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

Anonymous CVPR submission

Paper ID 18

Abstract

Assessing learner competency in clinical simulation requires expert observation that is time-intensive, difficult to scale, and subject to inter-rater variability. While multimodal learning analytics has emerged as a promising direction, the contributions of individual modalities remain underexplored. We investigate what vision perspective alone can reveal by proposing a three-stage framework that (1) extracts action timelines from egocentric nursing simulation video using frozen visual encoders and few-shot learning, (2) derives sequence-level features and per-session recognition metrics, and (3) relates these to instructor-rated competency. Across 22 densely annotated sessions (3.8 hours, 493 actions), a frozen DINOv2 backbone with HMM Viterbi decoding achieves 57.4% MOF in leave-one-out 1-shot recognition. Surprisingly, we observe a negative trend between recognition accuracy and competency ($\rho = -0.524$, $p = 0.012$ for mIoU), robust to six confound controls: more competent students produce diverse, harder-to-classify workflows, while simple sequence features show no such relationship. Per-item analysis identifies patient safety protocols and team communication as the expected behaviors most reflected in this pattern, and process model comparisons reveal that high-competency students exhibit more protocol-consistent action transitions. These findings suggest that recognition accuracy may complement predicted action timelines as a pedagogically informative signal in automated competency assessment.

1. Introduction

Across education and workforce training, a central goal is to determine whether learners have developed the knowledge, skills, and judgment needed to perform effectively in practice, a quality broadly termed *competency* [32, 46]. In domains defined by skilled physical performance, competency assessment requires expert observation of context-dependent behaviors that unfold over time [10]. The consequences of undetected gaps are especially acute in clinical education, where medication administration errors remain among the most common preventable adverse events, often rooted in procedural lapses missed during training [17]. Simulation-based learning addresses this by letting students practice clinical skills without risk to real patients [17, 22], but competency encompasses not just executing procedures correctly but doing so in an appropriate sequence, with complete safety checks and fluid task transitions [10, 26]. Instructors assess each session using standardized instruments such as the Creighton Competency Evaluation Instrument (C-CEI) [44] that map observable behaviors to competency constructs (e.g., clinical judgment, patient safety) [1, 26]. This model faces two structural constraints: expert observation cannot scale with growing cohorts [3], and inter-rater reliability remains only moderate to substantial even among trained faculty [20, 23]. Multimodal Learning Analytics (MMLA) [6, 9] has emerged as a promising response, integrating video, audio, physiological signals, and interaction logs to study learning processes at scale [2, 37]. However, in the context of nursing simulation competency, the contributions of individual modalities remain underexplored. As a first step, we focus on vision [43] to investigate what this modality can reveal.

Video captures what learners do, how they move, and what objects they interact with, all without requiring instrumented environments. Recent advances in first-person video understanding have made egocentric recordings especially compelling [15]. Head-mounted cameras provide an unobstructed, hands-proximal view of what a student attends to and acts upon, and when paired with gaze sensing [7, 21], can reveal attentional patterns linked to errors in skilled activities [27]. As shown in Fig. 1, this perspective captures the full behaviors of a clinical encounter, from how students hold a medication bottle to which device they use for dosage calculation, preserving precisely the signatures that distinguish novice from expert workflows.

Medication administration is an ideal proving ground because competence is inherently sequential: correct actions in the wrong order, or with safety steps omitted, consti-

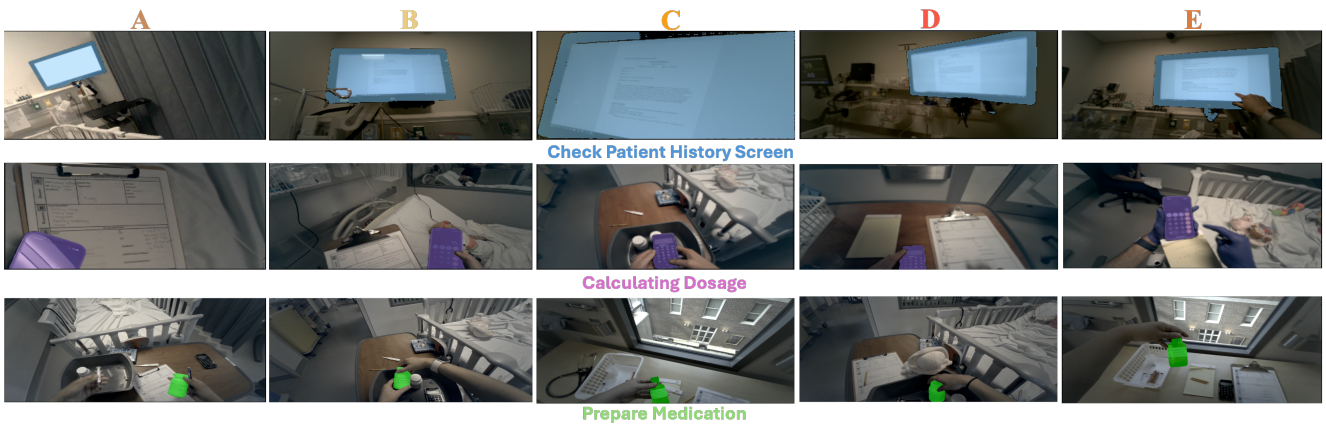


Figure 1. Example images of checking the patient screen, calculating dosage, and preparing medication from simulation videos of five nursing students. Students A, B, and E use phones for dosage calculation, whereas students C and D use handheld calculators. During medication preparation, students C and E use a dark brown medicine bottle, while students A, B, and D each use a different bottle and hold it differently. More details of the simulation procedure are in App. A.

tute a clinical error [22]. The annotation codebook used in this study was developed around the medication administration workflow, drawing on established nursing competency measurement tools [36], and captures fine-grained actions such as dosage calculation. The C-CEI rubric maps expected behaviors to broader competency constructs across the full clinical encounter; vision-based analysis can speak directly to the former, while the latter requires instructor judgment. Applying egocentric video understanding to this setting, however, introduces domain-specific challenges: clinical action vocabularies are absent from standard benchmarks [5, 8], cohorts are small due to privacy constraints, and models pre-trained on real bodies face a domain gap with simulation mannequins. While surgical education has demonstrated that AI can recognize operative phases and classify skill levels from laparoscopic recordings [24], that work assumes data-rich, fixed-camera environments. Nursing simulation differs in that the recording is egocentric, cohorts are governed by IRB constraints, and competency is holistic rather than tied to a single technical procedure. Yet these difficulties may not be purely noise. Research in surgical skill assessment has found that automated classifiers perform worse on higher-skilled practitioners [41], and that temporal patterns of action execution capture skill level better than outcome measures alone [47]. This raises the possibility that recognition accuracy itself carries a pedagogical signal, with lower accuracy reflecting the diverse workflows of more competent students.

