
2. CHAPTER 2: LITERATURE REVIEW

2.1. Theoretical Perspective

2.1.1. Foundational Hip-Hop Movements Taxonomy

Hip-hop dance encompasses a wide variety of movement styles, each with distinct technical demands and cultural roots. To provide a coherent analytical foundation for this study, it is useful to first situate the three movements under investigation within the broader landscape of hip-hop movement vocabulary. Scholars and practitioners have grouped foundational hip-hop moves into several pedagogically recognized categories (Art Gharana Academy, 2025; Fink, 2025). Understanding this taxonomy clarifies why the present research deliberately limits its scope to three specific movements, and it provides context for future system extensions.

The first category, groove and foundation moves, encompasses the rhythmic pulse that underpins all hip-hop performance. The Bounce, a repetitive vertical oscillation produced primarily through knee flexion, exemplifies this category, as does the Rock and the Head Nod. These movements are foundational precisely because they train the body to internalize musical meter before any footwork or isolation is introduced. From a computational standpoint, groove moves are well-suited to rhythm-based analysis, since their periodicity renders them amenable to frequency-domain techniques such as Fast Fourier Transform (FFT) correlation.

Footwork-based moves form the second category and introduce weight transfer, spatial directionality, and foot placement as core technical concerns. The Side Step, the Two-Step, and the Running Man require the dancer to shift body weight laterally or anteroposteriorly in time with the music. These movements present a particular challenge for pose-based analysis because ankle and foot landmarks are frequently occluded when limbs cross or when the dancer's lower body moves out of the primary camera frame. This limitation reinforces the importance of applying a landmark visibility threshold when computing skeletal features, as is done in the present system.

Body isolation and control moves constitute the third category and demand a high degree of spatial precision, requiring the dancer to mobilize specific joints while

keeping adjacent body segments relatively still. Chest pops, shoulder isolations, and body waves are representative examples. Evaluating these movements algorithmically requires a multi-joint skeletal representation, since any system that monitors only a single landmark would be unable to distinguish a genuine chest isolation from general upper-body displacement.

The fourth category comprises style-based techniques derived from distinct hip-hop sub-genres: Popping, Locking, and Boogaloo. These movements require advanced neuromuscular control and are typically mastered only after the foundational categories have been internalized. Algorithmically, they pose the most significant challenge because they involve abrupt velocity changes - rapid muscle contractions and sudden freezes - that demand a temporal alignment algorithm capable of handling non-linear time distortions between reference and student sequences.

Although the taxonomy above describes the full breadth of hip-hop movement vocabulary, this research deliberately narrows its scope to three foundational movements: Bouncing, Stepping, and Sliding. This selection was validated in consultation with an expert dance coach with two decades of professional teaching experience, who confirmed that these three movements represent the essential technical building blocks for beginner-level hip-hop learners.

Each movement also lends itself to distinct, quantifiable biomechanical criteria: vertical knee oscillation for Bouncing, foot-strike timing and lateral weight transfer for Stepping, and horizontal displacement linearity with hip stability for Sliding - making them appropriate targets for computational motion analysis in a lightweight, web-based educational context. The remaining movement categories identified in this taxonomy are acknowledged as directions for future system extension.

2.1.2. Wiraga and Wirama: A Two-Pillar Evaluation Framework

The evaluation philosophy of the present system draws upon Indonesian classical dance theory, which articulates three interconnected dimensions of performance quality: Wiraga, Wirama, and Wirasa (Robby, 2025). While these concepts originate within a distinct cultural and artistic tradition, the underlying constructs - bodily correctness, rhythmic synchronization, and expressive feeling - are

sufficiently universal to generalize to the assessment of any choreographed movement form, including contemporary hip-hop. Adopting this framework provides a principled structure for translating raw computational measurements into rubric-aligned performance scores, giving the system's output an interpretive clarity that a purely statistical metric would not provide.

