Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.088326

ARTICLE

Neuro-Symbolic Reasoning for 3D Human Pose Analysis

Yucheng Huang[1,*], **Xingyu Gao**[2], **Jianze Wei**[2] **and Hong Yan**[1]

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China
[2]Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China

*Corresponding Author: Yucheng Huang. Email: yuchhuang9-c@my.cityu.edu.hk

Received: 01 July 2026; Accepted: 02 September 2026; Published: 28 September 2026

ABSTRACT: Current 3D Human Pose and Shape estimation methods predominantly focus on numerical coordinate regression, treating pose analysis as a geometric mapping task rather than a structured reasoning problem. This reliance on quantitative output creates an interpretability gap, as these models fail to provide high-level, qualitative justifications for their predictions, such as why a specific pose is physically unstable or biomechanically incorrect. To bridge this gap, we propose a novel framework that leverages neuro-symbolic reasoning for the analysis of human poses from multimodal inputs. Our approach integrates symbolic expressions and logical deduction rules into a large language model-based reasoning pipeline. Specifically, the system first translates natural language descriptions of human interactions into structured symbolic representations, then derives a step-by-step plan to reason about 3D poses and verify natural language statements using symbolic logical rules. A verification mechanism ensures the reliability of both the symbolic translation and the reasoning process. Despite its simplicity, our approach captures the physical characteristics of pose and understands the pose structure via neuro-symbolic reasoning. Evaluations demonstrate that the integration of Large Language Models (LLMs) with neuro-symbolic logic provides a reliable and interpretable approach for automated posture balance assessment and anomaly identification.

KEYWORDS: Large Language Model (LLM); human pose reasoning; Chain-of-Thought (CoT); neuro-symbolic AI; 3D human pose estimation; spatial reasoning

1 Introduction

The emergence of LLMs has fundamentally redefined the landscape of artificial intelligence, transitioning the field from simple pattern matching to complex, multi-step reasoning [1,2]. Through massive-scale pre-training, these models have demonstrated an unprecedented capacity for abstraction, commonsense deduction, and the ability to follow intricate instructions. This "reasoning-first" paradigm has naturally paved the way for Vision-Language Models (VLMs), which bridge the gap between visual perception and linguistic understanding [3–5]. In the context of 3D human pose analysis, VLMs offer a transformative opportunity: they allow machines to interpret human movement not just as a collection of pixel coordinates, but as a semantically meaningful sequence described through natural language.

Several recent works have attempted to connect 3D Human Pose and Shape (HPS) estimation with VLMs, seeking to move beyond purely visual processing toward models capable of explicit multimodal understanding [6–8]. These methods typically introduce textual descriptions as auxiliary inputs to refine mesh parameters or provide global contextual cues. While effective in improving estimation accuracy, often measured by metrics such as Mean Per Joint Position Error (MPJPE), most existing approaches treat language

as an additional feature modality rather than a reasoning medium. As a result, they largely overlook the principal advantage of VLMs: the ability to perform structured logical inference. Consequently, complex biomechanical relationships and physical constraints are only implicitly learned, limiting interpretability and reliability.

A key limitation of current VLM-based approaches lies in their reliance on Chain-of-Thought (CoT) reasoning [9–11]. Standard CoT generates reasoning steps through sequential token prediction, lacking explicit external mechanisms to enforce geometric boundaries or physical laws. Although deterministic generation can be enforced using greedy decoding, the underlying reasoning process remains grounded in statistical co-occurrence rather than verifiable deduction. This design choice is misaligned with the requirements of HPS, where correctness depends on precise spatial relationships, kinematic feasibility, and physical consistency. Consequently, the design choice often leads to logical hallucinations, where a model generates a linguistically plausible narrative that fails to reflect physical reality, resulting in 3D reconstructions that are visually convincing but physically impossible or inherently unstable.

This limitation reveals that leveraging language in HPS is not merely a matter of appending textual features to a visual backbone. Human pose understanding requires structured reasoning over anatomy, joint connectivity, and spatial constraints. Visual cues alone are often ambiguous due to depth uncertainty, self-occlusion, or viewpoint variation-challenges that are well known in monocular 3D pose estimation. Language, however, can provide high-level semantic constraints that reduce this ambiguity. For instance, an instruction such as "the left hand is raised above the shoulder and touches the right ear" specifies explicit entities, relative spatial relations, and contact constraints. Correctly using such information requires decomposing the sentence into symbolic components and reasoning over their geometric implications. This form of compositional and constraint-aware reasoning is essential for pruning the large space of feasible 3D poses while maintaining physical and kinematic validity.

To achieve the rigor required for reliable pose reasoning, we must move beyond the informal deductions of standard Chain-of-Thought (CoT) prompting and toward formal symbolic representations. While CoT provides step-by-step natural language explanations, it lacks mathematical grounding and deterministic verification. Symbolic logic addresses this limitation by providing a deterministic scaffold that maps subjective linguistic descriptions to precise mathematical predicates. Although discretizing continuous human movement into logical units presents inherent challenges, forcing the deduction through a symbolic framework provides a structured mechanism to better align each inference step with physical and kinematic constraints.

Inspired by this observation, we propose a novel framework that integrates neuro-symbolic reasoning into 3D human pose analysis. As illustrated in Fig. 1, while conventional methods rely purely on quantitative coordinate regression without explicit explanations, our approach decomposes complex postures into verifiable logical predicates. The system operates on the premise that symbolic expressions and logical deduction rules provide an essential structural scaffold for LLMs, enabling them to accurately interpret language and translate visual observations into actionable pose constraints. Specifically, the system first parses natural language descriptions into structured symbolic representations, then derives a coherent, step-by-step reasoning plan using symbolic logic rules to infer pose validity or reconstruction constraints. A dedicated verification mechanism is incorporated to ensure the faithfulness and reliability of both the symbolic translation and the subsequent reasoning process.

