Comparison of Pose Estimation Models toward Nonverbal Feedback for Pre-service Teachers

Shota Shirasaka[1], Takahisa Imagawa[2], Shuichi Enokida[2]

[1] Fukuoka Institute of Technology, Japan
[2] Kyushu Institute of Technology, Japan

Abstract

Teachers' nonverbal behavior is a strong predictor of instructional effectiveness, yet manual observation is difficult to scale. Deep learning-based pose estimation offers automated extraction of body posture from classroom videos, but existing studies adopt single models without comparative evaluation. This exploratory, single-cohort study compared four pose estimation models—RTMPose-m (CNN/SimCC), ViTPose-b (Vision Transformer/heatmap), YOLOv8m-pose (CNN single-stage), and MediaPipe Pose (BlazePose, CPU)—applied to 27 micro-teaching videos (~294,000 frames) under two human detector conditions (RTMDet-tiny, YOLOv8n). Friedman tests revealed significant differences on all stability metrics (all $p < .001$). RTMPose-m achieved the highest detection rate ($M = .857$), while ViTPose-b showed the lowest normalized jitter and was the only model whose detection rate did not significantly decline during back-facing/board-writing intervals. The choice of human detector showed minimal differences in confidence-based output stability, with significant effects limited to temporal jitter for two models. These findings inform the design of automated nonverbal feedback systems in teacher education, where model choice may bear substantially on the reliability of downstream behavioral analysis.

Keywords: *pose estimation; teacher education; nonverbal feedback; classroom video analysis; micro-teaching*

1. Introduction

Teachers' nonverbal behavior—gestures, posture, movement, and body orientation—contributes substantially to instructional effectiveness. [1] showed that observers can predict teaching evaluations from 30-second silent video clips ($r = .76$), with accuracy remaining significant even for two-second clips ($r = .71$). Nonverbal immediacy is associated with student motivation [2], and nonverbal communication correlates strongly with achievement [3]. Systematic analysis of teachers' physical behavior could thus provide actionable feedback for professional development, particularly in micro-teaching exercises within teacher education programs.

However, manual observation is labor-intensive and difficult to scale. Deep learning-based pose estimation offers automated posture extraction from classroom video recordings. A systematic review identified 80 studies applying computer vision to classroom behavior recognition [4], and several studies have applied skeleton-based approaches to teacher behavior detection [5], [6], [7]. These studies typically adopt a single pose estimation model suited to their specific objective. Given that pose estimation models differ substantially in architecture and output characteristics [8], [9], a comparative evaluation under classroom-specific conditions—persistent occlusion, back-facing postures, and fixed-camera setups—could provide useful guidance for application-oriented researchers.

Automated analysis is also employed as a component of multi-source feedback designs intended to support reflection in micro-teaching. Three-point comparison feedback [10], which presents automated, peer, and self-evaluations side by side, frames discrepancies among sources as material for reflection, while noting that the educational value of such discrepancies may depend on whether they arise from differences in perspective or from measurement instability. Characterizing which body parts each model can stably measure under classroom conditions is therefore important for designing a reliable automated-feedback component in multi-source feedback. In teacher education, this technical reliability matters because unstable pose outputs can lead to misleading feedback about posture, gesture, or orientation, even when the downstream pedagogical interpretation is carefully designed.

This study compares four architecturally distinct pose estimation models applied to 27 micro-teaching videos under two detector conditions. We address:

- RQ1: How do models differ in detection stability?
- RQ2: How does performance change between front-facing and back-facing postures?
- RQ3: What are the practical trade-offs in processing speed?

- RQ4: To what extent does detector choice influence confidence-based output stability?