This gap motivates the present work to investigate whether egocentric video, analyzed through frozen visual encoders and few-shot learning, can serve as the basis for automated competency assessment in nursing simulation. Specifically, we ask whether action timelines can be reliably extracted from first-person clinical video under extreme low-data constraints; whether the resulting sequence features and recognition difficulty relate to instructor-rated competency; and whether temporal action patterns differ between high- and low-performing students. To answer these questions, we propose a three-stage framework that extracts action timelines, analyzes their sequential structure, and relates sequence features and recognition accuracy to competency scores across 22 densely annotated sessions.

Research Questions.

RQ1. To what extent can few-shot action recognition identify clinical actions in egocentric nursing simulation video? (§5.1)

RQ2. To what extent do automatically extracted action sequences and recognition difficulty relate to instructor-rated competency, and which expected behaviors on the C-CEI are most reflected in vision-based action analysis? (§5.2)

RQ3. What temporal action patterns distinguish high- from low-performing students? (§5.3)

Contributions. This work advances competency assessment from egocentric video through three contributions:

1. **Few-shot clinical action recognition.** We show that frozen DINOv2 features with HMM Viterbi decoding achieve 57.4% MOF in leave-one-out 1-shot action recognition of egocentric nursing simulation video, establishing feasibility under extremely low-data conditions without any fine-tuning.
2. **Classification difficulty to competency.** We observe a negative trend between recognition accuracy and instructor-rated competency ($\rho = -0.524$, $p = 0.012$ for mIoU), robust to six confound controls. Per-item analysis identifies expected behaviors related to patient

safety protocols and team communication as the C-CEI items most reflected in this pattern.

3. **Temporal workflow analysis.** Process model comparisons from ground-truth action sequences show that high-competency students exhibit more diverse, protocol-consistent action transitions, delineating the boundaries of unimodal video-based assessment.

2. Related Work

Computer vision for clinical skill assessment. Deep learning has been applied extensively to surgical workflow recognition and skill evaluation from operative video, including tool detection and phase recognition [45], direct skill classification from video [13], fine-grained action triplet recognition [29], and scalable objective assessment of technical skill [14, 16]. These studies highlight the promise of video-based clinical assessment, but most are developed in data-rich surgical settings with fixed cameras and relatively controlled workflows. In contrast, our setting is egocentric and small-scale, and the goal is to assess holistic nursing competency rather than isolated technical skill. **Computer vision for education and learning analytics.** MMLA integrates video, audio, physiological signals, and interaction logs to study learning processes [30]. Within this paradigm, vision has been used to detect learning-relevant affective states [4], align neural attention with human gaze [42], analyze embodied classroom learning [11, 12], and model student interaction sequences [40]; a recent review surveys multimodal methods across adult training environments, including nursing simulation [38]. We focus on a single modality (egocentric video) to establish what vision alone can reveal about clinical competency. **Few-shot and temporal action recognition.** Prototype networks enable classification with minimal labeled examples. Temporal action segmentation has advanced rapidly on standard benchmarks [8], and large-scale egocentric datasets [15] together with self-supervised encoders such as DINOv2 [31] provide strong frozen representations. We combine prototype matching with HMM Viterbi decoding [33] for temporal segmentation under extreme low-data clinical conditions.

3. Problem Formulation

Let $\mathcal{V} = \{V_1, \ldots, V_N\}$ denote a set of N egocentric video sessions, where each session $V_i = (f_i^{(1)}, f_i^{(2)}, \ldots, f_i^{(T_i)})$ consists of T_i ordered frames, where $i \in \{1, \ldots, N\}$ indexes the session and $t \in \{1, \ldots, T_i\}$ indexes the frame. Each video is associated with an instructor-assigned competency score vector $\mathbf{c}_i \in \mathbb{R}^{23}$ across 23 expected behaviors on the C-CEI rubric (App. B), of which 11 correspond to video-observable actions; the mean of these 11 items yields the overall competency percentage used for association analyses. Not all items are rated for every session, so some entries of $\mathbf{c}_i$ are missing.

Stage 1: Action Recognition. Given a clinically grounded action taxonomy $\mathcal{A} = \{a_1, \ldots, a_K, a_\varnothing\}$ comprising $K{=}16$ clinical action classes and one background class $a_\varnothing$ (17 labels total), the goal is to assign each frame a label $y_i^{(t)} \in \mathcal{A}$, producing a frame-level prediction $\hat{\mathbf{y}}_i = (\hat{y}_i^{(1)}, \ldots, \hat{y}_i^{(t)})$. A frozen encoder ϕ extracts per-frame features $\mathbf{z}_i^{(t)} = \phi\big(f_i^{(t)}\big)$, which are matched against class prototypes computed from a support set $\mathcal{S}$ of labeled exemplars sampled from held-out sessions:

$$\hat{\mathbf{y}}_i = \mathrm{Decode}\big(\mathrm{sim}(\phi(V_i),\ \mathcal{P}(\mathcal{S}))\big), \tag{1}$$

where $\mathcal{P}(\mathcal{S})$ computes class prototypes from the support set [39] (Sec. 4.1) and Decode applies HMM Viterbi decoding [33] to enforce temporally coherent label sequences. We evaluate $\hat{\mathbf{y}}_i$ against ground-truth annotations $\mathbf{y}_i^*$ using frame-level accuracy (MOF), mean intersection-over-union (mIoU), and segmental F1 (RQ1).

