

# ***Intelligent Assessment of Body Posture and Movement Timing in University Fencing Instruction: Development and Validation of a Phase-Aware, Confidence-Gated Model***

Huijian Deng<sup>1</sup>, Zijing Zhang<sup>2</sup>, Zheng Luo<sup>3,\*</sup>

<sup>1</sup>Department of Physical Education, Chongqing University of Technology, Chongqing, 400054, China

<sup>2</sup>Department of Physical Education, Chongqing University of Science and Technology, Chongqing, 401331, China

<sup>3</sup>Hunan Institute of Science and Technology, Yueyang, Hunan, 414006, China

\* Corresponding author

**Keywords:** Fencing instruction; Movement technique assessment; Human pose estimation; Movement timing; Formative feedback

**Abstract:** This study aimed to develop an intelligent assessment method for evaluating body posture and movement timing during the stationary straight thrust in épée and to examine its performance, criterion validity, and instructional application. Fifteen experts participated in two rounds of Delphi consultation to establish the assessment framework. The model was developed using 1,200 videos collected from 60 students and was independently validated using 120 videos from an additional 20 students. To evaluate the instructional application of the model, 56 students were stratified by sex and pretest score and then randomly assigned to either a conventional-instruction group or a group receiving conventional instruction supplemented with intelligent feedback for eight weeks. The proposed model integrated single-person COCO-17 keypoint extraction, phase feature encoding, confidence gating, and multitask regression for posture and movement-timing assessment. The FPT+CPG model achieved a phase-recognition accuracy of 83.03%, a macro-averaged F1 score of 82.43%, a scoring mean absolute error (MAE) of 3.85 points, and a coefficient of determination ( $R^2$ ) of 0.845. The model also demonstrated strong agreement with the mean total scores assigned by experts, with an intraclass correlation coefficient (ICC(2,1)) of 0.925 and an MAE of 2.80 points. After adjustment for pretest scores, students who received intelligent feedback scored 6.08 points higher at posttest than those who received conventional instruction alone (95% confidence interval, 3.98–8.17;  $p < 0.001$ ). The intelligent-feedback approach also reduced the mean feedback time to  $14.49 \pm 3.02$  s, compared with  $46.12 \pm 6.01$  s for conventional instruction ( $p < 0.001$ ). These findings indicate that the joint modeling of movement-phase information and keypoint confidence can support the assessment of body posture and movement timing during the stationary straight thrust in épée. The results further suggest that instructor-reviewed intelligent feedback may improve movement performance and reduce feedback time when used as an adjunct to formative assessment in university fencing instruction.

## 1. Introduction

Assessment of movement technique provides a basis for judging learning quality in physical education and informs demonstrations, error correction, and subsequent practice. Fencing movements are rapid and brief. In large classes, instructors must attend to several students simultaneously and can often observe only selected attempts, making continuous recording of each practice attempt difficult. The principal challenge is therefore a mismatch between observational capacity, feedback timeliness, and class size, rather than a lack of professional judgment. Markerless computer vision can extract information about body movement using ordinary cameras, offering a minimally intrusive means of recording and quantitatively analyzing technique [1].

Human pose estimation converts video into joint coordinates, joint connections, and detection confidence scores. Skeleton-based action recognition then learns movement representations from spatial topology and temporal changes [2,3]. Compared with direct processing of RGB images, keypoint sequences depend less on clothing color and background texture. Their structure corresponds to human anatomy, making it easier to translate model outputs into joints, movement phases, and deviations that instructors can interpret. However, two-dimensional pose estimation remains vulnerable to occlusion, motion blur, perspective distortion, and keypoint ambiguity. Reviews of sport-specific movement recognition consequently emphasize the need to report data partitioning, robustness, and validity for the intended task alongside model performance [4].

Computer vision research in fencing has gradually expanded from action recognition to movement quality analysis. Malawski and colleagues used multimodal cues to recognize fencing action dynamics [5] and subsequently assessed fencing footwork skill and performance using pose information [6]. The FACTS dataset treats tactical actions such as attacks, counterattacks, ripostes, and remises as fine-grained classification targets [7]. These studies demonstrate the feasibility of applying computer vision to rapid fencing movements. However, competition or tactical classification primarily identifies which action occurred. Classroom assessment must also establish how well an action was performed, the phase in which a deviation occurred, and whether feedback can improve the next attempt.

University classes present additional challenges, including other students entering the frame, self-occlusion, and substantial variation in beginners' movement patterns [8]. Models such as ST-GCN, CTR-GCN, and PoseC3D represent spatiotemporal skeletal information through graph topology, channel relationships, and heatmap volumes, respectively [9–11]. Hierarchical representation learning provides the technical foundation for these approaches [12]. Nevertheless, classification accuracy on general benchmarks does not directly establish validity for classroom assessment. If data are partitioned by video rather than by student, the same student's physical characteristics may appear in both training and test sets, inflating performance estimates. Similarly, reporting correlation alone cannot establish numerical agreement between a model and experts.

The stationary straight thrust in *épée* is a foundational exercise for beginners learning to sequence the weapon arm and control the body. Here, stationary means that foot positions remain essentially unchanged; lunges, crossover steps, and advances or retreats are excluded. Assessment is restricted to body posture and movement timing that can be inferred from one person's body keypoints. It does not evaluate weapon-tip trajectories, target contact location, engagement distance, points won or lost, or tactical choices. This boundary aligns conclusions with observable evidence and prevents a body-keypoint model from being interpreted as a comprehensive assessment of fencing competence.