2. Related Work

Computer vision for classroom behavior analysis has expanded rapidly, with YOLO-based methods dominating and physical action recognition as the most common task [4]. Several studies have applied pose estimation to teacher behavior: [5] used OpenPose with graph convolutional networks for classifying five nonverbal behavior categories (81.6% accuracy); [6] used HRNet with rule-based indicators for six behavior types; [7] combined AlphaPose with Faster R-CNN for behavior monitoring. These studies advance specific aspects of classroom analysis using a pose estimation model chosen for the task at hand. Complementary work has examined adjacent components: [11] compared detection architectures for classroom interactions, and [12] evaluated an integrated body-language assessment system.

Pose estimation has advanced through multiple paradigms [8], [9]: top-down two-stage methods including heatmap-based ViTPose [13] and coordinate-classification RTMPose [14], single-stage YOLOv8-pose [15], and lightweight MediaPipe Pose [16]. These are typically evaluated on benchmarks such as MS COCO [17], but benchmark performance does not necessarily translate to applied settings [18], and temporal instability (jitter) remains a recognized challenge that per-frame metrics cannot capture [19].

A systematic comparison of multiple pose estimation models under classroom-specific conditions would complement these efforts by clarifying model selection trade-offs, characterizing behavior under domain-specific occlusion patterns, and informing downstream system design. This study examines four architecturally distinct models across 27 videos (~294,000 frames), evaluating detection stability, orientation robustness, speed, and detector dependency.

3. Methodology

3.1 Participants and Data

Twenty-seven third-year undergraduate students enrolled in a teacher education program participated in the study. As part of a course on Tokubetsu Katsudo (Special Activities in the Japanese curriculum), each student prepared a lesson plan and delivered a three-minute micro-teaching session on a topic of their choice. Sessions were video-recorded in December 2025 in a single classroom at 59.94 fps (1920 × 1080), with a whiteboard behind the speaker and a lectern partially occluding the lower body. This setup introduces domain-specific challenges for pose estimation, including persistent lower-body occlusion and frequent transitions between front-facing instruction and back-facing board-writing. The total dataset comprises 294,300 frames. Ethical approval was obtained from the institutional review board, and all participants provided informed consent. The same cohort and recordings have also been analyzed in Shirasaka et al. [10] under a different research question (an exploratory investigation of reflection on multi-source feedback). The present study focuses on the technical comparison of pose estimation models and does not analyze the participants' reflection narratives or cross-source evaluation comparisons reported in that paper.

3.2 Pipeline and Models

We adopted a top-down pipeline: human detection with single-object tracking, followed by per-frame pose estimation within the detected bounding box (Fig. 1). The tracker selects the person closest to a manually identified reference at video start, ensuring consistent identity across frames. This step was important because the classroom scene contains stable background objects and occasional non-target body parts, whereas the analysis requires a single continuous teacher trajectory rather than frame-wise independent detections. Two detectors were employed (RQ4): RTMDet-tiny [20] (45.6 ms/frame) and YOLOv8n [15] (21.5 ms/frame). Frames with zero detection confidence were not passed to pose estimators and were retained as zero-keypoint rows (< 0.1% for both detectors). Using a detector-first design also separated two questions: whether the pose estimator remained stable once a person region was supplied, and whether detector choice changed downstream stability. For the top-down models, the detected box was cropped and passed to the pose estimator; YOLOv8m-pose was run in bounding-box input mode to keep person regions comparable across models. Because these arrays were saved before analysis, all four stability metrics were computed from the same post-inference data structure rather than from framework-specific visualizations or logs.