Stage 2: Sequence Analysis. From the predicted frame-level labels $\hat{\mathbf{y}}_i$, we collapse contiguous same-label frames into an ordered action sequence $\mathbf{s}_i = \big((c_i^{(1)}, d_i^{(1)}), \ldots, (c_i^{(L_i)}, d_i^{(L_i)})\big)$ of L_i segments ($l \in \{1, \ldots, L_i\}$), where $c_i^{(l)} \in \mathcal{A} \setminus \{a_\varnothing\}$ is the action label and $d_i^{(l)}$ is the segment duration in frames. From this sequence we derive two families of features used in subsequent analyses: (1) action transition frequencies, which capture the pairwise flow between clinical actions, and (2) per-video recognition metrics (MOF, mIoU, F1), which summarize how well the classifier fits each session.

Stage 3: Competency Analysis. Given the small sample size ($N = 22$) and the pedagogical requirement for transparent feedback, we map sequence features to competency scores using Spearman rank association. To disentangle action detection errors from the intrinsic limits of vision-based assessment, we evaluate under both oracle (features from ground-truth $\mathbf{y}_i^*$) and predicted (features from $\hat{\mathbf{y}}_i$) conditions. Per-item analysis examines which expected behaviors are captured by action sequences alone, and comparison of ground-truth action transition graphs across performance groups identifies discriminative temporal patterns (RQ2, RQ3).

4. Method

We collect egocentric video from 22 nursing students, each performing a single standardized pediatric simulation on high-fidelity mannequins. Each session captures one student's complete first-person view of the clinical encounter, recorded via egocentric glasses at 25 FPS. Sessions range from 4 to 24 minutes (mean 10.5 min, total 3.8 hours).

Each session is annotated across three temporal layers by

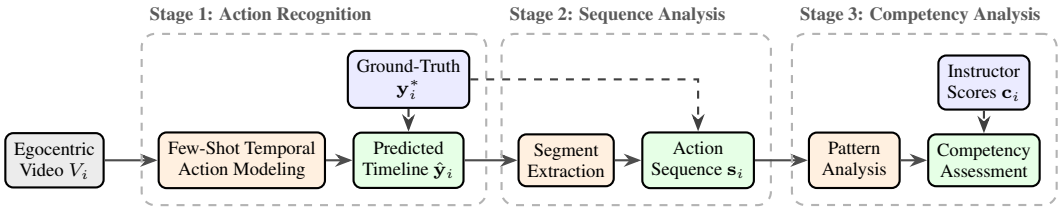


Figure 2. Overview of the proposed three-stage framework. Gray boxes denote inputs, orange boxes denote processing modules, green boxes denote outputs, and blue boxes denote supervision or reference signals. Solid arrows indicate the forward inference flow, while dashed arrows indicate supervision or oracle guidance. The three stages perform action timeline prediction, action sequence construction, and competency assessment, respectively.

a trained coder using the NOVA annotation system: (1) Behaviors (3 classes: Introduction, Assessment, Administration), (2) Actions ($K = 16$ fine-grained clinical classes plus one background class $a_\varnothing$ for unannotated frames, yielding 17 labels total; see App. C), and (3) Communication (Patient, Family, Provider). The Action layer, containing 493 annotated clinical segments, serves as the primary target for few-shot recognition. Each session is independently rated by an expert instructor across 23 expected behaviors using the C-CEI rubric (App. B), of which the 11 video-observable items yield the overall competency percentage used throughout this study. Inter-rater reliability was assessed on 3 stratified videos (low/median/high competency) independently annotated by a second rater, yielding substantial agreement (mean Cohen's $\kappa = 0.708$; App. F).

4.1. Stage 1: Action Recognition

Feature extraction. We extract frame-level features from each video using a frozen backbone encoder. For each frame $f_i^{(t)}$, we obtain a feature vector $\mathbf{z}_i^{(t)} = \phi\big(f_i^{(t)}\big) \in \mathbb{R}^D$, followed by L2 normalization: $\mathbf{z}_i^{(t)} \leftarrow \mathbf{z}_i^{(t)} / \|\mathbf{z}_i^{(t)}\|$. We evaluate three backbones: (1) ResNet-50 [18] (ImageNet-supervised, $D = 2048$), (2) DINOv2 ViT-B/14 [31] (ImageNet self-supervised, $D = 768$), and (3) CLIP ViT-B/16 [34] (vision-language contrastive, $D = 512$). All backbones are frozen with no fine-tuning.

Prototype computation. In the cross-sample (leave-one-out) setting, we construct class prototypes from the $N-1$ support sessions following the prototypical network paradigm [39]. For each support session V_j and each class k present in that session, we randomly sample n labeled frames and compute a per-session centroid $\boldsymbol{\mu}_{k,j}$, which is then L2-normalized. Because not every session contains every action class, the global prototype for class k is obtained by averaging the normalized centroids only over sessions that contain that class: $\mathbf{p}_k = \frac{1}{|\mathcal{J}_k|} \sum_{j \in \mathcal{J}_k} \frac{\boldsymbol{\mu}_{k,j}}{\|\boldsymbol{\mu}_{k,j}\|}$, where $\mathcal{J}_k$ is the set of sessions containing class k. The per-session normalization ensures each session contributes a unit-direction vector, preventing sessions whose sampled frames are more self-consistent from dominating the prototype direction. The aggregated prototype is then L2-normalized again to ensure it lies on the unit sphere, since the mean of unit vectors is not itself unit-length in general. As an alternative, we also evaluate a clustered strategy in which all support frames for each class are pooled and partitioned into k=3 sub-centroids via k-means; each query frame is then assigned to the class of its nearest sub-centroid.

Classification. Each query frame is scored against all prototypes (including the background class $a_\varnothing$) via cosine similarity. Rather than committing to a hard per-frame label at this stage, the continuous similarity scores are passed directly to the temporal smoothing step below, which jointly optimizes over the entire sequence.