Wiraga, derived from the Javanese root raga, meaning body - denotes the correctness, control, and spatial fidelity of physical form. Within the computational pipeline of the present system, this dimension is operationalized through angle-based kinematic feature extraction: per-frame joint angles are computed from a 13 analytical joint-angle features using the Law of Cosines applied to BlazePose landmark coordinates, and the resulting angular sequences are compared between the student and the reference instructor across three anatomical regions - arms, legs, and torso. A high Wiraga score therefore indicates that the learner has faithfully reproduced the reference posture, irrespective of the precise moment in the music at which the movement was performed. This separation of form assessment from timing assessment is intentional and reflects the pedagogical understanding that a beginner may execute a movement correctly in shape but inconsistently in timing, or vice versa.

Wirama, from irama, meaning rhythm or tempo - denotes the degree to which a dancer's movement is synchronized with the musical beat. In the present system, this dimension is operationalized through a two-stage temporal alignment pipeline. First, FFT-based audio cross-correlation is applied to identify the global temporal offset between the student's recording and the reference video, correcting for differences in start time that naturally arise when learners record themselves in unsupervised home environments. Second, the FastDTW algorithm is applied to the resulting aligned skeletal sequences to compute movement similarity in a manner that accommodates natural human timing flexibility - that is, the tendency of dancers to lead or lag slightly behind the beat without losing rhythmic intent. Together, these techniques allow Wirama to be measured as an independent dimension of performance quality, yielding actionable timing feedback that is distinct from feedback about movement form.

The third traditional dimension, Wirasa - from rasa, meaning feeling or sensibility - describes the expressive and emotional quality of a performance.

Expressive quality, however, cannot be reliably inferred from two-dimensional skeletal landmark data alone; its assessment requires facial affect analysis, dynamic effort cues, breath control information, and stylistic intent that are beyond the scope of pose estimation. Accordingly, Wirasa is deliberately excluded from the automated scoring pipeline and reserved as a direction for future multimodal system extensions. The present system therefore operates as a balanced two-pillar evaluation framework in which Wiraga and Wirama each contribute equally - 50% - to the final composite score, an apportionment validated by a certified dance instructor with twenty years of professional teaching experience. This weighting reflects the pedagogical consensus that for beginner learners, form and timing are the two most foundational competencies, and that expressive quality naturally develops only after both have been internalized.

Wirasa remains important in dance because it concerns expressive quality, emotional communication, stylistic intention, and the dancer's ability to give meaning to a movement. However, the present system intentionally excludes Wirasa because the available input is limited to two-dimensional pose information derived from BlazePose. The 33 BlazePose pose landmarks can describe body location, approximate posture, and movement trajectory, but they do not provide reliable information about facial nuance, gaze, muscle tension, breath control, emotional delivery, or culturally specific stylistic intention.

Therefore, including Wirasa in the current automated score would create a misleading impression of measurement accuracy. The system is designed to score only what it can observe and calculate with reasonable transparency: physical form through Wiraga and timing synchronization through Wirama. This decision is not intended to reduce the importance of expression in dance. Instead, it establishes a clear boundary between measurable technical features and expressive qualities that require richer multimodal information. Future work may explore partial proxies for Wirasa through facial-expression analysis, body-energy dynamics, acceleration contrast, movement amplitude variation, music interpretation, and expert annotations. However, such extensions would require a larger dataset and additional validation before being used as part of an automated score.

2.1.3. Human Pose Detection with BlazePose (MediaPipe)

The computational backbone of the present system is the BlazePose model, a real-time human pose estimation architecture developed by Google and distributed as part of the MediaPipe framework (Bazarevsky et al., 2020). BlazePose employs a two-stage detector-tracker pipeline: a lightweight body detector first localizes a region of interest around the human figure, after which a dedicated pose estimation network regresses the positional coordinates and visibility scores of 33 anatomical landmarks within that region. This architecture substantially reduces computational overhead compared to earlier full-image approaches, enabling inference at real-time frame rates on consumer-grade hardware without requiring dedicated GPU acceleration - a critical requirement for a system intended to operate as a browser-accessible web application.

A comparative evaluation of leading pose estimation frameworks - including OpenPose and AlphaPose - conducted by S. Lee et al. (2023) demonstrated that MediaPipe Pose achieves competitive landmark accuracy while maintaining substantially lower computational demands, making it particularly well-suited for deployment scenarios where processing resources are constrained. Mroz et al. (2021) similarly confirmed that BlazePose performs reliably for single-person tracking tasks under varied lighting conditions, though they noted accuracy degradations when joints are severely occluded or when the subject moves near the boundaries of the camera frame. These documented limitations directly informed two design decisions in the present system: the restriction to single-person evaluation and the application of a landmark visibility threshold of 0.60, below which landmark coordinates are excluded from kinematic computation to prevent low-confidence pose estimates from distorting the final assessment scores.

