AI-Assisted Competency Assessment from Egocentric Video in Simulation-Based Nursing Education

Supplementary Material

A. Simulation Scenario Summary

Scenario Context and Setting The simulation scenario and debrief were created by a nursing teaching instructor and have been used in nursing school classroom settings. The simulation scenario takes place in a high-fidelity pediatric emergency room bay. A standardized pediatric manikin representing a toddler and a faculty facilitator acting as the patient's caregiver are present at the bedside. The scenario is designed to evaluate pediatric assessment, weight-based medication administration, and caregiver communication competencies.

Anonymized Patient Profile The simulated patient is a 16-month-old toddler (9.6 kg, 76 cm) presenting with a primary diagnosis of croup (laryngotracheobronchitis). The caregiver reports a 3-day history of upper respiratory infection symptoms, with sudden overnight onset of a barking cough, hoarse voice, and inspiratory stridor. On arrival, the patient is placed on continuous pulse oximetry, heart rate, and respiratory rate monitoring, and maintained on humidified oxygen at 1 LPM via pediatric face mask.

Simulation Learning Objectives The scenario targets four core nursing competencies: (1) performing a focused pediatric assessment while maintaining age-appropriate patient safety; (2) recognizing pediatric fever and calculating accurate weight-based dosages for oral antipyretic medications; (3) preparing and administering pediatric oral suspensions safely; and (4) providing clear, developmentally appropriate education to caregivers regarding at-home medication administration.

Scenario Progression and Key Interventions The simulation unfolds across three phases. In **Phase 1**, the student initiates care by performing hand hygiene, verifying patient identification using two identifiers, and introducing themselves to the caregiver. Initial vitals reflect tachypnea and tachycardia consistent with the patient's respiratory distress. In **Phase 2**, the student performs a focused respiratory assessment, noting mild expiratory wheezing and intermittent barking cough. A bedside temperature check reveals a fever of $102.6°$F, prompting the student to review physician orders and perform a weight-based medication calculation for oral acetaminophen suspension (160 mg$/5$ mL):

$$\begin{aligned} 15 \text{ mg/kg} \times 9.6 \text{ kg} &= 144 \text{ mg}, \\ 144 \text{ mg} \times \frac{5 \text{ mL}}{160 \text{ mg}} &= 4.5 \text{ mL}. \end{aligned} \tag{3}$$

In **Phase 3**, after preparing the medication in an amber oral dosing syringe, the student engages in targeted caregiver education. Key instructional points include advising against using household spoons for measuring, demonstrating correct syringe administration technique to prevent choking, and establishing safe guidelines for dosing frequency at home.

B. Instructor Competency Rubric

The instructor rubric is an adapted version of the Creighton Competency Evaluation Instrument (C-CEI) [27], which maps 23 expected behaviors to broader concepts of competency (e.g., clinical judgment, patient safety, communication). Each item is rated on a 1–5 scale (Poor to Exceptional). Because our study uses egocentric video without audio, only the **11 video-observable items** (highlighted in Tab. 5) contribute to each student's competency percentage. The remaining 12 items require verbal or cognitive assessment not accessible from visual data alone.

Rationale for the video-observable subset. Items requiring verbal content (e.g., "Communicates effectively with team," "Provides evidence-based rationale") or internal cognitive processes (e.g., "Reflects on clinical experience") cannot be assessed from silent egocentric video. The 11 retained items correspond to physical actions and procedural behaviors that produce visible evidence in the video stream: checking wristbands, performing hand hygiene, documenting on screens, measuring vital signs, administering medications, and following safety protocols. This principled subset ensures that the competency score reflects only expected behaviors that our vision-based system could plausibly detect. Note that in the per-item association analysis (Tab. 4), we report associations for all 23 items to explore whether vision-based features carry any indirect signal for non-observable behaviors.

C. Action Annotation Rubric

The following is from our annotation codebook, used by trained coders to produce ground-truth action annotations, and is inspired by [23]. Actions represent discrete, observ-

Table 5. Full 23-item C-CEI rubric. Each item specifies an expected behavior; highlighted rows (✓) are the 11 video-observable items used for competency scoring.