Temporal smoothing. We apply *HMM Viterbi decoding* [33] to enforce temporally coherent label sequences. The transition matrix $\mathbf{A}$ is learned from support session label sequences with Laplace smoothing, prior probabilities $\boldsymbol{\pi}$ are estimated from action start frequencies, and emission log-probabilities are computed as temperature-scaled ($\tau = 5$, selected via grid search on held-out folds; values in the range 5–10 are standard for cosine-prototype matching in few-shot settings [39]) log-softmax of cosine similarities:

$$\log P(\mathbf{z}_i^{(t)} \mid a_k) = \tau \cdot \cos\big(\mathbf{z}_i^{(t)}, \mathbf{p}_k\big) - \log \sum_{k'} \exp\Big(\tau \cdot \cos\big(\mathbf{z}_i^{(t)}, \mathbf{p}_{k'}\big)\Big). \quad (2)$$

The Viterbi algorithm [33] then recovers the optimal label sequence $\hat{\mathbf{y}}_i = (\hat{y}_i^{(1)}, \ldots, \hat{y}_i^{(T_i)})$ by selecting, at each time step, the action label that jointly maximizes the cumulative sum of emission log-probabilities and transition log-probabilities over the entire sequence, thereby enforcing clinically plausible transitions rather than treating each frame independently. This smoothed timeline is the final output of Stage 1 (Sec. 4.1) and the prediction used in all subsequent analyses (Sec. 4.2–4.3).

4.2. Stage 2: Sequence Analysis

From the frame-level predictions $\hat{\mathbf{y}}_i$, we collapse contiguous same-label frames into an ordered sequence of action segments via run-length encoding, discarding segments below a minimum duration threshold and removing background segments. From the resulting clinical action sequence we compute two families of features. First, we tabulate pairwise action transition frequencies, which record how often each action class is followed by every other class within a session; these frequencies form the basis of the process model comparison in Sec. 5.3. Second, we retain the per-video frame-level recognition metrics (MOF, mIoU, F1) computed during Stage 1, which serve as summary measures of how well the classifier fits each session and are used in the competency analysis (Sec. 5.2).

4.3. Stage 3: Competency Analysis

Given the small sample size and the pedagogical need for transparency, we employ Spearman's rank correlation to test the relationship between sequence-level features and video-observable competency scores (11 items). To assess robustness, we compute partial associations controlling for potential confounds (annotation coverage, video duration, segment count). Per-item analysis examines which of the 23 expected behaviors are most reflected in recognition accuracy. Process model analysis (Heuristics Miner) visualizes action transition graphs for high- and low-competency groups to identify differences in clinical workflows.

5. Results

5.1. RQ1: Few-Shot Clinical Action Recognition

We evaluate few-shot action recognition under two settings: *within-sample*, where support and query frames are drawn from the same video, and *cross-sample* (leave-one-out), where the model must generalize to entirely unseen sessions. We report three standard temporal action segmentation metrics: mean-over-frames accuracy (MOF), mean intersection-over-union (mIoU), and segmental F1 score.

5.1.1. Within-Sample Evaluation

In the within-sample setting, for each of the 22 videos, n frames per action class are randomly sampled as support prototypes (where n denotes the shot count); the remaining frames serve as the query set. Tab. 1 reports within-sample performance across five shot counts. Recognition quality improves substantially with more support examples, with the largest gain between 1 and 3 shots. Performance plateaus around 10–15 shots, indicating that even a modest number of labeled exemplars enables reliable within-sample segmentation and that the bottleneck lies in cross-session generalization rather than representation capacity.

5.1.2. Cross-Sample Evaluation

The more challenging and practically relevant setting is cross-sample evaluation, where the model must generalize to entirely unseen sessions with unseen participants. We adopt a leave-one-out protocol across all 22 sessions: in each of 22 folds, one session is held out as the query video, and class prototypes are constructed from the remaining 21 support sessions using prototype computation (Sec. 4.1). We vary the shot count $n \in \{1, 3, 5, 10, 15\}$ frames sampled per class per session to examine how prototype quality scales with the support budget. The query video is classified frame-by-frame via cosine similarity followed by HMM Viterbi decoding.

Tab. 2 reports cross-sample performance for DINOv2, ResNet-50 and CLIP across five shot counts (1, 3, 5, 10, 15), comparing mean and clustered prototype strategies. DINOv2 with mean prototypes consistently outperforms all other configurations, achieving its best performance at 10 shots (65.6% MOF, 45.1% mIoU, 41.9% F1). Mean prototypes substantially outperform clustered prototypes across both backbones, indicating that splitting each class into multiple sub-centroids introduces false matches under the few-shot budget. DINOv2 consistently outperforms ResNet-50 and CLIP across all shot counts and metrics, suggesting that self-supervised vision transformer features offer better discrimination of fine-grained hand-object interactions in clinical settings.

Comparing Tab. 2 with Tab. 1, the cross-sample setting shows substantially lower performance than within-sample at the same nominal shot count n. Importantly, these are not directly comparable in terms of total support data: within-sample uses n frames per class from a single video, whereas cross-sample pools n frames per class from each of 21 sessions, yielding 20 times more support frames per class overall. Despite this $21\times$ larger support budget, cross-sample performance lags behind, underscoring that the gap is driven by participant-level domain shift (differences in student appearance, camera angle, pace, and workflow ordering across individuals) rather than insufficient support data. Yet the following sections show that this recognition variability is itself pedagogically informative.

5.2. RQ2: Action Sequences, Competency, and Per-Item Analysis

To investigate whether extracted action recognition metrics carry information related to instructor-rated competency, we compute Spearman rank associations between per-video recognition metrics obtained from cross-sample (leave-one-out) evaluation using the best model (DINOv2 + HMM, 10-shot) and the overall competency score (mean of the 11 video-observable rubric items; see Sec. 3).

Table 1. Within-sample few-shot action recognition. For each video, n frames per action class are sampled as prototypes, and the remaining frames serve as the query set. Results are reported as mean ± std over 22 videos. For all backbones, **bold** indicates the best-performing configuration per metric and underlined indicates the second-best. Higher is better for all metrics.