Of the 33 landmarks produced by BlazePose, the present system does not treat all landmarks as equally relevant to the selected beginner hip-hop movements. Instead, it derives 13 analytical joint-angle features from the landmark set: left and right elbow, shoulder, wrist, hip, knee, and ankle angles, together with one spine/posture angle. These analytical features are generated from connected landmark triplets using the Law of Cosines. Therefore, the term “13 analytical joint-angle features” in this study refers to 13 kinematic features used for analysis, rather than only 13 isolated raw landmark points.

This feature selection was based on the biomechanical requirements of the three target movements. Bouncing is primarily represented by knee flexion and lower-body vertical movement. Stepping is mainly represented by ankle motion, knee movement, and lower-body weight-transfer patterns. Sliding requires hip, ankle, and spine/posture information to represent lateral movement range and body stability. In contrast, facial landmarks, ear landmarks, hand-detail landmarks, and foot-detail landmarks beyond those required to form ankle angles do not directly contribute to the core scoring criteria for these three movements.

Using a focused feature set also reduces the effect of peripheral landmark noise and lowers computational complexity during video processing. This is important because the system is designed as a lightweight web-based prototype that operates on consumer hardware. The selected features remain sufficient for the current scope because they represent the body regions most relevant to the expert-defined criteria of Bouncing, Stepping, and Sliding. More complete landmark utilization may be required in future work for additional movement styles, upper-body isolations, hand-based choreography, facial-expression analysis, or expressive dance evaluation (Kim et al., 2023).

Table 2-1 Mapping of BlazePose Landmarks to the 13 Analytical Features

No.	Analytical Feature	BlazePose Landmark Components	Landmark IDs	Relevance to Current System
1	Left elbow angle	Left shoulder - left elbow - left wrist	11, 13, 15	Arm position and elbow control
2	Right elbow angle	Right shoulder - right elbow - right wrist	12, 14, 16	Arm position and elbow control
3	Left shoulder angle	Left hip - left shoulder - left elbow	23, 11, 13	Upper-body posture and arm placement
4	Right shoulder angle	Right hip - right shoulder - right elbow	24, 12, 14	Upper-body posture and arm placement

5	Left wrist angle	Left elbow - left wrist - left index	13, 15, 19	Wrist direction and arm extension
6	Right wrist angle	Right elbow - right wrist - right index	14, 16, 20	Wrist direction and arm extension
7	Left hip angle	Left shoulder - left hip - left knee	11, 23, 25	Weight transfer and lower-body posture
8	Right hip angle	Right shoulder - right hip - right knee	12, 24, 26	Weight transfer and lower-body posture
9	Left knee angle	Left hip - left knee - left ankle	23, 25, 27	Main feature for Bouncing depth
10	Right knee angle	Right hip - right knee - right ankle	24, 26, 28	Main feature for Bouncing depth
11	Left ankle angle	Left knee - left ankle - left foot index	25, 27, 31	Foot placement and lower-body movement
12	Right ankle angle	Right knee - right ankle - right foot index	26, 28, 32	Foot placement and lower-body movement
13	Spine/posture angle	Nose - midpoint of shoulders - midpoint of hips	0, 11, 12, 23, 24	Torso alignment and stability

2.1.4. View-Invariant Kinematic Feature Extraction

A fundamental challenge in video-based motion assessment is that raw skeletal landmark coordinates - expressed as pixel positions within the image plane - are inherently sensitive to extrinsic recording conditions: the distance between the subject and the camera, the resolution of the recording device, and the perspective angle all affect the absolute values of x and y coordinates produced by the pose estimator. Two recordings of the same movement performed identically but captured at different distances would yield substantially different coordinate values, rendering direct coordinate comparison unreliable as an assessment criterion.