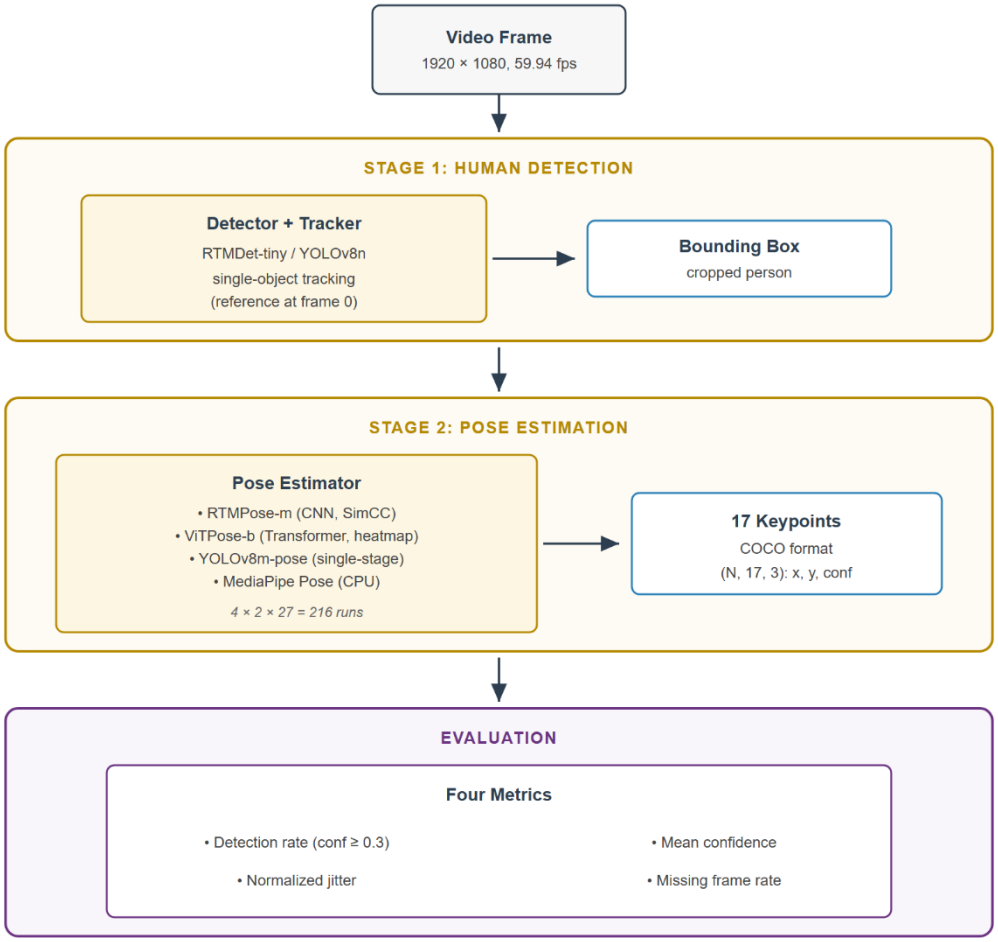

Fig. 1. Overall analysis pipeline from video frames to human detection, pose estimation, and the four stability metrics.

Four pose estimation models represented distinct paradigms (Table 1): RTMPose-m (CNN/SimCC), ViTPose-b (Vision Transformer/heatmap), YOLOv8m-pose (single-stage CNN, used in bounding-box input mode), and MediaPipe Pose (BlazePose heavy, CPU). Outputs were stored as 17-keypoint COCO-format arrays (N, 17, 3); MediaPipe's 33 landmarks were mapped to COCO by anatomical correspondence. The third channel was treated as a model-provided confidence-like score, with architecture-specific meanings (heatmap peak, SimCC softmax, YOLO keypoint confidence, and MediaPipe visibility) discussed in Section 5. Each model was applied under both detector conditions: 4 × 2 × 27 = 216 inference runs.

Table 1. Pose estimation models and pipeline performance.

Model	Type	Params	Pose (ms)	FPS w/ RTMDet	FPS w/ YOLOv8n
RTMPose-m	CNN (SimCC), 2-stage	13.6M	16.1	16.2	26.3
ViTPose-b	Transformer (Heatmap), 2-stage	~86M	16.3	16.2	26.1
YOLOv8m-pose	CNN single-stage	~26M	11.9	17.4	29.5
MediaPipe	CNN (BlazePose, heavy), CPU	N/A	60.9	9.4	12.8

Code for the analysis pipeline was developed with assistance from Claude (Anthropic); the first author reviewed and tested all code and takes responsibility for the content.