Shots (n)	DINOv2			ResNet50			CLIP (ViT-B/16)		
	MOF	mIoU	F1	MOF	mIoU	F1	MOF	mIoU	F1
1	0.555 ± 0.125	0.419 ± 0.108	0.522 ± 0.115	0.559 ± 0.150	0.416 ± 0.130	0.514 ± 0.140	0.556 ± 0.135	0.393 ± 0.113	0.495 ± 0.124
3	0.783 ± 0.083	0.632 ± 0.088	0.736 ± 0.077	0.786 ± 0.075	0.637 ± 0.078	0.737 ± 0.072	0.797 ± 0.081	0.642 ± 0.086	0.744 ± 0.076
5	0.847 ± 0.072	0.715 ± 0.099	0.805 ± 0.076	0.836 ± 0.079	0.699 ± 0.106	0.791 ± 0.084	0.828 ± 0.082	0.685 ± 0.104	0.779 ± 0.086
10	0.905 ± 0.054	0.797 ± 0.081	0.852 ± 0.067	0.896 ± 0.056	0.795 ± 0.074	0.852 ± 0.065	0.908 ± 0.041	0.798 ± 0.069	0.850 ± 0.067
15	**0.930 ± 0.042**	**0.846 ± 0.077**	**0.869 ± 0.078**	**0.923 ± 0.044**	**0.833 ± 0.090**	**0.856 ± 0.090**	**0.931 ± 0.030**	**0.844 ± 0.066**	**0.858 ± 0.064**

Table 2. Cross-sample few-shot action recognition (leave-one-out, 22 folds). For each held-out video, n frames per action class are sampled from the 21 support sessions to construct prototypes, and the held-out session is classified via HMM Viterbi decoding. Results are reported as mean ± std over 22 folds. **Bold** indicates the best-performing configuration per metric within each backbone, and underlined indicates the second-best. Higher is better for all metrics.

Shots (n)	Proto.	DINOv2			ResNet50			CLIP (ViT-B/16)		
		MOF	mIoU	F1	MOF	mIoU	F1	MOF	mIoU	F1
1	Mean	0.574 ± 0.121	0.337 ± 0.088	0.341 ± 0.092	0.467 ± 0.139	0.255 ± 0.083	0.256 ± 0.073	0.533 ± 0.134	0.306 ± 0.098	0.322 ± 0.095
	Clust.	0.472 ± 0.111	0.280 ± 0.088	0.245 ± 0.061	0.397 ± 0.126	0.225 ± 0.084	0.195 ± 0.055	0.464 ± 0.160	0.292 ± 0.105	0.255 ± 0.073
3	Mean	0.622 ± 0.128	0.433 ± 0.122	0.412 ± 0.106	0.507 ± 0.113	0.302 ± 0.078	0.287 ± 0.058	0.580 ± 0.137	0.405 ± 0.132	0.373 ± 0.111
	Clust.	0.569 ± 0.104	0.403 ± 0.103	0.342 ± 0.084	0.445 ± 0.106	0.257 ± 0.064	0.226 ± 0.063	0.502 ± 0.112	0.337 ± 0.108	0.279 ± 0.081
5	Mean	0.640 ± 0.119	0.436 ± 0.109	0.411 ± 0.096	**0.508 ± 0.139**	**0.327 ± 0.111**	**0.304 ± 0.087**	0.596 ± 0.115	0.405 ± 0.099	0.390 ± 0.099
	Clust.	0.594 ± 0.134	0.423 ± 0.115	0.372 ± 0.107	0.394 ± 0.110	0.247 ± 0.086	0.224 ± 0.061	0.540 ± 0.136	0.371 ± 0.098	0.333 ± 0.083
10	Mean	**0.656 ± 0.152**	**0.451 ± 0.128**	**0.419 ± 0.100**	0.473 ± 0.138	0.298 ± 0.085	0.279 ± 0.076	**0.618 ± 0.124**	**0.439 ± 0.098**	**0.414 ± 0.104**
	Clust.	0.612 ± 0.131	0.433 ± 0.117	0.407 ± 0.112	0.428 ± 0.123	0.282 ± 0.101	0.268 ± 0.089	0.574 ± 0.116	0.400 ± 0.084	0.374 ± 0.076
15	Mean	0.644 ± 0.148	0.446 ± 0.128	0.417 ± 0.106	0.496 ± 0.163	0.299 ± 0.116	0.290 ± 0.109	0.612 ± 0.134	0.426 ± 0.097	0.407 ± 0.088
	Clust.	0.594 ± 0.128	0.399 ± 0.106	0.391 ± 0.099	0.433 ± 0.146	0.256 ± 0.090	0.249 ± 0.078	0.550 ± 0.149	0.385 ± 0.096	0.364 ± 0.087

5.2.1. Overall Trends

Tab. 3 presents a notable pattern: all three recognition accuracy metrics show a negative trend as video-observable competency increases. The strongest observed relationship is for mIoU ($\rho = -0.524$, $p = 0.012$), which measures per-class balance. MOF ($\rho = -0.439$, $p = 0.041$) and F1 ($\rho = -0.433$, $p = 0.044$) show similar patterns. Neither the number of ground-truth action classes nor the number of labeled query frames shows any significant relationship with competency, ruling out annotation-count artifacts as a confounding explanation.

Table 3. Spearman ρ between per-video recognition metrics and overall video-observable competency score (11 items, $N = 22$). All accuracy metrics show a negative trend with competency.

Feature	Spearman ρ	p-value	Pearson r
mIoU (per-class accuracy)	**−0.524**	**0.012**	−0.469
MOF (frame accuracy)	**−0.439**	**0.041**	−0.436
F1 (macro)	**−0.433**	**0.044**	−0.407
Frame error rate (1−MOF)	+0.439	0.041	+0.436
# Action classes in GT	−0.087	0.701	−0.043
# Labeled query frames	−0.071	0.754	+0.016

This pattern is consistent with Moravec's insight [28]: the classifier performs better on the mechanical, templated workflows of lower-performing students, whereas the fluid, adaptive behaviors of higher competence prove harder to recognize. When sessions are split by median competency score, LOW-competency students have 9.5% higher MOF and 8.3% higher mIoU than HIGH-competency students (Fig. 3). One plausible interpretation, consistent with the motor learning principle of abundance [35], is that more competent students perform more diverse workflows with additional safety checks and fluid task transitions, producing greater visual diversity that makes classification harder while earning higher instructor marks. This converges with surgical skill assessment findings where classifiers achieve lower accuracy on higher-skilled practitioners [19, 41], suggesting that the negative trend in recognition accuracy *may* carry a pedagogically informative signal. We note, however, that with $N = 22$ sessions, this interpretation remains exploratory. Importantly, simple sequence features (screen time ratio, transition count, unique action count) extracted from both oracle and predicted timelines show no significant relationship with competency (all $p > 0.10$).