To address this limitation, the present system derives its primary kinematic features not from raw landmark positions but from joint angles computed via the Law of Cosines applied to triplets of adjacent skeletal landmarks. Because interior joint angles are determined by the geometric relationship between connected body segments rather than by their absolute position in the image, they remain approximately invariant to changes in recording distance and camera perspective. S. ; W. R. Lee (2023) provide empirical support for this approach, demonstrating that angular kinematics yield more consistent and reproducible pose assessments across varied recording environments than coordinate-based features. This property is especially important for a system intended for use in uncontrolled home recording environments, where students are unlikely to maintain a fixed camera distance or viewing angle.

The extracted joint angles are partitioned across three anatomical regions - arms, legs, and torso - to enable region-specific diagnostic feedback. A student whose leg angles align well with the reference but whose torso angles show significant deviation will receive targeted corrective guidance specific to upper-body form, rather than a single aggregated score that obscures the source of the error. This decomposed feature representation thus serves both the scoring function and the semantic feedback function, bridging the gap between quantitative kinematic measurement and qualitative pedagogical guidance.

2.1.5. Temporal Alignment: FastDTW and Audio Cross-Correlation

A persistent methodological challenge in movement comparison systems is that two recordings of the same dance sequence will rarely be precisely synchronized in time. Even when a student and a reference instructor perform the same movement to the same music, differences in recording start time, personal pacing preferences, and natural human timing variability mean that a naive frame-by-frame comparison of their skeletal sequences will yield poor alignment quality. Addressing this challenge requires both a global temporal alignment strategy - to correct for fixed start-time offsets - and a local temporal alignment strategy - to accommodate the elastic, non-linear time distortions introduced by individual performance style.

The global alignment problem is addressed through FFT-based audio cross-

correlation. By computing the cross-correlation function of the audio signals extracted from the student and reference recordings in the frequency domain, the system identifies the temporal lag that maximizes signal agreement. This lag is then used to trim or pad the video sequences before any skeletal comparison is performed, ensuring that the pose analysis begins from a synchronized starting point. This step is particularly important for a self-assessment tool used in home recording environments, where students are unlikely to begin recording at precisely the same moment as the reference video commences.

The local alignment problem is addressed through FastDTW, an $O(N)$ approximation of the classical Dynamic Time Warping algorithm. DTW-based alignment has been established as an effective technique for comparing time-series movement data that are similar in shape but vary in speed, as demonstrated by (Niu, 2024) in the context of sports posture analysis. Unlike simple Euclidean distance measures, FastDTW constructs an optimal warping path between two sequences, allowing one to be non-linearly stretched or compressed to best match the other. In the context of dance assessment, this means that a student who performs a stepping sequence slightly slower than the reference instructor - without changing the underlying form - receives a more accurate similarity score than would be yielded by frame-synchronous comparison. Wang et al. (2021) applied a similar DTW-based approach to dance performance evaluation and demonstrated strong agreement between algorithmic scores and instructor ratings, supporting the validity of this technique for the present application.

2.1.6. Pedagogical Principles of Semantic Corrective Feedback

Numerical performance scores alone are insufficient as educational outputs in a self-assessment system intended for beginner learners. Research in human-computer interaction and educational technology consistently demonstrates that users are better able to improve their performance when feedback is specific, actionable, and expressed in natural language rather than in abstract numerical terms (Yu et al., 2024). This principle is particularly salient in movement-based learning contexts, where the learner must interpret a computational judgment and translate it into a physical correction without the guidance of a present instructor.

Motivated by these findings, the present system incorporates a semantic corrective feedback module that translates quantified kinematic deviations into natural-language instructional cues. When the system detects, for example, that a student's knee angle during the Bouncing movement deviates significantly from the reference across multiple frames, it generates a targeted text message - such as "Try to bend your knees more consistently with each bounce" - rather than simply reporting a low score for the legs region. This design prioritizes explainability: the system's outputs are intended to be interpretable by a learner with no prior knowledge of biomechanics or computer vision. Alfredo et al. (2024) underscore the importance of human-centered design in AI-assisted learning systems, noting that learner trust and engagement are substantially higher when algorithmic outputs are accompanied by intelligible explanations.