3.3 Evaluation Metrics

The goal is to assess model suitability for classroom behavior analysis as a component of nonverbal feedback systems, where macro-level posture information matters more than millimeter-precise joint locations. This annotation-free deployment screening evaluates confidence-based proxies for output

availability and temporal stability rather than pixel-level accuracy (PCK, AP), which would require ground-truth annotations across approximately five million keypoint instances.
Four metrics were computed per video × model × detector condition:

- Detection Rate: proportion of keypoints with confidence ?? 0.3. Sensitivity analysis at thresholds 0.1–0.5 indicated stable model rankings for 0.1–0.4.
- Mean Confidence: average confidence across all keypoints and frames.
- Normalized Jitter: median frame-to-frame Euclidean displacement of each keypoint, divided by torso length (mean of left and right shoulder-to-hip distances when both sides are valid, otherwise the available side) for scale normalization. Only consecutive frame pairs where both keypoints exceeded the threshold were included.
- Missing Frame Rate: proportion of frames with no keypoint reaching the threshold.

All four metrics are computed from the full saved keypoint arrays. Frames where the human detector returned zero confidence (< 0.1% of frames) are stored with zero keypoint output and therefore contribute to detection rate, mean confidence, and missing frame rate as undetected/missing; normalized jitter additionally requires both frames of a consecutive pair to satisfy the confidence and torso-length criteria, which naturally excludes these zero-filled frames.
These metrics map onto feedback needs: face-keypoint availability for orientation, wrist/elbow stability for gestures, missing frames for movement summaries, and latency for offline processing.

3.4 Front/Back Classification

To compare performance across orientations (RQ2), we first attempted automatic classification by four-model consensus on face keypoints (nose and eyes): frames were labeled “front” when all four models detected all three points with confidence ?? 0.3, and “back” when three or more models reported all three below 0.3. This labeled 80.9% as front but only 0.1% as back (F1 = .028), with 19.0% ambiguous. Because 73.3% of ambiguous frames fell within manually annotated board-writing intervals, we adopted a hybrid scheme. The first author annotated 35 board-writing intervals across 25 of 27 videos; within each interval, frames where all four models detected face keypoints were excluded as clearly front-facing (4.6%), and the remaining interval frames were labeled back-facing (41,048 frames, 13.9%). All other frames were labeled front-facing. Labels differed by only 25 frames (0.06%) between detector conditions.

3.5 Statistical Analysis

All analyses used the RTMDet condition unless noted. RQ1: Friedman test (N = 27, k = 4) with Kendall’s W; Nemenyi post-hoc. RQ2: Friedman on degradation rates (N = 25, excluding two videos without back-facing frames); per-model Wilcoxon with rank biserial. RQ3: Descriptive statistics. RQ4: Wilcoxon (N = 27) for 16 comparisons (4 models × 4 metrics); Holm correction.

4. Results

4.1 Overall Model Comparison (RQ1)

Friedman tests revealed significant differences on all metrics (Table 2). Detection rate showed the strongest effect (W = .971): RTMPose-m ranked highest (M = .857), followed by ViTPose-b (.808), YOLOv8m-pose (.768), and MediaPipe (.710). All six Nemenyi pairwise comparisons were significant (all p < .05).

Table 2. Descriptive statistics and Friedman test results for the four stability metrics (N = 27, k = 4). Values are Mean (SD).