To address these issues, we developed a multitask model incorporating phase information and confidence gating. Movement phases first provide temporal priors; keypoint confidence then modulates phase-related features; and the model simultaneously outputs the phase class, body

posture score, and movement timing score. The study followed a sequence of indicator development, model development, independent validation, and instructional application to answer four questions: whether the model improves phase recognition and continuous scoring relative to a baseline; whether it retains a relative advantage when keypoints are missing; whether its scores achieve acceptable agreement with expert ratings; and whether intelligent feedback improves learning outcomes and reduces feedback time in instructor-led classes.

The study makes three contributions. In terms of assessment content, it decomposes the stationary straight thrust into four phases and ten observable indicators. Methodologically, it combines phase features and keypoint confidence in multitask assessment of posture and timing. For instructional application, it links algorithmic performance, professional judgment, and classroom outcomes through a participant-independent validation sample and an eight-week controlled teaching experiment, avoiding an inference of educational value from model accuracy alone.

## **2. Theoretical background and assessment framework**

### **2.1 Skill learning and feedback**

Fitts and Posner described cognitive, associative, and autonomous stages of skill learning [13]. University fencing beginners are generally in the cognitive or associative stage. They need to translate demonstrations into actionable rules for body control and gradually stabilize their movements through repeated practice. At these stages, knowledge of results that merely labels an attempt as good or poor provides insufficient guidance for correction. Identifying whether a deviation occurs during initiation, extension, or the final hold, and specifying the involvement of the trunk, shoulder, elbow, wrist, or supporting posture, provide information closer to knowledge of performance.

The value of formative feedback depends on its content, timing, and use by the learner. Hattie and Timperley emphasized that effective feedback connects learning goals, current performance, and subsequent action [14]. Accordingly, the system should provide component scores, low-confidence alerts, and reviewable indications of errors, rather than a total score alone. Conversely, overly frequent prompts may foster dependence. Instructors should adjust feedback frequency to students' skill levels and reduce external prompts as performance stabilizes, encouraging proprioception and self-assessment.

### **2.2 Integrating technology and pedagogy**

The technological pedagogical content knowledge framework emphasizes interactions among technological, content, and pedagogical knowledge [15]. In this study, pose estimation and skeletal temporal modeling represent technological knowledge; the structure and common errors of the stationary straight thrust represent fencing content knowledge; and demonstration, questioning, feedback timing, and practice organization represent pedagogical knowledge. Algorithmic outputs become instructionally useful only when these forms of knowledge are integrated. The model undertakes repeated observation, preliminary quantification, and error localization; the instructor interprets and verifies the outputs and selects interventions; and students test the feedback through further practice.

We therefore adopted a human-machine collaborative approach. When keypoint quality and model confidence are sufficient, the system offers component-level suggestions for posture and timing. When occlusion is severe, phase boundaries are unclear, or outputs approach category thresholds, the system flags the attempt for review instead of automatically issuing a definitive

judgment. This mechanism makes model confidence visible and leaves final responsibility with the instructor.

### 2.3 Movement phases and assessment indicators

Following the instructional sequence for the stationary straight thrust, the movement was divided into four phases: preparation, initiation, extension, and final hold. Preparation concerns body stability and initial alignment on the weapon side. Initiation concerns the onset of weapon-arm movement and any obvious preparatory swing. Extension concerns shoulder–elbow–wrist sequencing, continuity of elbow extension, and compensatory trunk movement. The final hold concerns body stability after the arm reaches its extended position. The phases localize deviations without imposing identical movement durations on all students.

The assessment comprises two primary dimensions, body posture and movement timing, each containing five secondary indicators (Table 1). Body posture concerns the trunk, head and neck, weapon-side joint alignment, balance of the non-weapon arm, and lower-limb support. Movement timing concerns movement onset, the sequence of arm extension, continuity of elbow extension, trunk–upper-limb coordination, and the final hold. The total score is 100 points and represents classroom technique within the scope of this rubric, rather than competitive ability.

Table 1. Assessment indicators for body posture and movement timing in the stationary straight thrust.

| Dimension       | Indicator                                       | Observation phase                  | Weight |
|-----------------|-------------------------------------------------|------------------------------------|--------|
| Body posture    | Upright trunk and lateral stability             | Preparation, extension, final hold | 0.111  |
| Body posture    | Head stability                                  | All phases                         | 0.092  |
| Body posture    | Weapon-side shoulder–elbow–wrist alignment      | Initiation, extension, final hold  | 0.106  |
| Body posture    | Balance of the non-weapon arm                   | All phases                         | 0.103  |
| Body posture    | Lower-limb support and center-of-mass stability | All phases                         | 0.090  |
| Movement timing | Identification of movement onset                | Preparation–initiation             | 0.111  |
| Movement timing | Sequence of arm-extension initiation            | Initiation–extension               | 0.096  |
| Movement timing | Continuity of elbow extension                   | Extension                          | 0.089  |
| Movement timing | Trunk–upper-limb coordination                   | Initiation, extension              | 0.097  |
| Movement timing | Movement completion and stable hold             | Final hold                         | 0.105  |

Note. Weights sum to 1.000: body posture, 0.502; movement timing, 0.498.

### 2.4 Research hypotheses

Four hypotheses were proposed. H1: Adding phase feature encoding and confidence gating would improve phase recognition and continuous scoring relative to the baseline. H2: The model incorporating phase information and confidence gating would show smaller performance declines in a stress test with randomly missing keypoints. H3: Model and expert scores would show high

agreement for body posture, movement timing, and the total score. H4: Adding instructor-reviewed intelligent feedback to otherwise equivalent conventional instruction would improve adjusted posttest performance and shorten feedback time per attempt.