It bears emphasis that the semantic feedback generated by the present system is a support tool, not a replacement for professional dance instruction. The system is designed to provide beginners with structured self-assessment opportunities when access to a qualified instructor is limited - for instance, during home practice sessions or in educational institutions without dedicated dance faculty. The feedback messages are derived from predefined rule-based mappings between joint angle deviation thresholds and pedagogically appropriate corrective cues, reviewed and validated by an experienced instructor. This bounded approach ensures that the system's guidance remains within a verified pedagogical scope rather than attempting to generate open-ended instructional content beyond its design parameters. Kang et al. (2023) argue that AI tools in dance education are most effective when positioned as facilitators of teacher-guided learning rather than autonomous instructors, a framing that is consistent with the design philosophy of the present system.

2.2. Review of Previous Studies

The following review examines six closely related studies that collectively establish the technical and pedagogical foundation for the present research. Each study is discussed in terms of its methodological contributions, documented limitations, and specific relevance to the design decisions made in the proposed system. Taken together, these works demonstrate that while the individual components of the proposed system - namely pose estimation, angular feature extraction, DTW-based temporal alignment,

and semantic feedback - have been validated in isolation and in closely related domains, their integration within a lightweight, browser-accessible educational platform targeted at beginner hip-hop learners remains underexplored.

2.2.1. BlazePose: On-Device Real-Time Body Pose Tracking

Bazarevsky et al. (2020) introduced BlazePose, a compact, single-person body pose estimation model designed to operate in real time on mobile devices. The architecture employs a two-stage pipeline in which a lightweight body detector first identifies a person-bounding region of interest, followed by a regression-based landmark network that estimates the three-dimensional positions and visibility scores of 33 anatomical keypoints. By constraining the estimation network to operate within the detected region rather than processing the full image, BlazePose achieves substantially lower latency than full-image approaches such as OpenPose, without sacrificing meaningful landmark accuracy for single-person tracking tasks.

A notable feature of the model is its per-landmark visibility score, which provides a confidence estimate for each keypoint independent of its positional prediction. This score enables downstream systems to selectively discard low-confidence landmark estimates (such as those arising from self-occlusion during complex movements) without discarding the entire pose frame. This design decision directly informs the 0.60 visibility threshold applied in the present system, which excludes any landmark whose confidence falls below this value from the kinematic computation pipeline. Despite its strengths, the authors acknowledge that BlazePose is optimized for single-person scenarios and may degrade in accuracy when the subject moves to the periphery of the camera frame or when clothing makes body contours difficult to distinguish. These limitations are consistent with the constraints of recording in uncontrolled home environments, and they reinforce the need for clear guidance to users regarding camera setup during self-assessment sessions.

2.2.2. Human Pose Estimation Using MediaPipe Pose and Optimization

Kim et al. (2023) investigated methods for improving the precision and temporal consistency of MediaPipe Pose landmark estimates by incorporating a humanoid biomechanical model as a post-processing constraint. Their approach applies

a kinematic optimization layer over the raw MediaPipe outputs, filtering high-frequency noise from landmark trajectories and enforcing anatomically plausible joint angle limits. The study demonstrated measurable improvements in joint angle measurement accuracy and a reduction in landmark jitter across successive frames, particularly for extremity joints such as the wrists and ankles, which tend to exhibit the greatest estimation variability.

The relevance of this work to the present study lies primarily in its validation of joint angle measurement as the core feature representation for pose-based assessment. Kim et al. (2023) confirm that angular kinematics derived from MediaPipe landmarks can reliably reflect biomechanically meaningful differences in movement quality when appropriate smoothing is applied, a finding that supports the use of angle-based features in the present system's Wiraga scoring pipeline. The humanoid model optimization employed by Kim et al. (2023) represents a more computationally intensive refinement than is feasible within the resource constraints of a lightweight web application; however, the underlying principle (that angle-based representations are more robust than raw coordinate comparisons) directly informs the feature extraction strategy adopted here.

2.2.3. Application of Improved DTW Algorithm in Sports Posture Analysis

Niu (2024) proposed an enhanced Dynamic Time Warping approach for assessing movement similarity in sports training scenarios, combining DTW with feature selection strategies and temporal segmentation to improve alignment accuracy for complex, multi-phase movement sequences. Their work demonstrated that DTW-based comparison can effectively handle the time-warping distortions introduced by individual differences in movement speed and pacing, producing similarity scores that correlate well with expert assessments of technique quality. Crucially, the study also confirmed that DTW performance scales appropriately when applied to skeletal pose sequences derived from video, rather than from dedicated motion capture hardware, which is a prerequisite for systems intended to operate on ordinary consumer video recordings.