Despite its conceptual simplicity, the proposed framework demonstrates a strong capacity to capture the physical characteristics and biomechanical structure of human poses. By harnessing the reasoning capabilities of LLMs within a symbolic logic framework, our method effectively converts subjective linguistic descriptions into objective, tractable constraints for 3D pose analysis. Extensive experimental evaluations show that the proposed approach provides a reliable and interpretable alternative for applications

such as posture correction, while reducing dependence on costly motion capture systems and extensive manual annotations.

To summarise, the main contributions of the paper are:

1. We introduce a neuro-symbolic reasoning framework for human pose analysis that integrates visual perception with formal logic. Specifically, our pipeline leverages a Visual Parser to extract structured symbolic predicates from visual inputs and instructions, and a Symbolic Reasoner to formulate First-Order Logic rules and multi-step deduction plans for transparent pose evaluation.
2. We incorporate a dedicated Integrity Verifier mechanism that strictly enforces logical consistency and ensures the faithfulness and reliability of the symbolic reasoning process, thereby substantially improving the trustworthiness and interpretability of the final pose reasoning outcomes.
3. We propose a protocol to generate stability-annotated benchmarks via directional perturbations of existing datasets. By constraining rotation magnitudes, we systematically synthesize physically plausible, unbalanced samples.

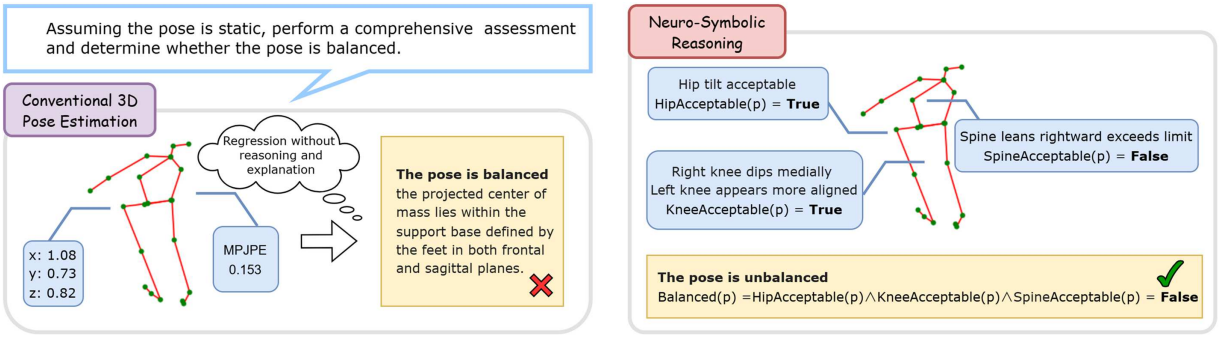

Figure 1: Comparison between conventional 3D pose estimation and Neuro-Symbolic Reasoning. (**Left**) Standard regression models often output numeric results without logical explanation, potentially leading to incorrect stability assessments. (**Right**) Our neuro-symbolic approach decomposes the pose into verifiable predicates.

2 Related Work

2.1 Physics-Aware Human Pose Estimation

Estimating human poses that are not only visually congruent but also physically plausible remains a challenge in computer vision. Early methods often focus on kinematic-based pose models [12–16], which, despite their precision in static pose regression, frequently suffer from "non-physical" artifacts such as jitter, foot-sliding, or the jarring phenomenon of ground penetration [17]. Recent graph-based approaches further improve pose estimation by modeling complex dependencies among skeletal joints [18,19]. Nevertheless, these methods primarily focus on improving coordinate estimation accuracy, while biomechanical validity and interpretable reasoning remain implicit.

To address these issues, subsequent research has pivoted toward the explicit integration of physical constraints. One prominent branch involves the use of foot-ground contact priors to anchor reconstructed motions to a global coordinate system [20–22]. Others have turned to Proportional-Derivative Control, simulating the underlying forces required to actuate a digital character toward target joint positions [23–27]. However, these simulators often present a trade-off: they are either non-differentiable, which can be hard to optimize, or overly simplified, which can only be applied for specific tasks. Although reinforcement learning offers a robust alternative by teaching agents to apply torques to a humanoid body [28,29], the inherent computational overhead and the need for retraining in new scenarios limit their generalization ability. In

a departure from these heavy-compute approaches, we introduce a training-free framework powered by neuro-symbolic reasoning.

2.2 Neuro-Symbolic Reasoning in Large Language Model

Chain-of-Thought prompting has bolstered the reasoning performance of Large Language Models by decomposing intricate queries into logical sub-steps [1,2,30,31]. CoT often struggles with the precision required for formal logical and spatial tasks, frequently succumbing to hallucinations or "shortcut learning", where models mirror statistical correlations rather than adhering to deductive truths [32].

To address these systemic vulnerabilities, Neuro-Symbolic Reasoning has emerged as a robust paradigm that integrates the flexible linguistic processing of deep learning with the rigorous, verifiable logic of symbolic representations [33–35]. Several frameworks have been developed to ground LLM reasoning in formal structures. For instance, Logic-LM and LINC utilize LLMs to translate natural language into symbolic formalisms like First-Order Logic (FOL), which are subsequently processed by external reasoning engines to ensure faithfulness [36,37]. Similarly, QuaSAR disentangles world knowledge from logical manipulation to mitigate content bias [38]. More sophisticated paradigms, such as the Safe framework or Graph-constrained Reasoning, introduce modular verifiers and machine-checkable proofs to audit the reasoning path [39,40]. Importantly, recent work has extended neuro-symbolic AI to context-aware decision making in real-time systems. A new Neuro-Symbolic framework [41] integrates deep neural representations of sensory perception with symbolic knowledge graphs to facilitate semantically based, explainable, and adaptable decision-making processes under uncertainty.