#	Item Description	Video?
1	Obtains pertinent data	✓
2	Performs follow-up assessments as needed	✓
3	Assesses the environment	×
4	Communicates effectively with team	×
5	Communicates effectively with patient	×
6	Documents clearly, concisely, and accurately	✓
7	Responds to abnormal findings appropriately	×
8	Promotes professionalism	×
9	Interprets vital signs	✓
10	Interprets laboratory results	×
11	Interprets subjective/objective data	×
12	Prioritizes appropriately	✓
13	Performs evidence-based interventions	✓
14	Provides evidence-based rationale for interventions	×
15	Evaluates evidence-based interventions and outcomes	×
16	Reflects on clinical experience	×
17	Delegates appropriately	×
18	Uses patient identifiers	✓
19	Utilizes standardized practices and precautions	✓
20	Administers medications safely	✓
21	Manages technology and equipment	✓
22	Performs procedures correctly	✓
23	Reflects on potential hazards and errors	×

able physical behaviors; verbal introductions are captured separately by the Communication layer.

General coding rules.

1. Code only what is directly observable; do not infer intent.
2. When in doubt, leave the segment unlabeled.
3. Annotations must not overlap within the Action layer.
4. Start when the action begins (first observable movement); end when it concludes (hands leave the object, body repositions away).
5. Brief interruptions ($<$2 s): code as one continuous segment.

Action definitions. Tab. 6 lists the K=16 fine-grained clinical action classes. Frames that do not correspond to any of these classes (e.g., walking, adjusting equipment, idle periods between clinical actions) are left unannotated and treated as the background class $a_\varnothing$, yielding K+1=17 labels in total for recognition.

Disambiguation guidelines. Several action pairs are visually similar and require explicit decision rules:

Lung Sounds (#8) vs. Apical Pulse (#9): Stethoscope on the back or moved across multiple chest positions is coded as #8. Stethoscope held at the left chest apex in one position for $\geq$15 s is coded as #9. If placement is unclear, default to #8 and flag for review.

Calculator (#13) vs. Phone (#14): Tapping numbers on a calculator app or physical calculator is #13. Scrolling, reading, or swiping on a phone (non-calculator) is #14.

Patient History Screen (#4) vs. Vital Signs Screen (#6): If the screen shows waveforms or real-time numeric readings, code as #6. If it shows text-based records, history, or medication orders, code as #4. Pressing a button on the vitals monitor to initiate a BP measurement is coded as #11.

Table 6. The K=16 clinical action classes used for temporal annotation and few-shot recognition, with brief operational definitions. An additional background class $a_\varnothing$ (not shown) captures all non-clinical frames, yielding 17 labels total.

ID	Action Class	Definition
1	Perform Hand Hygiene	Uses hand sanitizer or washes hands at sink
2	Put on Gloves	Retrieves and dons disposable gloves
3	Check Patient Wristband	Visually inspects or scans patient wristband
4	Check Patient History Screen	Reads electronic health record on screen
5	Examine Med Bottle	Picks up and reads medication label
6	Review Vital Signs Screen	Reads the vital signs monitor (HR, BP, SpO_2)
7	Assess Vital Signs (Palpate Wrist)	Manually palpates radial pulse
8	Auscultate Lung Sounds	Places stethoscope on chest/back for breath sounds
9	Measure Apical Pulse	Places stethoscope at heart apex, held $\geq$15 s
10	Measure Temperature	Uses thermometer (oral, tympanic, temporal)
11	Measure Blood Pressure	Initiates BP reading via monitor or manual cuff
12	Writing	Pen-to-paper: notes, calculations, forms
13	Use Calculator	Computes dosage on physical or phone calculator
14	Check Phone	Interacts with phone for non-calculator purposes
15	Prepare Medication	Draws syringe, crushes tablet, mixes solution
16	Apply Medication to Patient	Administers medication: oral, IV, injection, topical

D. Process Model Details

The five key structural differences between higher- and lower-performing students process models (Fig. 4) are elaborated below.