The present system adopts the FastDTW variant of DTW (an $O(N)$)

approximation algorithm) rather than the full quadratic DTW implementation, in order to maintain computational tractability within the constraints of a Python-based web server. Niu (2024) did not evaluate FastDTW specifically; however, their findings regarding the general effectiveness of DTW-family algorithms for video-derived pose sequence comparison provide a relevant empirical basis for the FastDTW adoption decision in the present work. The combination of global audio cross-correlation alignment (to correct for fixed temporal offsets) followed by FastDTW local alignment (to accommodate elastic timing variation) mirrors the two-stage temporal normalization logic advocated by Niu (2024) and directly addresses the starting-time synchronization problem inherent in self-recorded student submissions.

2.2.4. Comparative Study of Human Motion Recognition Based on Pose Estimation

S. Lee et al. (2023) conducted a systematic comparative evaluation of three widely used pose estimation frameworks (MediaPipe Pose, OpenPose, and AlphaPose) across a range of motion analysis tasks, assessing each system on landmark accuracy, processing speed, and robustness under varying conditions of occlusion, lighting, and camera perspective. Their results consistently demonstrated that MediaPipe Pose achieves the most favorable trade-off between accuracy and computational efficiency: while OpenPose and AlphaPose produced marginally higher landmark precision in controlled laboratory settings, MediaPipe Pose maintained competitive accuracy at substantially lower processing costs, making it the only framework among those evaluated that is realistically deployable on standard consumer devices without dedicated acceleration hardware.

Additionally, S. Lee et al. (2023) validated the use of angular kinematics (derived from detected landmark positions using the Law of Cosines) as a view-invariant feature representation for movement assessment. Their findings confirmed that angular features produced more consistent and repeatable pose comparisons than raw coordinate features across recordings made at different camera distances and angles. Mroz et al. (2021) further corroborated this comparative advantage of BlazePose for constrained motion tracking tasks, noting particular strength in scenarios involving a single visible subject performing moderate-complexity movements,

conditions that closely match the use case of the present system. Both studies collectively justify the selection of MediaPipe BlazePose as the foundational pose estimation layer of the proposed architecture, and the adoption of angle-based kinematic features as the primary input to the Wiraga scoring module.

2.2.5. Online Dance Learning with Teacher-AI Cooperation

Kang et al. (2023) conducted a qualitative study examining the experiences of dance learners and instructors engaged in online learning environments augmented by AI-assisted feedback tools. Their findings highlighted a consistent theme: learners perceived AI tools as most beneficial when those tools were explicitly positioned as supplements to instructor-led guidance rather than as autonomous teaching agents. Instructors, for their part, reported that AI feedback was most effective when it addressed technically definable criteria (such as joint alignment and timing accuracy) rather than expressive or aesthetic qualities that resist algorithmic formalization. The study also noted that transparency regarding the source and basis of AI-generated feedback was associated with greater learner trust and willingness to act on the system's recommendations.

These findings carry direct implications for the design philosophy of the present system. The restriction of automated scoring to the Wiraga and Wirama dimensions (while explicitly excluding the expressive Wirasa dimension) is consistent with Kang et al.'d (2023) observation that AI feedback is most credible and pedagogically useful when confined to technically measurable criteria. The decision to generate natural-language corrective feedback rather than presenting only numerical scores reflects the same study's finding that actionable, specific guidance is associated with higher learner engagement. Finally, the framing of the present system as a self-assessment support tool rather than a teaching agent aligns with the teacher–AI cooperation model advocated by Kang et al. (2023), in which AI tools augment self-directed practice without supplanting the instructional relationship.