However, these methods are predominantly designed for general commonsense reasoning, mathematical problem-solving, or broad-scope real-time control. They have not been systematically adapted for the fine-grained spatial and kinematic constraints inherent in 3D human pose analysis, where biomechanical validity, joint connectivity, and physical plausibility must be rigorously enforced.

2.3 Language Priors on Human Pose

The bridge between 3D coordinates and natural language has been significantly strengthened by the arrival of specialized datasets like PoseScript and PoseFix, which pair 3D poses with fine-grained textual modifiers [42,43]. Building on this, recent efforts like HUMANML3D++ incorporate scene-based texts and large-scale web video data to support temporal and 3D-free motion generation [44].

These resources have catalyzed a shift from simple embedding-based retrieval to multimodal reasoning. ChatPose embeds 3D human poses into a multimodal Large Language Model to enable the semantic understanding and reasoning-based generation of human postures from both text and visual inputs [45]. PoseLLaVA and UniPose refine this by integrating pose encoder-decoders and tokenizers that quantize 3D data into discrete tokens, utilizing mixed-attention mechanisms to handle the spatial nature of joint data [46,47]. Although direct 3D tokenization may preserve depth and out-of-plane relationships more completely, it requires additional tokenizer training and is tied to a specific pose representation. In contrast, our framework is training-free and accepts either skeleton or mesh renderings. Notably, an embodied AI-based athlete posture estimation approach [48] leverages physical interaction priors to refine pose analysis in dynamic sports scenarios. Their work demonstrates that embedding physical context—such as athlete-environment interaction—can significantly improve estimation robustness under motion ambiguities. However, this approach remains grounded primarily in implicit neural feature learning rather than explicit symbolic reasoning and does not provide verifiable justifications for its biomechanical assessments.

The integration of language priors has also redefined generative trajectories through diffusion-based frameworks. FinePOSE utilizes part-aware prompts to "imagine" missing joints with high biological plausibility [49], while GENMO treats pose estimation itself as a form of "constrained motion generation" [50]. To overcome data scarcity, OOHMG investigates "wordless training" to allow text-to-pose generators to generalize to real-world descriptions without explicit paired training texts [51]. Yet, a subtle gap remains: most frameworks rely on extensive pre-existing text annotations and struggle to maintain precision in complex, multi-person scenes where visual cues are frustratingly sparse.

3 Problem Formulation

The task is defined as a visual reasoning and verification problem. Given an input image I depicting a human pose and a natural language statement S describing that pose, the objective is to determine the truth value of S based on the visual evidence provided by I.

To enable structured reasoning, we explicitly define a symbolic pose state space. Let

$$\mathcal{J} = \{j_1, j_2, \ldots, j_N\} \tag{1}$$

denote the set of N anatomical joints of the human body, where each joint $j_i \in \mathcal{J}$ is associated with a 3D position $\mathbf{p}_i \in \mathbb{R}^3$. The joint set includes major body landmarks such as the head, shoulders, elbows, wrists, hips, knees, and ankles. Based on these joints, we define a finite set of symbolic spatial and biomechanical relations

$$\mathcal{R} = \{r_1, r_2, \ldots, r_M\}, \tag{2}$$

where each relation $r_k \in \mathcal{R}$ represents a semantically meaningful geometric or biomechanical relationship between joints or body parts. Each relation is instantiated as a predicate

$$r_k(j_a, j_b, \theta), \tag{3}$$

where $j_a, j_b \in \mathcal{J}$ and θ denotes an optional numerical parameter such as distance, angle, or tolerance.

The symbolic pose state is represented as a set of grounded predicates

$$\mathcal{S} = \{r_k(j_a, j_b, \theta) \mid j_a, j_b \in \mathcal{J},\ r_k \in \mathcal{R}\}, \tag{4}$$

which collectively encode the observable spatial and biomechanical properties of the pose. These predicates form the foundational premises for symbolic reasoning.

The system performs reasoning through the following structured process:

1. **Predicate Formation:** A Large Language Model, guided by a structured prompt, analyzes the image I to extract a set of foundational predicates $\mathcal{P} \subseteq \mathcal{S}$. These predicates describe observable joint configurations and spatial relationships in symbolic form.
2. **Logical Inference:** Using Neuro-Symbolic Reasoning, the system performs step-by-step logical deductions based on the premises $\mathcal{P}$ and a predefined set of logical and biomechanical rules.
3. **Statement Verification:** The goal is to derive a conclusion regarding the statement S. The final output is a ternary verdict:
   - **T (True)** if S can be logically entailed from $\mathcal{P}$;
   - **F (False)** if the negation of S ($\neg S$) can be logically entailed from $\mathcal{P}$;
   - **U (Unknown)** if the available premises are insufficient to prove either S or $\neg S$.

Formally, the task is to learn a function

$$f(I, S) \rightarrow \{\mathrm{T}, \mathrm{F}, \mathrm{U}\}, \tag{5}$$

where the decision is derived through an explicit symbolic reasoning chain rather than an opaque, end-to-end classification. This formulation emphasizes verifiability, faithfulness, and interpretability in visual reasoning for human poses.

4 Method

4.1 Framework Overview

Our proposed framework is supported by a Large Language Model and is structured around four core modules: the Visual Parser, the Symbolic Reasoner, the Logic Solver, and the Integrity Verifier, as illustrated in Fig. 2. This modular neuro-symbolic architecture ensures a transparent, step-by-step reasoning process from visual input to final verification. The roles of these modules are elaborated as follows.

