. (a) Comparison of discrepancy representations and integration strategies. (b) Removing reverse-time modeling and using raw prediction errors improve binary discrimination while degrading fine-grained ordinal grading.

**Discrepancy evidence is useful, but its integration matters.** As shown in Fig. 8 (a), removing discrepancy conditioning reduces BAcc from 0.410 to 0.332 and BAcc@0.5 from 0.512 to 0.417, despite retaining the observed and normative state representations. Additive injection recovers a substantial portion of this loss, reaching 0.383 BAcc and 0.461 BAcc@0.5, indicating that observed–normative discrepancy provides useful assessment evidence beyond the state representations themselves. However, treating discrepancy as explicit tokens performs even worse than removing discrepancy conditioning, with BAcc dropping to 0.282. These results suggest that discrepancy should not simply be exposed as an additional input stream; how this relational evidence is integrated into the corresponding observation representation is important for effective assessment.

**Binary deviation detection and fine-grained grading favor different evidence.** Raw prediction errors exhibit a distinct performance profile. They achieve a higher MCC than ATV (0.372 vs. 0.312), while reducing BAcc@0.5 from 0.512 to 0.478. A similar trade-off appears independently when the reverse branch is removed: MCC increases to 0.339, whereas BAcc@0.5 decreases to 0.468. Figure 8 (b) makes this common direction explicit. The recurrence of this pattern across two different ablations suggests that evidence sufficient for distinguishing broadly correct from incorrect executions is not necessarily sufficient for estimating the degree of procedural deviation.

**Structured discrepancy primarily benefits fine-grained assessment.** Taken together, the results support a distinction between detecting, characterizing, and evaluating procedural deviations. Raw prediction errors retain strong cues for whether an execution departs from expectation, whereas bidirectional structured discrepancy better preserves information about how the observed transition differs from normative evolution. Rubric-conditioned reasoning can then interpret this structured deviation evidence with respect to the target scoring criterion. This provides an empirical explanation for why ATV benefits from structured discrepancy reasoning rather than reducing procedural verification to generic prediction-error detection.

All component-level variants retain the same scoring backbone, hierarchical ordinal objective, optimization protocol, and checkpoint-selection criterion, differing only in the corresponding normative-dynamics or discrepancy component.

## G Experiments Details

All baseline models are evaluated on the same train/validation/test split of 11,048/1,566/3,161 samples and are trained with AdamW in BF16 precision using 8 Nvidia A800 GPUs. For all MLLM baselines, we retain the model-specific visual preprocessing and input formatting provided by the corresponding implementations, and fully fine-tune each model with DeepSpeed. All MLLM baselines use the same rubric-conditioned scoring prompt (as shown in Fig 9) to ensure a consistent evaluation protocol. For each baseline, we follow the model-specific fine-tuning recipe recommended by its official implementation, including the training duration, learning-rate schedule, and optimization strategy whenever available.

**Open-source Video-MLLMs.** We evaluate LLaVA-OneVision-7B, LLaVA-Video-7B, VideoLLaMA3-7B, Qwen3-VL-8B, Qwen3-VL-30B-A3B, and Molmo-2-8B. Videos are uni-

**Step-wise Clinical Scoring Prompt**

```
## Your Task
Evaluate the clinical performance strictly according to the following rubric, using only performance evidence visible in the provided video.
{rubric_text}
Allowed score labels are: {A=..., B=..., ...}.
## Mandatory Endpoint Veto Rule
MANDATORY ENDPOINT VETO RULE: Neither endpoint is a default. Output the highest label only after directly verifying every individual rubric requirement,
including quality, sequence, completeness, and safety where applicable. A generic impression, a plausible action, partial completion, or any unclear
requirement automatically vetoes the highest label: choose the closest internal (non-endpoint) label. Output the lowest label only after directly verifying
that the required performance is absent, fundamentally incorrect, or materially unsafe. Uncertainty, missing detail, mixed evidence, or an incomplete-looking
attempt does not justify the lowest label; choose the closest internal label instead. Before emitting either endpoint, silently name the specific evidence
that passes its gate; if you cannot name it, the endpoint is forbidden.
## Global Anti-Collapse Rule
GLOBAL ANTI-COLLAPSE RULE: This is one item in a large batch. It is forbidden to use either endpoint as a universal default across cases or rubrics: do not
make the whole batch all maximum labels and do not make the whole batch all minimum labels. Independently assess this video's evidence; an endpoint is
permitted only after its specific evidence gate is met, and an internal label is required whenever its evidence is stronger than either endpoint gate.
## Evaluation Procedure
Before answering, silently perform two checks for this one rubric:
  (1) identify the concrete visual evidence for each required component;
  (2) compare its quality, completeness, sequence, and safety against the rubric, then map that evidence to the nearest allowed score.
Score this criterion independently; do not infer quality from the candidate, the setting, another action, or an overall impression. A visible attempt or a
completed-looking action is not automatically full credit: reserve the maximum label only for clear evidence that every stated requirement is satisfied
correctly. Use the minimum only when the required performance is absent. If evidence is partial, incomplete, unsafe, incorrectly sequenced, or mixed, select
the corresponding intermediate label. Never default to either endpoint.
## Output Format
Return exactly one allowed label and no other text.
```