Metric	RTMPose-m	ViTPose-b	YOLOv8m-pose	MediaPipe	Chi2	W
Detection Rate	.857 (.029)	.808 (.038)	.768 (.042)	.710 (.034)	78.64	.971
Mean Confidence	.636 (.027)	.686 (.029)	.703 (.035)	.685 (.035)	56.29	.695
Norm. Jitter	.0055 (.0010)	.0050 (.0013)	.0081 (.0017)	.0067 (.0012)	75.58	.933
Missing Rate	.00053 (.00131)	.00038 (.00111)	.00191 (.00229)	.00165 (.00240)	38.50	.475

All Friedman tests: df = 3, p < .001.

Mean confidence showed a reversed top ranking: YOLOv8m-pose was highest (.703) and RTMPose-m lowest (.636), consistent with architecture-specific confidence distributions. Sensitivity analysis showed stable detection-rate rankings across thresholds 0.1–0.4 (all W > .90), with RTMPose-m and ViTPose-b reversing only at 0.5. Temporal stability formed two groups: two-stage models (ViTPose-b, RTMPose-m; jitter ~.005) were more stable than MediaPipe (~.007) and YOLOv8m-pose (~.008), with significant between-group but not within-group differences; missing frame rate followed the same pattern.
Per-keypoint analysis (Fig. 2) showed that ankles were detected in at most 5% of frames across models, consistent with lower-body occlusion by the lectern. Knees varied substantially under the same condition (RTMPose-m: 81%; MediaPipe: <10%), and MediaPipe also showed lower elbow and wrist detection rates (70–81%) than the other models (95%+).

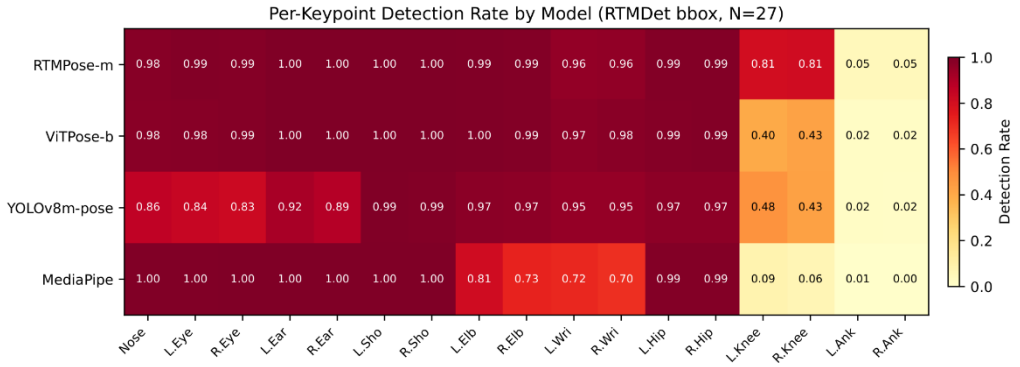

Fig. 2. Per-keypoint detection rate by model (mean across 27 videos, RTMDet condition). Ankles were near zero across models because the lectern physically occluded the lower body; model differences were largest for knees and wrists.

4.2 Front-Facing vs. Back-Facing Performance (RQ2)

Degradation rates differed significantly across models for all three metrics (Friedman, all $p < .001$; Table 3). Two-stage models showed minimal detection rate degradation (RTMPose-m: 2.2%, ViTPose-b: 1.8%), while YOLOv8m-pose dropped 17.0% and MediaPipe 13.7%. Per-model Wilcoxon tests indicated significant front-to-back differences for all models except ViTPose-b ($p = .127$, $r_b = .35$)—the only model whose detection rate did not significantly decline during back-facing/board-writing intervals. Mean confidence declined significantly for all models (all $p < .001$), with YOLOv8m-pose showing the largest drop (24.8%). Jitter worsened for all models, most severely for YOLOv8m-pose (66.5%, SD = 61.6%).

Table 3. Detection rate, confidence, and normalized jitter for front-facing and back-facing frames, with degradation rates (N = 25). Values are Mean (SD).