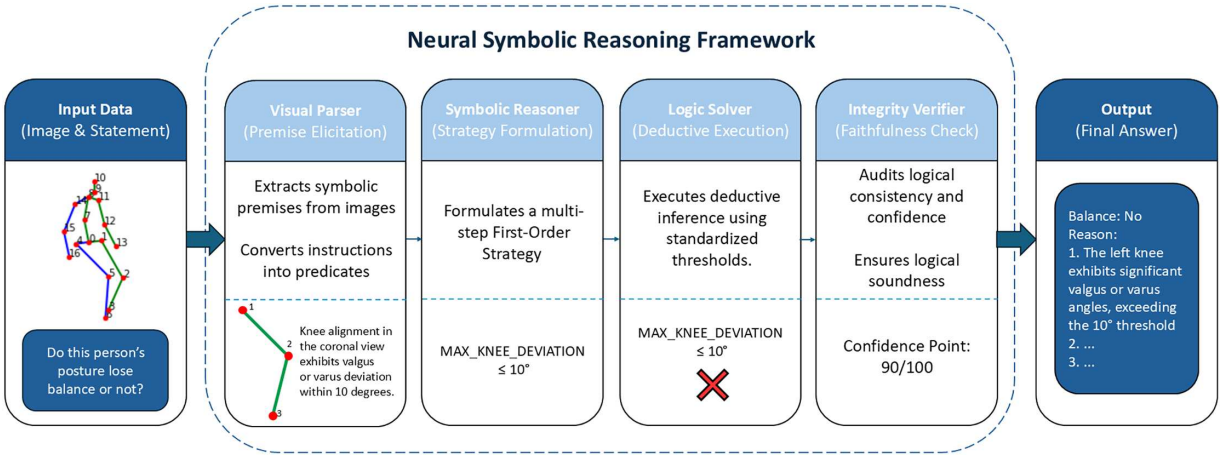

Figure 2: Proposed Symbolic Neuro-Symbolic Reasoning framework for human pose analysis. Given 3D mesh or skeleton input, the system translates biomechanical criteria into symbolic predicates, formulates a first-order logic reasoning plan, performs deductive inference, and verifies logical consistency before producing the final decision.

4.1.1 Visual Parser

The Visual Parser module is responsible for converting the raw visual information from the input image I into a structured set of premises $\mathcal{P}$. Guided by a carefully designed prompt, the LLM acts as a visual interpreter, identifying key body joints, their configurations, and spatial relationships (e.g., 'left hand is above shoulder', 'right foot is touching the ground'). The output of this module is a formal, language-based representation that serves as the foundational facts for all subsequent logical reasoning, bridging the gap between pixel-level data and symbolic processing.

4.1.2 Symbolic Reasoner

Given the premises $\mathcal{P}$ and the statement S under verification, the Symbolic Reasoner module devises a step-by-step reasoning strategy. It decomposes the complex task of verifying S into a sequence of simpler, logically sound sub-tasks or deduction steps. This involves determining the order in which rules (e.g., spatial transitive relations, biomechanical constraints) should be applied to infer new facts from $\mathcal{P}$ that lead to

a conclusion about S. The output is a reasoning plan, which outlines the logical pathway from premises to conclusion.

4.1.3 Logic Solver

The Logic Solver module executes the reasoning plan generated by the Symbolic Reasoner. It follows the plan step-by-step, applying the specified logical and kinematic rules to the premises $\mathcal{P}$ to derive new intermediate conclusions. This process continues iteratively until no new facts can be deduced or until a definitive conclusion about the statement S (or its negation $\neg S$) is reached. The Logic Solver is the core deduction engine that performs the actual symbolic inference.

4.1.4 Integrity Verifier

The Integrity Verifier module is a critical component for ensuring the reliability and robustness of the entire reasoning pipeline. It operates at two levels:

1. **Internal Consistency Check:** It verifies the logical soundness of each deduction step in the Logic Solver's output, ensuring that no rules are misapplied and that the conclusions validly follow from the premises.
2. **Final Output Validation:** It cross-checks the final verdict (T, F, U) against the original premises $\mathcal{P}$ to guard against logical contradictions and to confirm that the "Unknown" (U) status is justified by a genuine lack of evidence rather than a reasoning failure.

This module adds a crucial layer of faithfulness, making the system's conclusions more trustworthy.

4.2 Visual Parser (Premise Elicitation)

The input to the Visual Parser consists of standardized 2D multi-view image projections—specifically, the orthogonal Front (Coronal) View and Side (Sagittal) View—rendered directly from either a 3D skeleton coordinate configuration or a 3D Skinned Multi-Person Linear (SMPL) mesh reconstruction. By rendering these complex, high-dimensional 3D representations into dual-perspective 2D images, the framework bypasses the challenge of having a text-based language model interpret raw spatial coordinate matrices or dense mesh vertex tensors. To extract these specific findings, a detailed, expert-guided prompt is fed into the Large Vision-Language Model alongside the dual Front and Side view images. The prompt instructs the model to act as a clinical biomechanics analyst, evaluating the rendered skeletal or mesh profiles against explicit structural guidelines. Rather than predicting raw 3D coordinates, the model performs localized spatial reasoning over the rendered geometries to check for threshold compliance, outputting a structured set of premises $\mathcal{P} = \{p_1, p_2, ..., p_n\}$, where each premise p_i is a direct finding from the image against a specific criterion. The partial set of premises used for this demonstration is defined as follows:

- p_1(CoG Sagittal): Center of gravity (CoG) in the sagittal view lies within the base of support.
- p_2(CoG Coronal): CoG in the coronal view lies within the base of support.
- p_3(Spinal Inclination): Forward spinal inclination in the sagittal view exceeds the threshold (Measured: $\approx 22°$; Limit: 15°).
- p_4(Knee Alignment): Knee alignment in the coronal view exhibits excessive valgus deviation in the left knee (Measured: ≈12–15°; Limit: 10°).
- p_5(Hip Tilt): Hip tilt remains within the acceptable range (Measured: $\approx 7°$; Limit: ±10°).
- p_6(Ankle Dorsiflexion): Ankle dorsiflexion in the sagittal view does not exceed the threshold (Measured: ≈18.4°; Limit: 25°).
- p_7(Arm Asymmetry): Arm asymmetry is present but does not indicate significant imbalance as the CoG remains centered and no external load is detected.