Table 2-2 Comparison of Reviewed Studies

Study	Pose Model	Temporal Alignment	Feedback Type	Deployment	Dance Domain	Gap Addressed by Present Study
Bazarevsky et al. (2020)	BlazePose (33-KP)	None	None	On-device (mobile)	General / Generic	Provides pose engine; present study extends with visibility gating and kinematic scoring
Kim et al. (2023)	MediaPipe Pose	None	None	Research lab	General	Validates angle-based features; present study applies these within a lightweight web pipeline
Niu et al. (2024)	Skeleton (motion capture / video)	Improved DTW	None	Desktop / Research	Sports (general)	Validates DTW for video pose; present study adopts FastDTW with audio pre-alignment
S. Lee et al. (2023)	MediaPipe, OpenPose, AlphaPose (comparison)	None	None	Research lab	General motion	Justifies MediaPipe + angle features; present study deploys these in an uncontrolled home setting

Kang et al. (2023)	Not specified	Not specif ied	AI- assist ed (quali tative)	Online learnin g	Hip-hop (online)	Informs feedback design philosophy; present study implements and evaluates a rule- based semantic feedback engine
-----------------------	------------------	----------------------	--	------------------------	---------------------	---

2.3. Research Gap

The literature reviewed in the preceding sections reveals a consistent pattern of progress in the individual components relevant to automated movement assessment: pose estimation frameworks have become sufficiently accurate and efficient for consumer-grade hardware deployment; angular kinematic features have been validated as view-invariant and biomechanically meaningful representations of movement quality; DTW-based temporal alignment has demonstrated reliable performance for video-derived skeletal sequence comparison; and semantic feedback design principles have been articulated for motion-based learning contexts. However, the review also reveals that these advances have not yet been integrated in a manner that directly addresses the specific needs of beginner hip-hop learners engaged in self-directed home practice.

Existing systems for automated dance performance evaluation (including the work of Wang et al. (2021)) have predominantly been developed and evaluated in controlled studio environments using professional recording equipment, consistent lighting, and fixed camera configurations. These conditions differ substantially from the uncontrolled home recording contexts in which self-directed learners typically practice. When a student records themselves using a smartphone or laptop webcam in a domestic setting, the resulting video is subject to variable lighting quality, non-standardized camera distance and viewing angle, background clutter, and inconsistent audio quality. No reviewed study has specifically investigated the performance of a pose-based movement assessment pipeline under these ecologically realistic conditions, nor has any study proposed explicit system design adaptations (such as

visibility thresholding and audio-based temporal synchronization) to mitigate the impact of such environmental variability.

A second gap concerns the accessibility of existing systems for self-directed learners. The majority of evaluated motion assessment tools require dedicated software installation, specialized hardware, or institutional computational resources. Wang et al. (2021), for instance, implemented their system as a desktop application; S. Lee et al. (2023) evaluated pose estimation frameworks in research laboratory contexts; and Kang et al. (2023), while addressing online learning environments, did not report on technical implementation or deployment architecture. No reviewed study has described the implementation of a complete movement assessment and feedback pipeline as a lightweight web application deployable on a standard Flask server and accessible through a browser without requiring any client-side software installation. This accessibility gap is particularly significant for educational contexts in developing regions or under-resourced institutions, where dedicated computing infrastructure may not be available.

A third gap concerns the specificity of the movement domain and learner population addressed. Existing automated dance assessment studies have primarily focused on classical or structured dance forms (such as ballet, ballroom, or traditional folk dance), where movement vocabularies are precisely codified and evaluation criteria are well-established in formal pedagogy. Hip-hop movement assessment has received considerably less attention, and foundational hip-hop movement assessment for beginner learners has received virtually none. This is a meaningful omission, because hip-hop's emphasis on rhythmic groove, weight transfer, and spatial flow introduces evaluation challenges (including the importance of auditory beat alignment as a distinct assessment criterion) that are not fully captured by posture-focused systems designed for other genres. The Wirama dimension of the present system's evaluation framework is a direct response to this gap, operationalizing rhythmic accuracy as a first-class scoring dimension alongside spatial form.

A fourth gap relates to the educational design of feedback outputs. While Yu et al. (2024) and Alfredo et al. (2024) articulate design principles for corrective feedback in extended reality and AI-assisted learning contexts respectively, few studies have

reported concrete implementations of rule-based semantic feedback engines in pose-based dance assessment systems, nor have they evaluated the comprehensibility or pedagogical utility of such feedback for beginner learners without prior dance training. The present system addresses this gap by implementing and evaluating a feedback module that maps region-specific kinematic deviations to validate.