### 3. Materials and methods

#### 3.1 Study design and participants

The study used four sequential stages. First, assessment indicators were developed from the literature, curricular standards, and expert consultation. Second, stationary straight-thrust videos were collected and annotated to develop the multitask model. Third, agreement between model and expert scores was evaluated using videos from an entirely independent student sample. Fourth, classroom learning outcomes were compared between conventional instruction and conventional instruction supplemented with intelligent feedback. Participants in model development, independent validation, and the teaching experiment did not overlap. These samples were used for model development, independent criterion validation, and evaluation of instructional effects, respectively, reducing performance inflation from data leakage and repeated use of participants.

Fifteen experts participated in two consultation rounds. The development sample comprised 60 students enrolled in elective fencing courses, each contributing 20 valid videos, for a total of 1,200 videos. The independent validation sample comprised another 20 students and 120 videos. The teaching experiment included 56 students who were stratified by sex and pretest score and randomly assigned to conventional instruction or intelligent feedback, with 28 students per group. Development data were divided into training, validation, and test sets by participant rather than by video; no student appeared in more than one set. The independent validation sample was not used for model training or hyperparameter selection.

Students were eligible if they could complete the course exercises and had no acute injury affecting upper- or lower-limb movement. Exclusion criteria were missing video data, prolonged loss of keypoint visibility, or failure to complete the prescribed practice.

#### 3.2 Indicator development

The initial indicator pool was jointly reviewed by fencing instructors, specialists in physical education measurement and evaluation, and computer vision researchers. Experts rated indicator importance, observability, and classroom interpretability on five-point scales and commented on definitions and scoring anchors. We calculated the valid response rate, expert authority coefficient, indicator means, coefficients of variation, and Kendall's coefficient of concordance. Indicators were retained if their mean importance rating was at least 4.00 and their coefficient of variation was no greater than 0.20. We normalized second-round importance scores to determine indicator weights and examined whether the weights of the primary dimensions aligned with instructional priorities.

Each indicator included four specifications: the technical requirement, observable evidence, common deviations, and scoring anchors. For example, trunk stability did not require the body to remain completely motionless. Instead, it concerned whether forward lean, lateral lean, or compensatory rotation during arm extension exceeded the range considered acceptable for instruction. Similarly, the final hold did not prescribe a fixed duration; it concerned whether a clearly identifiable stable configuration could be maintained briefly after completion. These operational definitions informed expert scoring and the generation of interpretable model feedback.

### 3.3 Video acquisition and pose data

Videos were recorded from a fixed camera position at  $1920 \times 1080$  pixels and 30 frames/s. The camera was approximately 4 m from the student and 1.2 m above the ground. An oblique frontal view from the weapon side captured the entire body. Each clip began approximately 0.5 s before the preparatory posture and ended approximately 0.5 s after the final hold. Floor markings maintained camera placement across sessions.

The pose estimator retained only the primary performer and returned horizontal coordinates, vertical coordinates, and detection confidence for the COCO-17 keypoints. Translation was normalized to the hip center, and scale was normalized to shoulder width. Missing keypoints were interpolated linearly over short intervals while retaining a missingness mask; clips exceeding the threshold for consecutive missing observations were not scored automatically. Sequences were resampled to 64 frames over the valid movement interval. The model input is specified in Equation (1), where  $C$ ,  $T$ , and  $V$  denote the number of channels, sequence length, and number of joints, respectively, and  $s$  denotes detection confidence.

$$\mathbf{X} \in \mathbb{R}^{C \times T \times V}, \quad \mathbf{H}_{C,T,V} = \mathbf{H}_{64,17,1}, \quad \mathbf{x}_{t,v} = \mathbf{H}_{t,v} \mathbf{y}_{t,v} \mathbf{s}_{t,v} \quad (1)$$

The recording area and floor markings constrained variation in viewing angle. Normalization of two-dimensional coordinates reduced variation arising from recording distance and body size, but could not eliminate perspective error. COCO-17 has no keypoints for the blade or weapon tip. We therefore did not extrapolate weapon-tip position from the wrist or equate wrist movement with successful target contact.

### 3.4 Model architecture

The model comprised pose preprocessing, a spatiotemporal graph backbone, phase feature encoding, confidence gating, and multitask outputs (Figure 1). The backbone used natural anatomical connections to construct the graph. Spatial graph convolutions aggregated adjacent-joint information, and temporal convolutions represented movement over time. To support classroom assessment beyond action classification, the shared representation was followed by three task heads: a four-phase classification head, a body posture regression head, and a movement timing regression head.

The phase feature encoding module generated probabilities  $P$  for the four phases from wrist and elbow velocities, elbow angle, changes in shoulder and hip axes, and support stability. Phase probabilities did not directly constitute scores; they supplied temporal context for the shared features. For example, the same elbow-angle deviation has different instructional implications during preparation and the final hold. Expert-annotated phase boundaries provided training supervision. At inference, the model produced framewise phase probabilities, with duration and transition-order constraints reducing implausible phase switching.