Rather than computing precise joint coordinate angles or center-of-gravity trajectories through an explicit physics simulation engine, the framework relies on the advanced spatial-perceptual capabilities of the Large Vision-Language Model (LVLM). The LVLM performs qualitative perceptual estimation via visual inspection of the multi-view rendered images. To ground this estimation, the model is provided with explicit, text-based structural guidelines. This transforms a heavy numerical computation task into a structured visual-verification problem, mimicking how a human clinical analyst visually screens a pose for overt biomechanical deviations. Thus, we form this structured set $\mathcal{P}$, which serves as the foundational, objective evidence for all subsequent logical reasoning and verification.

4.3 Symbolic Reasoner (Strategy Formulation)

The Symbolic Reasoner module takes the structured premises $\mathcal{P}$ from the Visual Parser and the target statement S (e.g., "The pose is functionally stable") as input. Its role is to generate a reasoning plan $R = [r_1, r_2, ..., r_k]$ that outlines the logical path to verify S. This involves defining the formal logic framework and the sequence of deductions required.

For instance, given the biomechanical stability task, the Visual Parser would generate a multi-step strategy analogous to the example output:

4.3.1 Formal Logic Framework Definition

The Symbolic Reasoner first establishes the logical vocabulary required for reasoning. To support symbolic reasoning, the Symbolic Reasoner defines a set of pose-related predicates that encode spatial and biomechanical properties of a human pose. Let P denote a pose instance corresponding to a specific subject.

- CoGWithinBaseOfSupport(P): A boolean predicate indicating that the CoG of pose P lies within the subject's base of support, as determined from multi-view observations.
- VisibleCoGShift(P): A boolean predicate indicating that pose P exhibits a visually observable shift in the center of gravity, typically induced by body or load asymmetry and detectable from multi-view observations.
- DynamicPose(P): A predicate indicating that pose P corresponds to a dynamic configuration, characterized by non-static body motion or transitional joint configurations rather than a steady-state posture. In the current single-frame implementation, this predicate is not derived from temporal motion trajectories. Instead, the prompt explicitly instructs the model to consider that dynamic poses may exhibit temporary center-of-gravity shifts that remain functionally acceptable.
- SpinalAlignment(P, A): A predicate specifying that the spinal alignment of pose P deviates by an angle A from the neutral upright posture, where A is measured in degrees and constrained by an application-dependent threshold.
- ArmAsymmetry(P, W): A predicate denoting that the arm configuration of pose P results in an asymmetric load distribution corresponding to a weight ratio W, expressed as a percentage of the subject's body weight.
- FunctionalStability(P): A high-level predicate denoting that pose P satisfies the biomechanical conditions required for functional stability, as determined through symbolic inference over multiple lower-level spatial and alignment predicates.

The Symbolic Reasoner also includes domain-specific constants which define biomechanical thresholds

- MAX_SPINAL_ANGLE = 15°,
- MAX_KNEE_DEVIATION = 10°,
- MAX_ANKLE_DORSIFLEXION = 25°,
- MAX_ARM_ASYMMETRY_WEIGHT = 0.05 × BodyWeight, etc.

4.3.2 Domain Knowledge Encoding

Biomechanical knowledge is encoded as First-Order Logic (FOL) rules. For example,

- $\forall P \forall W\,(\text{ArmAsymmetry}(P, W) \land W \leq \text{MAX\_ARM\_ASYMMETRY\_WEIGHT} \implies \text{FunctionalStability}(P))$
- $\forall P\,(\text{DynamicPose}(P) \land \text{VisibleCoGShift}(P) \implies \text{FunctionalStability}(P))$

4.3.3 Goal Formulation

The target statement S is formalized as a composite logical condition defined by the conjunction of required criteria:

- $\forall P\,(\text{CoGWithinBaseOfSupport}(P) \land \text{SpinalAlignment}(P, A) \land \quad \ldots \implies \text{FunctionalStability}(P))$

4.3.4 Deduction Plan Generation

Based on the defined predicates and rules, the Symbolic Reasoner produces an ordered deduction plan that specifies how the Logic Solver should instantiate the general FOL rules with the concrete premises $\mathcal{P}$, evaluate each logical condition against the defined constants, and apply valid inference rules (e.g., modus ponens) to derive a conclusion regarding S.

This structured plan R transforms the abstract verification goal into a concrete, executable sequence of symbolic operations for the Logic Solver.

4.4 Logic Solver (Deductive Execution)

The Logic Solver module executes the formal reasoning plan R generated by the Symbolic Reasoner. It instantiates the general First-Order Logic rules with the specific symbolic premises $\mathcal{P}$ provided by the Visual Parser. The execution is a process of logical inference and constraint checking.

For the biomechanical stability example, the Logic Solver performs the following sequence of operations:

4.4.1 Check Criteria Against Premises

It systematically evaluates each condition of the comprehensive FOL formula against the fact set $\mathcal{P}$:

- Checks if CoGWithinBaseOfSupport(P) is TRUE, given p_1 and p_2.
- Checks if SpinalAlignment(P, A) is TRUE with $A \leq 15$, given p_3.
- Similarly checks knee, hip, and ankle alignment predicates against premises p_4, p_5, p_6.
- Evaluates the arm asymmetry condition against premises p_7.