5.2.2. Per-Item Analysis

To identify which facets of clinical competency are most accessible through vision-based action analysis, we examine

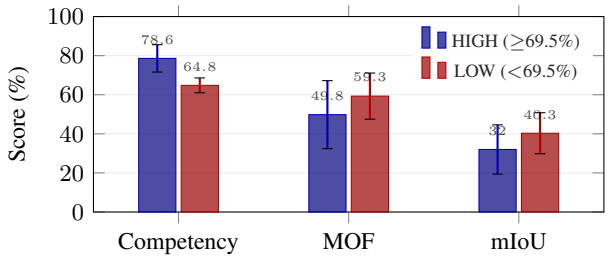


Figure 3. Group-level comparison: sessions split by median video-observable competency score (11 items). Despite higher instructor ratings, HIGH-competency students exhibit *lower* classification accuracy. Error bars show ±1 std.

Spearman associations between per-video MOF and each of the 23 instructor rubric items. Tab. 4 reports the five items with the strongest associations, ordered by magnitude.

Table 4. Top-5 Spearman ρ between per-video MOF and individual rubric items, ordered by magnitude. Shaded rows denote items with visually observable behaviors. N varies across items because instructors may omit ratings when a behavior is not observed or not applicable during a particular session.

#	Rubric Item	ρ	p	N
4	Communicates effectively with team	−0.470	**0.049**	18
18	Uses patient identifiers	−0.455	**0.033**	22
19	Utilizes standardized practices	−0.377	0.083	22
21	Manages technology and equipment	−0.350	0.110	22
12	Prioritizes appropriately	−0.343	0.118	22

The per-item trends are broadly similar in magnitude across items, consistent with the expectation that with $N = 22$ sessions, individual rubric items lack sufficient power to differentiate statistically from one another. Two items reach nominal significance: Item 18 ("Uses patient identifiers," $\rho = -0.455$, $p = 0.033$) and Item 4 ("Communicates effectively with team," $\rho = -0.470$, $p = 0.049$). Patient safety protocols may produce the strongest pattern because students who excel in this expected behavior tend to perform additional wristband-checking and verification steps, generating visually diverse frame sequences that are inherently harder to classify.

Items related to purely procedural tasks performed in a static, repetitive manner (e.g., Item 1 "Obtains pertinent data," $\rho \approx 0$) show no association, consistent with the expectation that these actions appear visually similar regardless of competency level. Overall, these patterns suggest that vision-based analysis is most informative for expected behaviors tied to procedural diversity and protocol complexity, but is fundamentally limited in capturing the *content* of verbal communication and clinical reasoning.

5.3. RQ3: Temporal Patterns of High vs. Low Performers

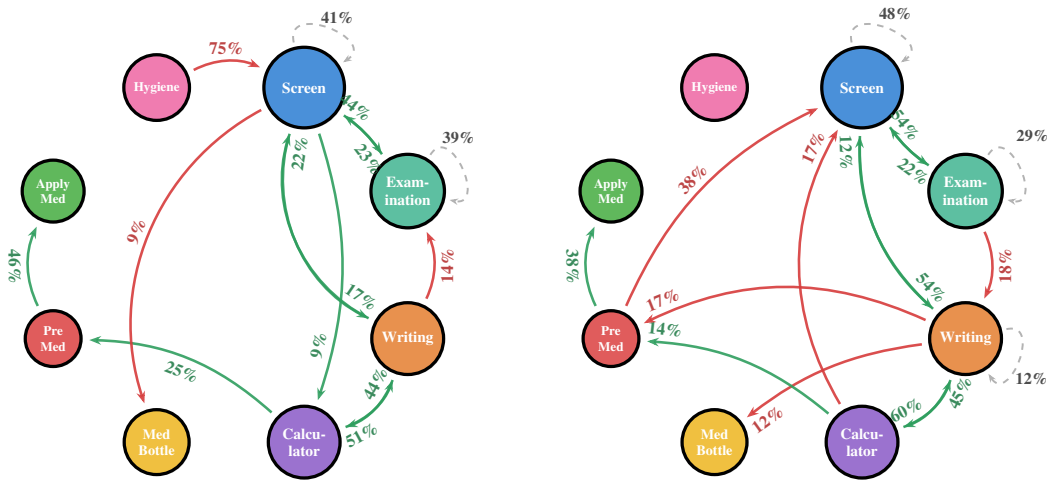
To understand what distinguishes high- from low-performing students, we partition sessions by median competency score and construct process models from ground-truth action sequences (Fig. 4).

Several structural differences emerge (detailed analysis in App. D). Low performers show a higher Screen self-loop (48% vs. 41%), reflecting more time lingering on the bedside monitor, a visually uniform action that inflates MOF. High performers distribute transitions more evenly across Examination, Writing, and Calculator. The medication pathway also differs: high performers show a direct Prep Med → Apply Med transition (46%), while low performers route through Screen (38%), suggesting workflow hesitation. High performers engage in more Examination actions (36 vs. 29), which involve diverse movements that are inherently harder to classify, consistent with the observed negative trend between accuracy and competency. Furthermore, the low-performer model contains more group-unique (red) transitions, indicating irregular workflow paths, while high performers follow a more protocol-consistent progression. Finally, high performers exhibit a strong Hygiene → Screen transition (76%), suggesting more consistent infection-control practices.

To rule out annotation artifacts as the source of this negative trend, we perform a partial association analysis controlling for six potential confounders (annotation coverage, segment count, unique action types, average segment duration, video duration, and total annotations). The MOF–competency association persists across all controls and *strengthens* when controlling for annotation coverage (ρ: $-0.439 \rightarrow -0.546$, $p = 0.009$); full results are reported in App. E.