Keypoint-confidence gating combined phase probabilities  $P$  with joint confidence  $S$ , as shown in Equation (2). Here,  $g(\cdot)$  maps phase probabilities to joint-channel weights,  $\odot$  denotes elementwise multiplication,  $F$  represents shared features, and  $F'$  represents gated features. The factor  $0.5 + 0.5S$  retained limited information from low-confidence keypoints instead of setting their contributions entirely to zero, avoiding abrupt feature changes caused by occasional confidence reductions. Phase weights also enhanced responses directly relevant to assessment, such as those of the weapon-side shoulder, elbow, and wrist during extension.

$$\mathbf{G} = g(\mathbf{P}) \odot \mathbf{H}_{0.5 + 0.5S}, \quad \mathbf{F}' = \mathbf{F} \odot \mathbf{H} + \mathbf{G} \quad (2)$$

The total loss is given in Equation (3).  $\mathcal{L}_{\text{phase}}$  denotes the phase cross-entropy loss;  $\mathcal{L}_{\text{posture}}$  and  $\mathcal{L}_{\text{timing}}$  are Huber losses for the two continuous scores; and  $\mathcal{L}_{\text{smooth}}$  penalizes abrupt changes in phase probabilities and intermediate features between adjacent frames. The coefficients  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  were selected using the validation set. Body posture and movement timing were each scored from 0 to 50, and a total score from 0 to 100 was obtained using the secondary-indicator weights in Table 1. Both dimension scores and the total score were retained to prevent an aggregate score from obscuring specific deviations.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{phase}} + \lambda_1 \mathcal{L}_{\text{posture}} + \lambda_2 \mathcal{L}_{\text{timing}} + \lambda_3 \mathcal{L}_{\text{smooth}} \quad (3)$$

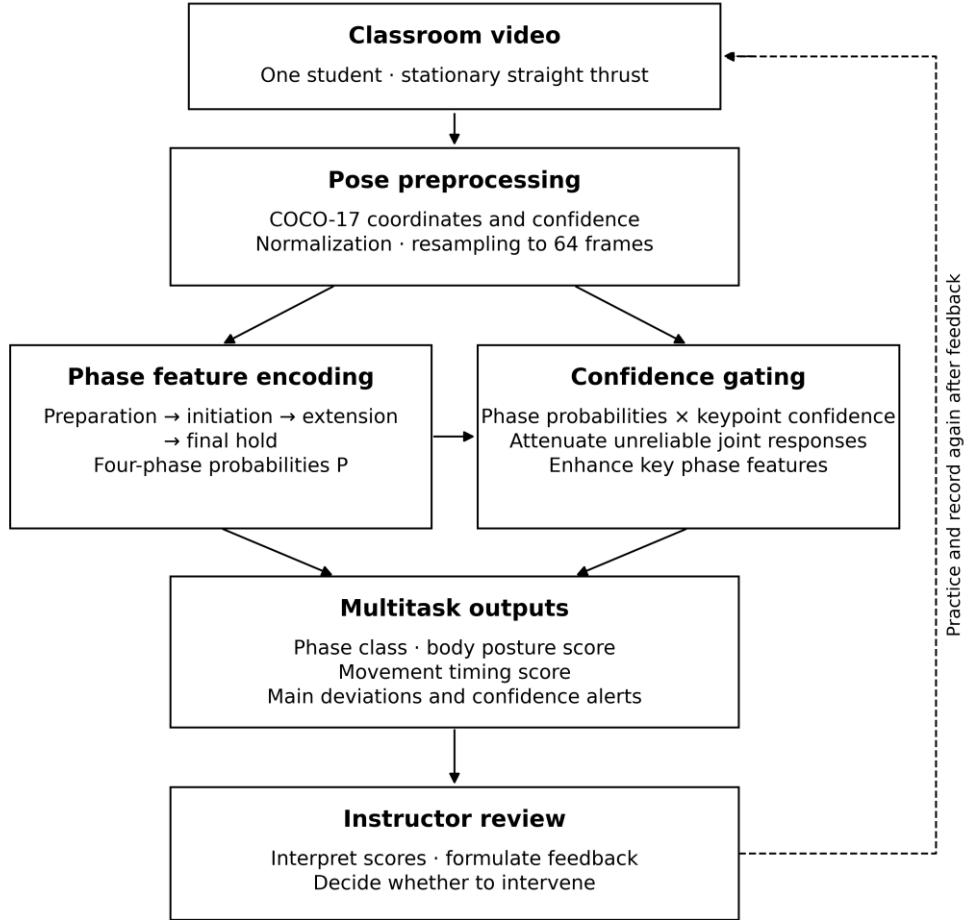


Figure 1. Multitask assessment workflow incorporating phase information and confidence gating.

After feedback, the student practices again and another clip is recorded. The instructor retains final judgment.

We compared BaselineGCN, FPT with phase encoding alone, and FPT+CPG with both phase encoding and confidence gating. Models were trained using AdamW with an initial learning rate of 0.001, a batch size of 32, and a maximum of 100 epochs. Early stopping was based on the combined validation loss. Small translation, scale, and temporal perturbations were applied to the coordinates. Each model was run with three random seeds.

Phase recognition was evaluated using accuracy and macro-averaged F1 (macro-F1). Continuous scoring was evaluated using mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination ( $R^2$ ). To assess sensitivity to unreliable keypoints, we artificially removed 0%, 10%, or 20% of keypoint coordinates in the test set and reduced their confidence

accordingly. This stress test characterized sensitivity to discrete missing detections; it did not replace testing under actual continuous occlusion, motion blur, or interference from background participants.

### 3.5 Validation against expert ratings

Three experts with experience in fencing instruction or coaching scored the videos independently while blinded to model outputs and student identity. Before formal scoring, example videos were used to align indicator definitions and rating anchors; the independent validation sample was not discussed. The 120 validation videos were excluded from training, validation during model development, and hyperparameter selection. The mean of the three expert scores served as the criterion. Interrater agreement was examined before comparing the model with expert ratings. Expert scores were treated as a testable reference linked to professional classroom judgment, rather than as absolute ground truth.