4.4.2 Derive Intermediate Conclusions

For each criterion, the Logic Solver produces a boolean value (TRUE/FALSE) based on the evaluation.

4.4.3 Aggregate Results for Final Verdict

The Logic Solver applies the logical conjunction from the comprehensive formula. If all individual criteria are satisfied (evaluate to TRUE), it derives the final conclusion C = "yes" (True) for the balance statement S. In cases where a violation is detected—such as a threshold breach in spinal or knee alignment—the solver short-circuits the positive verdict and generates a structured report. As shown in Fig. 3, this output

explicitly maps failed predicates to their physical measurements, providing a transparent, interpretable trace of the reasoning process:


```
{
  "balance":  "no",
  "reasons":  [
  "Center of Gravity (CoG) is within the base of support in both front and
side views.",
  "Spinal alignment in the sagittal view shows a forward lean of
approximately 22 deg, which exceeds the acceptable limit of 15 deg.",
  "Knee alignment in the coronal view exhibits a valgus deviation (approx.
12–15 deg) in the left knee, exceeding the 10 deg threshold.",
  "Hip alignment is neutral with a tilt of approximately 7 deg, remaining
within the +/-10 deg range.",
  "Ankle dorsiflexion is within the acceptable range at approximately 18.4
deg, staying below the 25 deg limit.",
  "Arm asymmetry is present but does not indicate significant imbalance as
the CoG remains centrally aligned.",
  "There is no evidence of external load or a shift in CoG resulting from
limb asymmetry."
  ]
}
```


Figure 3: The raw JSON output of the balance assessment system.

This output represents the final fact set F_k, which is passed to the Integrity Verifier.

4.5 Integrity Verifier (Faithfulness Check)

The Integrity Verifier module performs a critical audit of the entire reasoning pipeline to ensure its reliability. It analyzes the outputs from the previous modules against the original input and domain knowledge, focusing on three key aspects.

4.5.1 Symbolic Context Consistency

The Integrity Verifier checks that the symbolic representations (predicates, constants) and logical structure generated by the Symbolic Reasoner faithfully capture the semantics and intent of the original natural language instructions and domain criteria. For instance, it validates that MAX_SPINAL_ANGLE = 15 correctly represents "≤15 degree forward/backward lean is acceptable" and that the FOL formula for FunctionalStability(P) comprehensively incorporates all specified biomechanical criteria.

4.5.2 Logical Validity of the Solving Step

The Integrity Verifier examines the Logic Solver's execution trace to ensure that the deduction process is sound. It confirms that the general FOL rules were correctly instantiated with the specific premises $\mathcal{P}$, that all logical operations (e.g., conjunctions, implications) were applied validly, and that the final JSON verdict (e.g., "balance: yes") is a necessary logical consequence of the satisfied criteria. It specifically checks for errors in handling complex clauses, such as those accounting for dynamic poses or acceptable asymmetries.

4.5.3 Confidence Assessment

To compute the quantitative confidence score ($C \in [0, 100]$) and determine final verdict overrides, the Integrity Verifier applies a deterministic deductive rubric based on a strict confidence threshold:

- **Standard Verification ($C \geq 70$):** The reasoning chain is deemed logically sound and consistent. The final ternary verdict (T or F) derived by the Logic Solver is maintained and exported.
- **Fallback and Rejection ($C < 70 \implies$ Verdict = U):** If the verifier flags an internal logical contradiction—such as the Logic Solver declaring "balance: no" while all underlying biomechanical predicates are logged within safe thresholds—the confidence score falls below 70. Under this condition, the verifier automatically overrides the system conclusion and strictly returns Unknown (U) to prevent hallucinated anomalies from leaking into the final report.

5 Experiments

To rigorously evaluate the effectiveness of our proposed Neuro-Symbolic Reasoning framework for human pose reasoning, we conducted a series of experiments designed to answer the following key research questions:

- **RQ1:** How accurately does the Neuro-Symbolic Reasoning framework verify biomechanical statements?
- **RQ2:** How does each module contribute to the overall performance and reliability?
- **RQ3:** Does the generated reasoning align with the underlying biomechanical perturbations?
- **RQ4:** Can Neuro-Symbolic Reasoning generalize to 3D mesh-based representations and reason about complex biomechanical properties?

5.1 Model

To execute the Neuro-Symbolic Reasoning, we utilize Qwen2.5-VL-72B [52] as the core LVLM. The model is deployed to perform multi-modal logical deduction by integrating multi-view projections with the extracted biomechanical predicates. To ensure the reproducibility of our experimental results and facilitate a fair comparative analysis across different models, we adopted a standardized inference protocol:

- **Deterministic Decoding:** We primarily employed greedy decoding by setting the sampling temperature to $T = 0$. This configuration ensures that the model consistently selects the token with the highest logit probability, thereby minimizing stochastic variance in the logical reasoning outputs.
- **In-Context Learning:** To maintain parity in prompting conditions, a uniform set of in-context examples was utilized for all tested models. These few-shot demonstrations provide a consistent semantic baseline for the premise elicitation and reasoning tasks.
- **Fail-Safe Mechanism:** In instances where the output was filtered due to sensitive content or where the model returned an "Unknown" (U) classification, a fallback protocol was triggered. Under these conditions, the temperature was reset to the model's default value, and the inference was re-executed to obtain a valid logical deduction.