6. Discussion: Rethinking Eval Metrics

Our findings raise the question of whether higher frame-level accuracy is always the appropriate optimization target for action recognition in educational settings. In our data, the classifier tends to perform better on sessions with repetitive actions, whereas the more fluid and adaptive workflows associated with higher instructor-rated competency appear harder to recognize. This asymmetry is partly rooted in the prototype-based design: each action class is represented by a single centroid, which favors within-class visual consistency. Students who perform an action similarly across instances produce tighter feature clusters that are easier to match, whereas students who vary their approach across instances produce more dispersed features that weaken prototype fit. One interpretation is that competency, as assessed by clinical educators, includes behavioral diversity and procedural flexibility that current vision models do not fully



(a) High Performer Group (b) Low Performer Group

Figure 4. Process models comparing high- and low-competency groups from the ground truth of actions. The 16 fine-grained clinical actions (App. C; background excluded) are aggregated into 8 macro-categories: Examination (Palpate Wrist, Apical Pulse, Lung Sounds, Temperature, Blood Pressure), Hygiene (Hand Hygiene, Gloves), Screen (Patient History, Vital Signs), Writing, Calculator, Med Bottle, Prep Med, and Apply Med. Green edges denote transitions shared by both groups; Red edges are unique to one group. Percentages indicate transition probabilities in the arrow's direction.

capture. This view also aligns with broader work suggesting that the extent to which an individual's responses align with a group can relate to learning and memory outcomes; here, the analogous notion of "fit" is between a student's action sequence and a prototype-based model constructed from peers through leave-one-out cross-sample prototypes.

This observation suggests a practical two-tier approach to automated assessment: (1) the predicted action timeline provides a coarse behavioral summary of what the student did, and (2) the recognition *difficulty* of each session, quantified by mIoU or F1, may serve as a complementary signal of holistic competency. For medication administration, where competency is inherently sequential and correct actions performed in the wrong order can still constitute clinical error, this perspective may help educators identify students whose workflows deviate from the expected procedural pathway even when checklist ratings appear similar. The limited variability in instructor C-CEI ratings further suggests that checklist-based instruments may lack the granularity to distinguish students with clustered overall scores; recognition accuracy may complement such instruments by capturing differences in *how* workflows are executed. More broadly, both rubric-based assessment and vision-based analysis are limited to observable behavior and do not capture the clinical reasoning behind procedural choices. Combining recognition accuracy with post-simulation reflection data, such as structured debriefs or self-assessments, may therefore provide a more complete view of student ability across behavioral and cognitive dimensions. However, with only $N = 22$ sessions, this observation remains exploratory, and recognition accuracy should be viewed as one potential indicator rather than a definitive measure.

7. Conclusion

We presented a three-stage framework for automated competency assessment from egocentric nursing simulation videos. Our results suggest that recognition accuracy may carry a pedagogically informative signal: more competent students produce diverse workflows that are systematically harder to classify. This suggests a two-tier assessment in which predicted action timelines provide a behavioral summary and recognition difficulty provides a complementary competency signal, establishing the utility and boundaries of unimodal video-based assessment.

The primary limitation is the sample size, which reflects the constraints of privacy-regulated clinical data collection and expert annotation; larger multi-site cohorts are needed for generalizability. Vision alone cannot capture verbal communication or the clinical reasoning underlying procedural choices. Integrating audio, gaze, and physiological modalities, along with post-simulation reflection data such as debrief transcripts, could yield a more complete picture of competency across behavioral, cognitive, and metacognitive dimensions. Whether the relationship between recognition accuracy and competency extends to other educational domains remains an open question.

References

[1] Katie Anne Adamson. *Assessing the reliability of simulation evaluation instruments used in nursing education: A test of concept study*. Washington State University, 2011. 1

[2] Paulo Blikstein and Marcelo Worsley. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2):220–238, 2016. 1

[3] Brandon M Booth, Nigel Bosch, and Sidney K D'Mello. Engagement detection and its applications in learning: a tutorial and selective review. *Proceedings of the IEEE*, 111(10): 1398–1422, 2023. 1

[4] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 379–388, 2015. 3

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 2

[6] Clayton Cohn, Caitlin Snyder, Joyce Horn Fonteles, Ashwin TS, Justin Montenegro, and Gautam Biswas. A multimodal approach to support teacher, researcher and ai collaboration in stem+ c learning environments. *British Journal of Educational Technology*, 56(2):595–620, 2025. 1

[7] Eduardo Davalos, Yike Zhang, Ashwin T S, Joyce Horn Fonteles, Umesh Timalsina, and Gautam Biswas. 3d gaze tracking for studying collaborative interactions in mixed-reality environments. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pages 175–183, 2024. 1

[8] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1011–1030, 2024. 2, 3

[9] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015. 1

[10] K Anders Ericsson, Ralf Th Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3):363–406, 1993. 1

[11] Joyce Fonteles, Eduardo Davalos, TS Ashwin, Yike Zhang, Mengxi Zhou, Efrat Ayalon, Alicia Lane, Selena Steinberg, Gabriella Anton, Joshua Danish, et al. A first step in using machine learning methods to enhance interaction analysis for embodied learning environments. In *International Conference on Artificial Intelligence in Education*, pages 3–16. Springer, 2024. 3

[12] Joyce Horn Fonteles, Clayton Cohn, Efrat Ayalon, Mengxi Zhou, Ashwin TS, Eduardo Davalos, Zhijian Li, Surya Rayala, Divya Mereddy, Austin Coursey, et al. Analyzing embodied learning in classroom settings: A human-in-the-loop ai approach for multimodal learning analytics. *Learning and Instruction*, 103:102274, 2026. 3

[13] Isabel Funke, Sjoerd T Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3D convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 14(7):1217–1225, 2019. 3

[14] Emmett D Goodman, Krishna K Patel, Yilun Zhang, William Locke, Chris J Kennedy, Rohan Mehrotra, Stephen Ren, Melody Y Guan, Maren Downing, Hao Wei Chen, et al. A real-time spatiotemporal ai model analyzes skill in open surgical videos. *arXiv preprint arXiv:2112.07219*, 2021. 3

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 1, 3

[16] Daniel A. Hashimoto, Guy Rosman, Daniela Rus, and Ozanan R. Meireles. Artificial intelligence in surgery: Promises and perils. *Annals of Surgery*, 268(1):70–76, 2018. 3