Interrater reliability was assessed using a two-way random-effects, absolute-agreement, average-measures intraclass correlation coefficient, ICC(2,k). Model–expert agreement was assessed using the two-way random-effects, absolute-agreement, single-measure ICC(2,1), with 95% confidence intervals (CIs) [16]. We also calculated MAE, RMSE, correlation coefficients, and the Bland–Altman mean difference and 95% limits of agreement [17]. Correlation reflects rank-order association, whereas ICC and limits of agreement address numerical agreement; these measures are not interchangeable.

### 3.6 Teaching experiment

The teaching experiment lasted eight weeks, with two sessions per week and 20 min of stationary straight-thrust practice per session. Both groups were taught by the same instructor using the same instructional progression, demonstrations, number of practice attempts, and facilities. The conventional-instruction group was observed by the instructor and received verbal correction. In addition to conventional instruction, the intelligent-feedback group received phase localization, separate posture and timing scores, low-confidence alerts, and one principal improvement suggestion. The instructor reviewed the output and decided whether to present it to the student. Groups practiced at different times to reduce contamination. Experts blinded to group allocation scored both pretests and posttests. The primary outcome was the total stationary straight-thrust score. Secondary outcomes were posture and timing scores, feedback time per attempt, and students' five-point ratings of comprehensibility, actionability, and satisfaction. The system did not contribute to summative course grades.

Intelligent feedback followed the principle of one attempt, one main issue, and one subsequent attempt, avoiding several simultaneous prompts. If confidence was insufficient or the model output conflicted with the instructor's judgment, the instructor could stop automated feedback and record the reason. The intervention was therefore instructor-reviewed feedback enabled by reduced observation and error-localization time, rather than algorithm use in isolation.

### 3.7 Statistical analysis

Statistical analyses were conducted in SPSS 27.0. Continuous variables were examined for distributional properties, outliers, and homogeneity of variance and were summarized as means  $\pm$  standard deviations (SDs). Delphi results were summarized using valid response rates, expert authority coefficients, coefficients of variation, and Kendall's *W*. Model performance was reported as the mean  $\pm$  SD across three runs. Given the small number of runs, between-model comparisons

were interpreted descriptively without overstating statistical significance. ICC point estimates and 95% CIs were calculated using the prespecified models. Bland–Altman analysis examined fixed bias, outliers, and score-dependent heteroscedasticity.

For instructional outcomes, baseline comparability and the group  $\times$  pretest interaction were examined first. Analysis of covariance (ANCOVA) then used the posttest total score as the dependent variable, group as the fixed factor, and pretest score as the covariate. This strategy controls for baseline differences more effectively than comparing change scores alone [18]. We reported the adjusted between-group difference, its 95% CI, the  $F$  statistic, and partial  $\eta^2$ ; Cohen's  $d$  was additionally reported for change scores. All tests were two-sided, with  $\alpha = 0.05$ . The primary outcome and analysis methods were not changed according to the direction of the results.

## 4. Results

### 4.1 Expert consultation

All 15 experts returned valid questionnaires in both consultation rounds, yielding a 100% valid response rate in each round. The expert authority coefficient was  $0.834 \pm 0.040$ . Kendall's  $W$  increased from 0.241 in the first round ( $\chi^2 = 32.581$ ,  $p < 0.001$ ) to 0.544 in the second round ( $\chi^2 = 73.453$ ,  $p < 0.001$ ), indicating greater concordance among experts. Second-round mean importance ratings ranged from 4.00 to 5.00, and coefficients of variation ranged from 0 to 0.115. All ten indicators met the prespecified retention criteria. The weights for body posture and movement timing were 0.502 and 0.498, respectively (Table 1).

### 4.2 Model performance

Relative to BaselineGCN, FPT showed only small changes in phase-recognition metrics, but scoring MAE decreased from 5.90 to 5.12 points. With CPG added, FPT+CPG achieved an accuracy of 83.03% and a macro-F1 of 82.43%, exceeding BaselineGCN by 4.17 and 4.14 percentage points, respectively. Scoring MAE decreased from 5.90 to 3.85 points, a reduction of 2.05 points (34.7%); RMSE decreased by 2.09 points; and  $R^2$  increased from 0.696 to 0.845 (Table 2). Descriptive results across three random seeds suggest that phase feature encoding primarily reduced continuous-scoring error, whereas combining phase encoding with confidence gating improved both classification and regression. These findings provide descriptive support for H1.

Table 2. Phase-recognition and scoring performance of the models.

| Model       | Accuracy (%)     | Macro-F1 (%)     | MAE (points)    | RMSE (points)   | $R^2$             |
|-------------|------------------|------------------|-----------------|-----------------|-------------------|
| BaselineGCN | $78.86 \pm 0.41$ | $78.29 \pm 0.33$ | $5.90 \pm 0.19$ | $7.13 \pm 0.16$ | $0.696 \pm 0.015$ |
| FPT         | $79.08 \pm 0.26$ | $78.82 \pm 0.43$ | $5.12 \pm 0.13$ | $6.54 \pm 0.21$ | $0.748 \pm 0.008$ |
| FPT+CPG     | $83.03 \pm 0.14$ | $82.43 \pm 0.57$ | $3.85 \pm 0.03$ | $5.04 \pm 0.12$ | $0.845 \pm 0.010$ |

Note. Values are means  $\pm$  SDs across three random seeds per model. The number of runs supports descriptive comparisons only. MAE, mean absolute error; RMSE, root mean square error;  $R^2$ , coefficient of determination.