Metric	Model	Front	Back	Degradation (%)
Detection Rate	RTMPose-m	.860 (.030)	.840 (.041)	+2.2 (4.9)
	ViTPose-b	.810 (.041)	.796 (.045)	+1.8 (6.0)
	YOLOv8m-pose	.788 (.044)	.654 (.051)	+17.0 (7.3)
	MediaPipe	.725 (.039)	.623 (.037)	+13.7 (7.8)
Mean Confidence	RTMPose-m	.649 (.025)	.555 (.039)	+14.6 (5.4)
	ViTPose-b	.698 (.029)	.616 (.041)	+11.6 (6.0)
	YOLOv8m-pose	.729 (.032)	.549 (.043)	+24.8 (6.0)
	MediaPipe	.701 (.039)	.590 (.028)	+15.5 (7.0)
Norm. Jitter	RTMPose-m	.0053 (.0011)	.0069 (.0015)	+35.9 (50.5)
	ViTPose-b	.0049 (.0014)	.0065 (.0016)	+48.8 (74.2)
	YOLOv8m-pose	.0077 (.0017)	.0122 (.0030)	+66.5 (61.6)
	MediaPipe	.0065 (.0011)	.0099 (.0052)	+52.8 (84.9)

4.3 Practical Feasibility (RQ3)

Replacing RTMDet-tiny with YOLOv8n reduced pipeline latency by 26–41% (Table 1). The fastest combination was YOLOv8n + YOLOv8m-pose (33.9 ms, 29.5 FPS); YOLOv8n + RTMPose-m (38.0 ms, 26.3 FPS) balanced speed and output availability. MediaPipe, though slowest (78–107 ms total), is the only pose stage that runs on CPU. All combinations are feasible for offline analysis of three-minute videos.

4.4 Detector Dependency (RQ4)

Of 16 Wilcoxon tests, only two reached significance after Holm correction: jitter for YOLOv8m-pose (RTMDet: .00806; YOLOv8n: .00813; $p = .006$, $rb = -.74$) and MediaPipe (RTMDet: .00674; YOLOv8n: .00703; $p < .001$, $rb = -.95$). Detection rate and confidence showed no significant differences (all $p_{holm} > .19$, mean differences < .003). The statistically significant jitter differences were small in absolute terms (≤ .0003), suggesting that bounding-box fluctuations can propagate to temporal stability without materially changing per-frame proxy metrics.

5. Discussion

The results point to a task-dependent model-selection strategy rather than a single best model. RTMPose-m maximized output availability; ViTPose-b minimized jitter and showed the smallest, non-significant detection-rate decline; YOLOv8m-pose prioritized speed; and MediaPipe provided a CPU-based option. For offline micro-teaching analysis, model choice can be guided by the behavioral features of interest rather than benchmark performance alone.

Two broader points follow. First, confidence distributions differ systematically between SimCC and heatmap architectures, so confidence-based metrics are not directly comparable across architectures. Researchers reporting confidence values benefit from specifying the architectural class to support cross-study interpretation. Second, the selective detector effect on jitter, despite minimal effects on per-frame metrics, shows that bounding-box fluctuations can propagate to temporal stability. Evaluations targeting downstream temporal analyses may benefit from including jitter-aware metrics rather than relying solely on per-frame accuracy [19].

The failure of automatic front/back classification ($F1 = .028$) highlights that current models cannot reliably distinguish orientation from detection failure in classroom settings, motivating domain-specific model development.

The per-keypoint heatmap (Fig. 2) further shows that models differ markedly in which body parts they detect reliably. Near-zero ankle detection reflects the filming configuration rather than model capability; analyses requiring lower-body information would require repositioning the camera or lectern. Within the visible body, model differences for knees (.09–.81) and wrists (.70–.97) indicate that model selection may benefit from considering the body parts relevant to the intended analysis.