Figure 9: Rubric-conditioned scoring prompt used for all MLLM baselines.

formly sampled to 32 frames. Training uses a per-GPU batch size of 1 and gradient accumulation of 2, with model-specific learning rates and scheduler settings following the corresponding implementations. Models are optimized with causal-LM loss over the target score label. We use a single seed 42, with inference performed on a single GPU.

**Open-source Omni-MLLMs.** We evaluate Qwen2.5-Omni-7B, Qwen3-Omni-30B-A3B and Baichuan-Omni-1.5-7B using only the visual stream. We use each model's default frame-sampling strategy. Qwen2.5-Omni and Qwen3-Omni use a learning rate of $2 \times 10^{-5}$, Baichuan-Omni use $1 \times 10^{-5}$s.

**Medical Video-MLLMs.** We evaluate uAI-NEXUS-MedVLM-1.0a-7B-RL, UniMed-VL-14B, and SurgVidLM Stage-1-7B. uAI-NEXUS and SurgVidLM sample videos at 2 FPS with at most 64 frames, whereas UniMed-VL uniformly samples 8 temporally ordered RGB frames. Training uses DeepSpeed ZeRO-3. uAI-NEXUS uses learning rates of $5 \times 10^{-7}$ and $1 \times 10^{-6}$ for the language and visual components, respectively, while SurgVidLM and UniMed-VL use $1 \times 10^{-5}$.

**Latent Predictive Models.** We evaluate V-JEPA 2 ViT-g/16 (1B), V-JEPA 2.1 ViT-G/16 (2B), and VL-JEPA [7]. Since latent predictive models do not natively produce rubric-constrained clinical scores, we retain their pretrained visual backbones and attach a lightweight ordinal scoring head. For V-JEPA 2 and V-JEPA 2.1, video tokens are projected into a shared latent space and aggregated by cross-attention with a learned task-step query, while the prediction combines a task-step prior with a video-conditioned residual. Open-VLJEPA retains its released V-JEPA2 encoder and Llama predictor initialized from the Stage-B checkpoint, with the predictor jointly processing rubric text and visual tokens before ordinal readout.

All three models are adapted in two stages: a frozen-probe stage that trains only the ordinal head, followed by full adaptation of the head and pretrained backbone; for Open-VLJEPA, the predictor is also updated. V-JEPA 2 and V-JEPA 2.1 sample 64 frames, resize and center-crop them to $384 \times 384$, and apply ImageNet normalization, whereas Open-VLJEPA uses 8 frames at $256 \times 256$. The ordinal head predicts 13 score levels from 0 to 6 at 0.5-point intervals and is optimized with masked cross-entropy plus $0.25$ times normalized SmoothL1 loss.

[7] We use the open-source Open-VLJEPA implementation (1.6B), as access to the official VL-JEPA model weights was not granted at the time of our experiments.


[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。<br>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 10: Screenshot of the IRB approval document.

## H Extended Ethical Considerations

ClinSkill-V is built from recordings of medical students performing clinical skills during authentic examinations conducted at our collaborating hospitals. The collection, annotation, and use of these recordings for research were conducted under protocols approved by the Institutional Review Board (IRB) of the collaborating institution. The videos were obtained in standardized educational and examination environments under institutional supervision and were restricted to educational and research applications. Throughout data processing, safeguards were adopted to protect participant privacy, remove identifying information where appropriate, limit data access, and prevent unauthorized use. We further retained only the information required for rubric-based analysis of clinical skill performance. The corresponding IRB approval document is shown in Fig 10.

The dataset may facilitate future development of automated methods for clinical skill assessment, but such applications also raise important ethical considerations. In particular, automated assessment systems may inherit biases from expert annotations or local evaluation standards, and their predictions may be overinterpreted beyond their validated scope. Accordingly, potential downstream applications should incorporate continued expert supervision, clearly communicate the criteria underlying assessment, and use model-generated scores or feedback only as supporting evidence rather than as a substitute for professional judgment.