### 4.3 Robustness to missing keypoints

Performance declined for both models as keypoint missingness increased, but the decline was smaller for FPT+CPG. From 0% to 20% missingness, BaselineGCN's macro-F1 fell by 31.1 percentage points and its scoring MAE increased by 6.1 points. For FPT+CPG, macro-F1 fell by

18.3 percentage points and scoring MAE increased by 3.9 points. At 20% missingness, FPT+CPG achieved a macro-F1 of 64.4%, 17.6 percentage points higher than BaselineGCN, and an MAE of 7.8 points, 4.1 points lower than the baseline (Table 3). H2 was therefore supported within the scope of the artificial random-missingness stress test; these findings do not directly represent performance under actual continuous occlusion.

Table 3. Model performance under keypoint missingness.

| Model       | Missing keypoints (%) | Macro-F1 (%) | Scoring MAE (points) |
|-------------|-----------------------|--------------|----------------------|
| BaselineGCN | 0                     | 77.9         | 5.8                  |
| BaselineGCN | 10                    | 62.4         | 8.4                  |
| BaselineGCN | 20                    | 46.8         | 11.9                 |
| FPT+CPG     | 0                     | 82.7         | 3.9                  |
| FPT+CPG     | 10                    | 74.8         | 5.7                  |
| FPT+CPG     | 20                    | 64.4         | 7.8                  |

#### 4.4 Agreement between model and expert scores

The ICC(2,k) for total scores across the three experts was 0.972. Agreement between model total scores and mean expert scores yielded an ICC(2,1) of 0.925 and a Pearson correlation of  $r = 0.925$  ( $p < 0.001$ ). Total-score MAE was 2.80 points and RMSE was 3.37 points. The Bland–Altman mean difference was  $-0.04$  points, with 95% limits of agreement from  $-6.67$  to  $6.58$  points. Errors and correlations for body posture and movement timing are reported in Table 4. Association, absolute error, and agreement results collectively support H3. However, the limits of agreement require interpretation against a prespecified tolerance for instructional use and do not justify complete replacement of expert scoring.

Table 4. Agreement between model scores and expert ratings.

| Dimension       | MAE (points) | RMSE (points) | Mean bias (points) | $r$   |
|-----------------|--------------|---------------|--------------------|-------|
| Body posture    | 2.07         | 2.55          | 0.19               | 0.871 |
| Movement timing | 1.98         | 2.55          | $-0.23$            | 0.853 |
| Total score     | 2.80         | 3.37          | $-0.04$            | 0.925 |

Note. The total score ranges from 0 to 100; each dimension ranges from 0 to 50. MAE, mean absolute error; RMSE, root mean square error;  $r$ , correlation coefficient.

#### 4.5 Instructional outcomes

The groups differed by only 0.17 points at pretest ( $p = 0.901$ ). After eight weeks, unadjusted posttest means were  $67.17 \pm 5.67$  points for conventional instruction and  $73.40 \pm 6.06$  points for intelligent feedback. The total-score gain was  $11.23 \pm 4.40$  points in the intelligent-feedback group, compared with  $5.18 \pm 3.47$  points in the conventional-instruction group. Gains in posture and timing followed the same pattern (Table 5). The group  $\times$  pretest interaction was not significant ( $p = 0.581$ ). After adjustment for pretest score, the intelligent-feedback group scored 6.08 points higher at posttest than the conventional-instruction group (95% CI, 3.98–8.17),  $F(1,53) = 33.792$ ,  $p < 0.001$ , partial  $\eta^2 = 0.389$ . Cohen's  $d$  for the difference in total-score gains was 1.53.

Table 5. Stationary straight-thrust assessment outcomes by group.

| Outcome              | Conventional instruction<br>( <i>n</i> = 28) | Intelligent feedback<br>( <i>n</i> = 28) | Between-group comparison |
|----------------------|----------------------------------------------|------------------------------------------|--------------------------|
| Pretest total score  | 62.00 ± 4.97                                 | 62.17 ± 5.45                             | <i>p</i> = 0.901         |
| Posttest total score | 67.17 ± 5.67                                 | 73.40 ± 6.06                             | ANCOVA in text           |
| Posture score gain   | 2.94 ± 2.05                                  | 5.43 ± 2.32                              | <i>p</i> < 0.001         |
| Timing score gain    | 2.24 ± 1.81                                  | 5.80 ± 2.44                              | <i>p</i> < 0.001         |
| Total score gain     | 5.18 ± 3.47                                  | 11.23 ± 4.40                             | <i>p</i> < 0.001         |

Note. Scores are presented as means ± SDs in points. Unadjusted posttest means are descriptive. The primary between-group effect was estimated by ANCOVA with posttest total score as the dependent variable and pretest score as the covariate. The adjusted between-group difference was 6.08 points (95% CI, 3.98–8.17).

Feedback time per attempt was 14.49 ± 3.02 s with intelligent feedback and 46.12 ± 6.01 s with conventional instruction (*p* < 0.001), a mean reduction of 31.63 s (68.6%). On five-point scales, the intelligent-feedback group rated comprehensibility, actionability, and satisfaction at 4.38 ± 0.40, 4.32 ± 0.46, and 4.30 ± 0.48, respectively. The movement outcomes and feedback-efficiency findings support H4.

## 5. Discussion

### 5.1 The role of phase information

The stationary straight thrust is more than a sequence of static poses; the technical meaning of a posture depends on the phase in which it occurs. Moderate elbow flexion is appropriate during preparation, whereas substantial flexion at completion may indicate insufficient extension. Brief changes in trunk angle must likewise be interpreted in relation to initiation and extension order. Phase feature encoding introduces this temporal context into the model, allowing the same joint configuration to receive different weights at different times. The phase module therefore offers more than improved classification accuracy: it provides a structured temporal reference for component scoring and error localization.