Screen self-loop. Lower-performing students exhibit a higher self-loop on the Screen action (48% vs. 41%), spending proportionally more time returning to the bedside monitor without transitioning to other clinical actions. Higher performing students distribute transitions away from Screen more evenly across Examination, Writing, and Calculator, reflecting a more fluid workflow. Screen actions are visually static and uniform, making them easy for the classifier and inflating MOF for the lower group.

Medication pathway. Higher-performing students show a strong direct Prep Med $\rightarrow$ Apply Med transition (46%), indicating a coherent prepare-then-administer sequence. Lower-performing students lack this link; instead, Prep Med routes back to Screen (38

Examination frequency. Higher performers engage in more Examination actions (36 vs. 29), while lower-performing students produce more Writing and Screen actions (42 and 79 vs. 37 and 74). Physical examination (lung sounds, blood pressure, palpation) involves diverse movements inherently harder to classify, consistent with the observed negative trend between accuracy and competency.

Transition irregularity. The lower-performing students model contains more group-unique (red) transitions, indi-

cating irregular workflow paths. Higher performers follow a more protocol-consistent progression with fewer idiosyncratic transitions.

Hygiene compliance. Hygiene actions connect to Screen with 76% probability in higher performing students, suggesting consistent hand hygiene before engaging with the patient monitor. This transition is less prominent in lower-performing students, pointing to less consistent infection control practices.

These process model comparisons offer actionable insight for clinical educators: the transition graphs visualize where each student's workflow diverges from the expected clinical pathway, enabling targeted remediation of specific procedural gaps.

E. Annotation Confound Analysis

A potential concern is that annotation artifacts drive the negative trend between classification accuracy and competency, since higher-competency sessions have lower annotation coverage (40% vs. 50%). We perform partial association analysis, controlling for six potential confounds (Tab. 7). If any drove the observed pattern, controlling for it would weaken or eliminate the effect.

Table 7. Robustness analysis. *Partial* ρ: Spearman association between MOF and competency after controlling for each variable. *Var* $\leftrightarrow$ *MOF*: bivariate association between each variable and MOF. The MOF–competency association persists across all controls and *strengthens* when controlling for annotation coverage (bolded). No control variable independently predicts MOF (all $p > 0.46$).

Control Variable	Partial ρ	p	Var $\leftrightarrow$ MOF ρ	p
None (baseline)	-0.439	0.041	—	—
Annotation coverage	$\mathbf{-0.546}$	**0.009**	-0.032	0.887
# GT action segments	-0.427	0.047	$+0.115$	0.611
# Unique GT action types	-0.438	0.041	$+0.027$	0.907
Avg segment duration	-0.437	0.042	-0.165	0.462
Video duration	-0.454	0.034	$+0.074$	0.744
# All annotations	-0.427	0.047	$+0.115$	0.611

The pattern persists across all controls. When controlling for annotation coverage, the effect *strengthens* (ρ: $-0.439 \rightarrow -0.546$), and no control variable independently predicts MOF (all $p > 0.46$), confirming that the negative trend reflects workflow complexity rather than annotation density.

F. Inter-Rater Reliability

A second rater independently annotated 3 stratified videos (low / median / high competency) to assess annotation reliability. Agreement was measured using frame-level Cohen's κ at 1 Hz resolution. To avoid inflation from unannotated frames, κ was computed only over frames where at least one rater placed a label.

Mean $\kappa = 0.708 \pm 0.199$ (substantial agreement; [15]). As a secondary metric, mean per-class IoU $= 0.697 \pm 0.143$, and both raters identified identical action type sets in all 3 videos (Jaccard $= 1.0$). Disagreements were predominantly in segment boundary placement, particularly action endpoints (mean $|\Delta| = 3.4$ s), rather than action identification or ordering. This pattern indicates that raters agree on *which* actions occur and *in what order*, with variability confined to the precise temporal boundaries, consistent with the known difficulty of endpoint annotation in temporal action segmentation [4].