5.2 Dataset Construction and Evaluation

To address the scarcity of benchmarks for natural language statement verification on 3D human kinematics, we curated a specialized evaluation dataset comprising 506 unique 3D human pose samples. This dataset is architected to evaluate the alignment between visual skeletal data and biomechanical logic through two distinct classes of samples.

5.2.1 Positive Sample Acquisition

The positive (balanced) samples are sourced from the Human3.6M dataset [53], specifically filtered for the Walking category. These samples represent naturally stable gait cycles with kinetically valid postures, providing a baseline for normal human locomotion.

5.2.2 Biomechanically Constrained Negative Synthesis

To generate high-quality negative samples that are "unbalanced" yet physically plausible, we developed a perturbation protocol grounded in authoritative biomechanical literature. The synthesis process adheres to the following constraints:

- **Feasibility Constraints:** The maximum range of motion for every joint is bounded by the NASA Handbook [54], ensuring that generated poses do not violate human anatomical limits (θ_{max}).
- **Stability Thresholds:** The definition of imbalance is derived from the Limits of Stability (LoS), the maximum angle from vertical that can be tolerated without a compensatory step strategy [55].

5.2.3 Perturbation Protocol

For each positive sample, we synthesize a corresponding unbalanced pose by applying a directional rotation vector. The procedure is defined as follows:

- **Direction Selection:** A perturbation direction d is randomly selected from the primary anatomical directions:
  $\{Anterior, Posterior, Left, Right\}$.
- **Rotation Magnitude:** A global rotation angle x is determined to force the center of gravity beyond the base of support while remaining anatomically intact. This is formalized by the inequality:

$$\text{LoS} < x < \theta_{max} \tag{6}$$

  where LoS represents the limit of stability threshold and θ_{max} represents the anatomical joint constraint.
- **Kinematic Chain Weighting:** To simulate realistic falling mechanics, the perturbation is applied to specific nodes in the kinematic chain with varying intensity coefficients (w) as illustrated in Table 1. The final pose P' is generated by rotating the subset of joints J by $w_j \cdot x$ in direction d.

Table 1: Joint perturbation intensity coefficients.

Joint Node (j)	Coeff. (w)	Role in Kinetic Chain
Spine 1	1.0	Primary Trunk Lean
L_Hip/R_Hip	1.0	Pelvic Girdle Rotation
L_Ankle/R_Ankle	–1.0	Counter-Rotation/Pivot
Spine 2	0.5	Thoracic Follow-through

The negative ankle coefficient does not represent a negative perturbation magnitude; instead, it indicates that the ankles rotate in the opposite direction to the imposed spinal and pelvic rotation. This counter-rotation approximates a distal pivot about the support point, producing a coordinated whole-body lean rather than a rigid translation of the entire skeleton. For example, when the trunk and hips are perturbed anteriorly, the ankles rotate posteriorly relative to the kinematic chain, representing an ankle-based compensatory response to the induced imbalance.

Because the proposed neuro-symbolic framework operates in a completely training-free, zero-shot inference paradigm, the 506 curated samples are treated strictly as an evaluation benchmark, omitting the need for traditional training or validation data splits. To ensure rigorous validation of reasoning faithfulness, the synthetic negative perturbations are engineered using a deterministic joint-rotation matrix. These modifications are strictly bounded by Feasibility Constraints and Stability Thresholds introduced in the previous section.

The purpose of this perturbation protocol is not to simulate the distribution of real-world falls or support clinical fall detection. Instead, it provides a controlled benchmark with known perturbation directions and biomechanical constraint violations, enabling evaluation of whether the generated reasoning is consistent with the underlying pose modification. Accordingly, the kinematic-chain weights are selected to produce anatomically feasible and directionally distinguishable samples rather than to reproduce clinical fall dynamics.

5.2.4 Evaluation Metrics

We employ the following metrics for a comprehensive comparison:

- **Accuracy:** The overall proportion of correctly classified statements (T, F).
- **Precision, Recall, F1-Score:** Calculated for each class (T, F) to assess per-class performance.
- **Expected Calibration Error (ECE):** To evaluate the reliability of the model's confidence estimates, we measure calibration using Expected Calibration Error. ECE quantifies the discrepancy between predicted confidence and empirical accuracy by partitioning predictions into confidence bins and computing the weighted average of the absolute difference between confidence and accuracy within each bin [56]. Lower ECE values indicate better-calibrated and more trustworthy decision-making, which is critical for interpretable reasoning systems.

5.3 Results and Analysis

5.3.1 Overall Performance (RQ1)

The classification performance of the proposed neuro-symbolic reasoning framework was quantitatively assessed using a confusion matrix across 506 test samples (Table 2). Based on these results, the framework achieved an overall accuracy of 83.4%.

Table 2: Pose balance confusion matrix.

Predicted	Expected	
	Balanced	**Unbalanced**
Balanced	187 Correct detection	18 Type I error
Unbalanced	66 Type II error	235 Correct rejection

Detailed performance metrics reveal the model's Precision of 91.2%, Recall of 73.9%, and an F1-score of 81.6%, further demonstrate that the framework maintains a strong balance between detection coverage and classification reliability.

The error distribution comprises 18 cases of Type I error (false positives) and 66 cases of Type II error (false negatives). Further analysis suggests that Type II errors typically occur in "borderline" samples where joint rotations are near the limits of stability but have not yet violated the equilibrium constraints defined in Section 2. Conversely, the low count of Type I errors suggests that the symbolic reasoning process effectively captures balance failure.

5.3.2 Ablation Study (RQ2)

To quantify the individual contributions of the core modules within our Neuro-Symbolic Reasoning framework, we conducted a systematic ablation study. The results, detailed in Table 3, underscore how the integration of visual-to-symbolic translation and logical verification is essential for both predictive accuracy and statistical calibration.