FPT produced limited improvements in classification metrics but more clearly reduced scoring error. Adding CPG further improved classification, regression, and performance in the artificial-missingness stress test. These results suggest that phase priors help distinguish movement processes, while confidence gating limits interference from poor-quality joint detections in phase recognition and scoring. The mechanisms appear complementary rather than merely representing the addition of more parameters.

However, the study did not include strictly parameter-matched controls, alternative gating functions, or a complete ablation analysis separately removing confidence and phase probabilities. Differences in parameter counts and training stability therefore cannot be fully excluded as explanations for the observed improvements. Future work should compare parameter-matched models under equal training budgets and report confusion matrices for preparation, initiation, extension, and final hold to determine whether improvements are concentrated in particular phases.

### 5.2 Confidence gating and classroom robustness

Classroom keypoint errors cluster by body region and time. Rapid motion of the weapon-side wrist may cause blur, the non-weapon arm may be occluded by the trunk, and background students

may cause identity switches. Feeding coordinates directly into a model risks treating detection errors as actual movement. Simply deleting low-confidence keypoints may instead disrupt movement continuity. The soft gating used here attenuates low-confidence information without abruptly setting it to zero and is therefore theoretically better suited to continuous technical movements.

FPT+CPG retained a relative advantage under artificial missingness, suggesting that confidence gating can reduce the impact of discrete missed keypoint detections [8]. Nevertheless, random artificial missingness differs substantially from real classroom occlusion. Actual errors often occur over consecutive frames and depend on movement speed, camera position, and body region. The robustness findings should therefore be confined to the stress-test conditions used here and should not be taken as evidence of suitability for all complex classroom environments.

Future external testing should address three conditions: occlusion by joint region, separately examining missing wrists, elbows, shoulders, and lower limbs; continuous temporal occlusion, simulating invisibility over successive frames; and generalization across classrooms, camera positions, and lighting conditions. In addition to average performance, studies should report the proportion of attempts for which the system declines automated scoring. Reliability requires recognizing insufficient evidence rather than forcing a score for every video.

### 5.3 Expert criteria and assessment boundaries

Training loss or correlation alone cannot establish assessment validity. High correlation can coexist with systematic overestimation or underestimation, which is why we used ICC, absolute errors, and Bland–Altman analysis together. The combination of high ICC, small mean bias, and the reported limits of agreement allowed model validity to be examined through rank-order association, absolute differences, and individual-level error.

Mean expert scores are not an error-free gold standard. Raters may interpret trunk compensation, the final hold, or movement rhythm differently. Training directly on uncalibrated expert scores could encode these disagreements into the model. Clear scoring anchors, interrater reliability estimates, and retention of disputed cases for collective review are therefore needed. Videos showing large model–expert differences should undergo error analysis to distinguish pose-detection failures, phase-classification errors, rubric ambiguity, and fluctuations in expert ratings.

The assessment boundary is particularly important. The model analyzes one student's body keypoints and cannot directly observe the weapon tip, blade contact, or hits. It also lacks information about an opponent and therefore cannot assess distance control, choice of timing, or tactical effectiveness. High posture and timing scores indicate only that the stationary straight thrust was performed well under the specified camera configuration and rubric. They do not establish comprehensive *épée* proficiency or replace assessment in actual bouts.

### 5.4 Mechanisms supporting instruction

Conventional classroom feedback requires instructors to observe, remember, judge, and articulate an evaluation. The potential advantage of the intelligent system is to reduce time spent on repeated observation and preliminary error localization, allowing instructors to focus on explanation and practice design. Phase localization can translate a general description, such as poor coordination, into a specific issue, such as a preparatory swing during initiation or trunk compensation during extension. Component scores also give students trackable improvement targets.

The intelligent-feedback group showed greater movement-score gains and shorter feedback time, consistent with formative feedback that is timely, specific, and directed toward subsequent action [14]. The control group also improved, confirming that demonstration, conventional correction, and

repeated practice remain fundamental to skill learning. With pretest adjustment and equivalent conventional instruction, the findings support the capacity of intelligent tools to make instructor feedback more timely and better localized. They do not imply that the system produces educational benefits independently of an instructor.

The adjusted between-group difference of 6.08 points and gain-score effect size of  $d = 1.53$  indicate a difference in this sample beyond statistical significance alone. However, a minimum educationally important difference for the stationary straight-thrust score was not prespecified, and no retention test was conducted after feedback withdrawal. A single posttest advantage therefore cannot be equated with durable learning or long-term educational benefit. The experiment provides preliminary evidence of classroom effectiveness and efficiency; retention and transfer require further validation.

Because the study included only an immediate posttest after the eight-week intervention, it cannot establish that learning had stabilized. Score improvements may partly reflect familiarity with the rubric, short-term adaptation to system prompts, or additional attention. Future research should include delayed tests after feedback withdrawal and transfer tests using different camera positions or continuous movement combinations. It should also record whether students can identify errors without score prompts. Evidence of both retention and transfer would provide a stronger account of instructional benefit.

## 5.5 Implementation considerations and limitations

Classroom videos contain identifiable information about human movement. Deployment should therefore follow data-minimization principles: collect only the clips needed for assessment, use anonymous identifiers, restrict access, specify retention periods, and allow students to withdraw. Outputs should display confidence information and assessment boundaries; opaque total scores should not directly determine course grades. Applications involving students who differ in sex, body type, weapon-hand dominance, or disability status should examine errors separately to avoid unfairness from uniform thresholds.