[17] Jennifer K Hayden, Richard A Smiley, Maryann Alexander, Suzan Kardong-Edgren, and Pamela R Jeffries. The NCSBN national simulation study: A longitudinal, randomized, controlled study replacing clinical hours with simulation in prelicensure nursing education. *Journal of Nursing Regulation*, 5(2):C1–S64, 2014. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[19] Sanchit Hira, Digvijay Singh, Tae Soo Kim, Shobhit Gupta, Gregory Hager, Shameema Sikder, and S Swaroop Vedula. Video-based assessment of intraoperative surgical skill. *International journal of computer assisted radiology and surgery*, 17(10):1801–1811, 2022. 6

[20] Sissel Eikeland Husebø, Febe Friberg, Eldar Søreide, and Hans Rystedt. Instructional problems in briefings: How to prepare nursing students for simulation-based cardiopulmonary resuscitation training. *Clinical Simulation in Nursing*, 9(8):e307–e318, 2013. 1

[21] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R Brockmole, and Sidney K D'Mello. Automated gaze-based mind wandering detection during computerized learning in classrooms: S. hutt et al. *User Modeling and User-Adapted Interaction*, 29(4):821–867, 2019. 1

[22] Pamela R Jeffries. A framework for designing, implementing, and evaluating simulations used as teaching strategies in nursing. *Nursing Education Perspectives*, 26(2):96–103, 2005. 1, 2

[23] Suzan Kardong-Edgren, Kathleen A. Adamson, and Cynthia Fitzgerald. A review of currently published evaluation instruments for human patient simulation. *Clinical Simulation in Nursing*, 6(1):e25–e35, 2010. 1

[24] Ahmad Khalifa, Owais Tahhan, Mohammed Albazooni, Mohammed Saeed, Ruha Hamdi, Megan Stanners, Amman Malik, and Adnan Malik. Automated and artificial intelligence

(ai)-derived performance assessment in surgical simulation: A systematic review. *Cureus*, 17(12), 2025. 2

[25] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174, 1977. 3

[26] Kathie Lasater. Clinical judgment development: Using simulation to create an assessment rubric. *Journal of Nursing Education*, 46(11):496–503, 2007. 1

[27] Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[28] Hans Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, Cambridge, MA, 1988. 6

[29] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seez, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. page 102433, 2022. 3

[30] Xavier Ochoa and Marcelo Worsley. Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, 3(2):213–219, 2016. 3

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3, 4

[32] James W. Pellegrino, Naomi Chudowsky, and Robert Glaser. *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academies Press, 2001. 1

[33] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 3, 4

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 4

[35] Rajiv Ranganathan, Mei-Hua Lee, and Karl M Newell. Repetition without repetition: Challenges in understanding behavioral flexibility in motor skill. *Frontiers in Psychology*, 11:2018, 2020. 6

[36] Ginger Schroers and Jill Pfieffer. Tool development and testing: An objective measurement of medication administration competency. *Nursing Education Perspectives*, 46(5):E37–E39, 2025. 2, 1

[37] Deborah Schwengel, Ignacio Villagrán, Geoffrey Miller, Constanza Miranda, et al. Multimodal assessment in clinical simulations: A guide for moving towards precision education. *Medical Science Educator*, 35(2):1025–1034, 2024. 1

[38] Saswat Shankar, Caitlyn Ruiz, Zijian Zheng, Marcelo Worsley, and Paulo Blikstein. Multimodal methods for analyzing learning and training environments: A systematic literature review. *arXiv preprint arXiv:2408.14491*, 2024. 3

[39] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4

[40] Caitlin Snyder, Nicole M Hutchins, Clayton Cohn, Joyce Horn Fonteles, and Gautam Biswas. Analyzing students collaborative problem-solving behaviors in synergistic stem+ c learning. In *Proceedings of the 14th learning analytics and knowledge conference*, pages 540–550, 2024. 3

[41] Abed Soleymani, Ali Akbar Sadat Asl, Mojtaba Yeganejou, Scott Dick, Mahdi Tavakoli, and Xingyu Li. Surgical skill evaluation from robot-assisted surgery recordings. In *2021 International Symposium on Medical Robotics (ISMR)*, pages 1–6. IEEE, 2021. 2, 6

[42] Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. Multimodal integration of human-like attention in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2648–2658, 2023. 3

[43] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2):1012–1027, 2021. 1

[44] Michael Todd, Julie A. Manz, Kathleen S. Hawkins, Michele E. Parsons, and Mary Hercinger. The development of a quantitative evaluation tool for simulations in nursing education. *International Journal of Nursing Education Scholarship*, 5(1), 2008. 1

[45] Andru Putra Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017. 3

[46] Caleb Vatral, Gautam Biswas, and Benjamin Goldberg. A theoretical framework for performance analysis in competency-based experiential learning environments. In *AI and Gamification Technologies for Complex Work*, pages 1–19. CRC Press, 2025. 1

[47] Hung-Hsuan Yen, Ming-Chih Ho, Yi-Hsiang Hsiao, and Chun-Chieh Huang. Surgical video-based temporal action analysis algorithm and competency assessment in laparoscopic cholecystectomy: development and exploratory evaluation. *Surgical Endoscopy*, 2025. 2

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 ($160\ \text{mg}/5\ \text{mL}$):

$$15\ \text{mg/kg} \times 9.6\ \text{kg} = 144\ \text{mg},$$
$$144\ \text{mg} \times \frac{5\ \text{mL}}{160\ \text{mg}} = 4.5\ \text{mL}. \quad (3)$$

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) [44], 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 [36]. 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 high- and low-performer process models (Fig. 4) are elaborated below.

Screen self-loop. Low-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. High performers 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 LOW group.

Medication pathway. High performers show a strong direct Prep Med $\rightarrow$ Apply Med transition (46%), indicating a coherent prepare-then-administer sequence. Low performers lack this link; instead, Prep Med routes back to Screen (38%), suggesting hesitation or uncertainty in the medication procedure.

Examination frequency. High performers engage in more Examination actions (36 vs. 29), while low performers 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 low-performer model contains more group-unique (red) transitions, indicating ir-

regular workflow paths. High performers follow a more protocol-consistent progression with fewer idiosyncratic transitions.

Hygiene compliance. Hygiene actions connect to Screen with 76% probability in high performers, suggesting consistent hand hygiene before engaging with the patient monitor. This transition is less prominent in low performers, 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 HIGH 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; [25]). 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 [8].