Persistent lower-body occlusion also complicates confidence interpretation. In the present data, this pattern is consistent with the possibility that heatmap peaks decrease more sharply under weak local visual evidence, whereas SimCC softmax scores over discretized coordinate distributions may retain moderate values even when the predicted location is partly inferred from pose context. This is not evidence of a general SimCC advantage for occluded regions: a supplementary inspection of face-keypoint outputs showed higher ViTPose-b confidence in both front-facing and back-facing intervals. The RTMPose-m advantage for lower-body keypoints therefore likely reflects confidence normalization under weak evidence, and the structural plausibility of such predictions requires ground-truth validation.

6. Limitations and Future Work

6.1 Limitations

This evaluation relies on confidence-based metrics rather than ground-truth annotations. Because these scores have architecture-specific meanings, detection rate and mean confidence should be interpreted as output-availability proxies rather than calibrated accuracy measures. While sensitivity analysis indicated stable rankings across thresholds, the absence of pixel-level accuracy measures means we cannot rule out the possibility that maintained confidence under occlusion corresponds to structurally implausible joint positions. All recordings used a single camera in a single classroom in a Japanese teacher education program. Generalizability to other educational contexts—different cultural settings, classroom configurations, age groups, or instructional formats—requires verification. The front/back classification depended on manual annotation of board-writing intervals with model-output filtering and is better regarded as an interval-level approximation than as frame-level orientation ground truth; fully

automatic orientation detection remains an open problem. The micro-teaching format also differs from authentic classroom instruction, and behavioral patterns observed here may not fully transfer to in-service teaching. Scale normalization of jitter relies on torso length computed from shoulder and hip keypoints; under persistent lower-body occlusion, uncertainty in hip-keypoint estimates may introduce minor noise into the normalization, though absolute jitter rankings remained consistent across models. Given the descriptive, single-cohort design, the findings should be interpreted as context-bound exploratory evidence for model selection rather than as definitive performance claims.

6.2 Future Work

Three directions follow from these findings. First, partial ground-truth annotation on a representative subset of frames could validate confidence values under persistent occlusion, particularly for SimCC-based lower-body outputs. Second, architecture-aware confidence interpretation—distinguishing outputs grounded in visual evidence from those reconstructed from pose context—could improve the reliability of downstream behavioral analysis. Third, evaluations across multiple classroom configurations and instructional formats would help characterize the boundary conditions of the present findings.

7. Conclusion

This exploratory, single-cohort study compared four architecturally distinct pose estimation models across 27 classroom videos under two detector conditions. All confidence-based stability metrics showed significant model differences, with detection rate ranking nearly perfectly consistent ($W = .971$). Two-stage models ranked higher on output availability and temporal consistency; detector choice had minimal association with the selected proxy metrics except for jitter.

These findings should therefore be read as an evaluation of the technical layer that supports automated feedback, not as an automated assessment of teaching quality itself. For automated nonverbal feedback in teacher education, ViTPose-b is a cautious choice when board-writing or orientation analysis is involved because its detection-rate decline was small and not significant. MediaPipe remains viable when GPU acceleration is unavailable, and RTMPose-m with YOLOv8n offers the highest detection rate at practical speed for balanced offline use. These findings can inform the automated-feedback component of multi-source feedback designs [10] and support selecting classroom pose-estimation pipelines by the body parts and behavioral features needed for feedback, not by benchmark performance alone.

Beyond model selection, two methodological considerations may apply across studies. First, confidence values are not directly comparable across architectural classes; the same numerical value can reflect different generation processes (heatmap peak vs. SimCC softmax vs. visibility score), so cross-study comparisons benefit from architectural disclosure. Second, detector and pose-estimator behavior can decouple on temporal-stability metrics even when per-frame metrics agree, so evaluations supporting downstream behavioral analysis benefit from including jitter-aware indicators alongside per-frame accuracy.

REFERENCES

[1] N. Ambady and R. Rosenthal, “Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness,” Journal of personality and social psychology, 1993, doi: 10.1037/0022-3514.64.3.431.