The study has six further limitations. First, single-camera two-dimensional data cannot accurately represent movement in depth. Second, a single-university sample of beginners limits external validity. Third, expert rubrics introduce instructional conventions and rater preferences into the model. Fourth, model comparisons lacked strict parameter matching and complete ablation analyses. Fifth, full blinding of instructors and students was not possible, leaving risks of expectancy effects and group contamination. Sixth, only an immediate posttest after eight weeks was available, without retention or transfer testing. Future work could address these limitations through multiple cameras or three-dimensional reference data, participant-independent multicenter testing, expert calibration and stratified error analysis, parameter-matched ablations, cluster-randomized designs, and delayed testing.

The study links assessment indicators, model development, independent validation, classroom feedback, and instructional outcomes, providing integrated empirical evidence for intelligent recognition and precise assessment of movement technique. However, a single movement performed by one person under a fixed camera view constitutes component-level validation. It does not establish the general effectiveness of a complete system across multiple actions, camera positions, and interference from other people. This scope should remain explicit when applying the findings.

## 6. Conclusions

For the stationary straight thrust in university épée instruction, we developed an assessment

framework comprising four phases (preparation, initiation, extension, and final hold), two dimensions (body posture and movement timing), and ten secondary indicators. The resulting workflow integrates single-person COCO-17 keypoint extraction, phase feature encoding, keypoint-confidence gating, and multitask posture and timing scoring. Phase-specific and dimension-specific outputs, accompanied by confidence alerts, support instructor-led formative feedback. They are not intended to replace instructor judgment or comprehensive assessment of fencing competence.

FPT+CPG achieved an accuracy of 83.03%, a macro-F1 of 82.43%, and a scoring MAE of 3.85 points. Agreement between model total scores and mean expert scores yielded an ICC(2,1) of 0.925. In the teaching experiment, the intelligent-feedback group scored 6.08 points higher at posttest after pretest adjustment, and mean feedback time per attempt was reduced by 31.63 s. These findings support joint modeling of phase information and keypoint confidence and suggest that instructor-reviewed intelligent feedback can improve movement performance and feedback efficiency. Conclusions are restricted to beginners from one university performing the stationary straight thrust under a fixed camera view; generalization to other movements, camera positions, and complex classroom conditions requires external validation.

## Acknowledgement

This work was supported by the 2025 General Project of the Teaching Reform Special Program under the Chongqing Education Science 14th Five-Year Plan, “Research on intelligent recognition and precise assessment of technical movements in physical education using deep learning” (grant number K25ZG2110228).

This work was also supported by the Key Research Project on Teaching Reform in General Colleges and Universities of Hunan Province (No. 202502001052). Project: Research on Construction and Application of a Digital-Enabled Precision Teaching Evaluation System for Physical Education.

## References

- [1] Sun D, Song Y, Cen X, et al. Research progress of computer vision-based markerless sports motion capture technology. *Journal of Shanghai University of Sport*. 2021;45(9):70–85. doi:10.16099/j.sus.2021.09.007. [In Chinese].
- [2] Yu L, Du C, Yan Z, et al. Review of 2D human pose encoding and decoding methods: from the perspective of ambiguity mitigation. *Journal of Image and Graphics*. 2024;29(11):3319–3344. doi:10.11834/jig.230648. [In Chinese].
- [3] Huang Q, Cui J, Li C. A review of skeleton-based human action recognition. *Journal of Computer-Aided Design & Computer Graphics*. 2024;36(2):173–194. doi:10.3724/SP.J.1089.2024.2023-00358. [In Chinese].
- [4] Cust EE, Sweeting AJ, Ball K, et al. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *Journal of Sports Sciences*. 2019;37(5):568–600. doi:10.1080/02640414.2018.1521769.
- [5] Malawski F, Kwolek B. Recognition of action dynamics in fencing using multimodal cues. *Image and Vision Computing*. 2018;75:1–10. doi:10.1016/j.imavis.2018.04.005.
- [6] Malawski F, Krupa M, Kapela K. Automatic assessment of skill and performance in fencing footwork. In: *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications: VISAPP*. Setúbal: SCITEPRESS; 2024. p. 637–644. doi:10.5220/0012577600003660.
- [7] Lai C, Mo J, Xia H, et al. FACTS: fine-grained action classification for tactical sports. *arXiv [Preprint]*. 2024 Dec 21 [cited 2026 Jul 16]. arXiv:2412.16454. doi:10.48550/arXiv.2412.16454.
- [8] Liu C, Zang Y, Lan C. Research on multiplayer pose estimation in complex sports scenes. *Journal of University of Electronic Science and Technology of China*. 2024;53(6):930–939. doi:10.12178/1001-0548.2023233. [In Chinese].
- [9] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32(1):7444–7452. doi:10.1609/aaai.v32i1.12328.
- [10] Chen Y, Zhang Z, Yuan C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. p. 13359–13368. doi:10.1109/ICCV48922.2021.01311.
- [11] Duan H, Zhao Y, Chen K, et al. Revisiting skeleton-based action recognition. In: *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition. 2022. p. 2969–2978. doi:10.1109/CVPR52688.2022.00298.
- [12] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444. doi:10.1038/nature14539.
- [13] Fitts PM, Posner MI. *Human performance*. Belmont, CA: Brooks/Cole; 1967.
- [14] Hattie J, Timperley H. The power of feedback. *Review of Educational Research*. 2007;77(1):81–112. doi:10.3102/003465430298487.
- [15] Mishra P, Koehler MJ. Technological pedagogical content knowledge: a framework for teacher knowledge. *Teachers College Record*. 2006;108(6):1017–1054. doi:10.1111/j.1467-9620.2006.00684.x.
- [16] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012.
- [17] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310. doi:10.1016/S0140-6736(86)90837-8.
- [18] Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. *BMJ*. 2001;323(7321):1123–1124. doi:10.1136/bmj.323.7321.1123.