[2] J. Frenkel, A. Cajar, R. Engbert, and R. Lazarides, “Exploring the impact of nonverbal social behavior on learning outcomes in instructional video design,” Scientific Reports, 2024, doi: 10.1038/s41598-024-63487-w.

[3] F. Bambaeeroo and N. Shokrpour, “The impact of the teachers’ non-verbal communication on success in teaching,” Journal of Advances in Medical Education & Professionalism, 2017.

[4] Q. Liu, X. Jiang, and R. Jiang, “Classroom behavior recognition using computer vision: A systematic review,” Sensors, 2025, doi: 10.3390/s25020373.

[5] S. Pang, S. Lai, A. Zhang, Y. Yang, and D. Sun, “Graph convolutional network for automatic detection of teachers’ nonverbal behavior,” Computers and Education: Artificial Intelligence, 2023, doi: 10.1016/j.caeai.2023.100174.

[6] Y. Ye, J. Wang, P. He, J. Nie, J. Xiong, and H. Gao, “An action analysis algorithm for teachers based on human pose estimation,” Computers and Electrical Engineering, 2023, doi: 10.1016/j.compeleceng.2023.108915.
[7] J. Huang, H. Hashim, H. Norman, M. H. Zaini, and X. Zhang, “Automatic detection of teacher behavior in classroom videos using AlphaPose and faster r-CNN algorithms,” PeerJ Computer Science, 2025, doi: 10.7717/peerj-cs.2933.
[8] C. Zheng et al., “Deep learning-based human pose estimation: A survey,” ACM Computing Surveys, 2023, doi: 10.1145/3603618.
[9] Z. Gao, J. Chen, Y. Liu, Y. Jin, and D. Tian, “A systematic survey on human pose estimation: Upstream and downstream tasks, approaches, lightweight models, and prospects,” Artificial Intelligence Review, 2025, doi: 10.1007/s10462-024-11060-2.
[10] S. Shirasaka, T. Imagawa, and S. Enokida, “Exploring pre-service teachers’ reflection on nonverbal behavior in microteaching through three-point comparison feedback,” Education Sciences, vol. 16, p. 760, 2026, doi: 10.3390/educsci16050760.
[11] A. Almubarak, A. Aljohani, N. Alolayan, and M. Aldharrab, “An AI-powered framework for assessing teacher performance in classroom interactions: A deep learning approach,” Frontiers in Artificial Intelligence, 2025, doi: 10.3389/frai.2025.1553051.
[12] E. Dimitriadou and A. Lanitis, “Evaluating the impact of an automated body language assessment system,” Education and Information Technologies, 2025, doi: 10.1007/s10639-024-12931-5.
[13] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “ViTPose: Simple vision transformer baselines for human pose estimation,” in Advances in neural information processing systems (NeurIPS), 2022.
[14] T. Jiang et al., “RTMPose: Real-time multi-person pose estimation based on MMPose.” arXiv preprint arXiv:2303.07399, 2023. doi: 10.48550/arXiv.2303.07399.
[15] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8.” Ultralytics YOLOv8 (software), 2023.
[16] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, “BlazePose: On-device real-time body pose tracking.” arXiv preprint arXiv:2006.10204, 2020.
[17] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in European conference on computer vision (ECCV), 2014. doi: 10.1007/978-3-319-10602-1_48.
[18] F. Roggio, B. Trovato, M. Sortino, and G. Musumeci, “A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review,” Heliyon, 2024, doi: 10.1016/j.heliyon.2024.e39977.
[19] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, “SmoothNet: A plug-and-play network for refining human poses in videos,” in European conference on computer vision (ECCV), 2022. doi: 10.1007/978-3-031-20065-6_36.
[20] C. Lyu et al., “RTMDet: An empirical study of designing real-time object detectors.” arXiv preprint arXiv:2212.07784, 2022. doi: 10.48550/arXiv.2212.07784.