Table 3: Ablation study on the effect of different modules. Bold text indicates the best performance for each metric. The downward arrow (↓) indicates that lower values correspond to better calibration performance.

Variant	Accuracy	Recall	Precision	F1 Score	ECE ↓
Neuro-Symbolic Reasoning	**0.834**	**0.739**	**0.912**	**0.816**	**0.118**
- w/o Symbolic Reasoner	0.783	0.672	0.864	0.756	0.152
- w/o Integrity Verifier	0.791	0.696	0.859	0.769	0.166
- w/o Symbolic Reasoner & Integrity Verifier	0.757	0.628	0.804	0.705	0.205

Eliminating the Symbolic Reasoner degrades both the Precision and the F1-scores. This decline highlights that direct inference lacks the deductive rigor required to navigate complex 3D human kinematics. The structured reasoner is essential for maintaining a logical chain of thought that links extracted predicates to a final stability classification.

The removal of the Integrity Verifier results in the most significant increase in Expected Calibration Error (ECE), rising from 0.118 to 0.166. This confirms that the Integrity Verifier acts as a critical regulatory mechanism, preventing the model from generating overconfident but logically inconsistent assertions and ensuring that predicted probabilities align more closely with actual accuracy.

The most severe performance degradation occurs when both the Symbolic Reasoner and the Integrity Verifier are bypassed, forcing the model to rely only on visual inputs. In this configuration, Accuracy drops to 0.757 and ECE surges to a maximum of 0.205. These results demonstrate that visual features fail to provide the stable, low-dimensional symbolic grounding necessary for reliable geometric reasoning, particularly in borderline balance cases.

Overall, the data suggest that while the Visual Parser and Logic Solver establish the foundational evidence, the Symbolic Reasoner and Integrity Verifier are what elevate the system from simple pattern recognition to robust, well-calibrated biomechanical analysis.

5.3.3 Evaluation of Reasoning Faithfulness (RQ3)

While overall accuracy and module utility are essential, they do not guarantee that the model is "right for the right reasons." To address **RQ3**, we evaluate the faithfulness of the generated reasoning by comparing the parsed justifications against the ground-truth perturbations applied during dataset synthesis.

For this analysis, we utilized 253 samples specifically designed with manual biomechanical perturbations. We developed a validation utility to scan the reasons list within the response text for direction-specific keywords, including "forward," "backward," "left," and "right."

To ensure the reasoning was anatomically specific, the utility further verified if these directionals were associated with relevant anatomical or geometric terms, such as "spinal," "hip," "sagittal," or "coronal." A justification is labeled as *Faithful* if the parsed direction and anatomical feature match the specific perturbation applied during the sample synthesis.

Table 4 compares the faithfulness of the full framework against the baseline.

Table 4: Comparison of reasoning faithfulness across 4-directional perturbations (n = 253).

Perturbation Direction	w/o Symbolic Reasoner & Integrity Verifier	Full Framework	Improvement
Forward (Sagittal)	75.9%	90.7%	+14.8%
Backward (Sagittal)	67.2%	87.5%	+20.3%
Left (Coronal)	54.7%	78.1%	+23.4%
Right (Coronal)	56.3%	80.3%	+24.0%
Average	63.5%	84.2%	+20.7%

The comparative analysis reveals distinct behavioral differences between the baseline and the proposed framework. We observe that the baseline model exhibits a significant bias toward the "forward" direction; it is more likely to incorrectly justify an instability as a forward spinal lean even when the perturbation is in a different direction. This suggests that without symbolic constraints, the model relies on linguistic frequency or common postural tropes rather than visual-physical evidence.

Furthermore, the baseline performs notably better on sagittal plane problems (Forward/Backward) than on coronal plane problems (Left/Right), likely because sagittal stability is more frequently discussed in general biomechanical literature. However, this discrepancy in the baseline is characterized more by "guessing" than accurate reasoning.

Our full framework mitigates these biases, showing a significant improvement in faithfulness. By utilizing the Symbolic Reasoner to map coordinates to specific plane-based rules and the Integrity Verifier to audit the final output, the system ensures that directional keywords are physically grounded. The improvement is most pronounced in the coronal plane, where the framework successfully identifies left/right shifts that the baseline frequently overlooks or misidentifies. This level of granularity is essential for reliable human pose analysis, where the specific direction of instability is as important as the stability verdict itself.

5.3.4 Evaluation on Image Input Based on SMPL Mesh Representations (RQ4)

To evaluate the robustness of the proposed framework and its ability to handle more complex biomechanical reasoning, we extended our evaluation to the HumanEva dataset [57]. This dataset provides a transition from coordinate-based skeletons to dense SMPL 3D Mesh representations [58]. By utilizing the same dataset construction and evaluation methodology as our previous experiments, we collect 1818 distinct SMPL mesh samples from the HumanEva dataset and assess whether the symbolic reasoning pipeline generalizes to higher-dimensional pose data.

As illustrated in the four-stage pipeline (Fig. 4), the Visual Parser (Stage 1) is tasked with mapping mesh surface data to established biomechanical criteria. While skeletal data is sparse, the SMPL mesh requires the parser to interpret volumetric information for premises such as center of gravity and infer arm asymmetry. The stability of the framework's performance suggests that the Symbolic Reasoner (Stage 2) effectively abstracts these complexities into a uniform logical framework regardless of the input modality.

As illustrated in Table 5, the proposed Neuro-Symbolic Reasoning framework demonstrates superior performance across all evaluation metrics compared to state-of-the-art LVLMs. The framework achieves an accuracy of **0.875**, representing a significant margin over the strongest baseline, Qwen2.5-VL-72B (0.757), and the more recent Kimi-k2.5 (0.742).

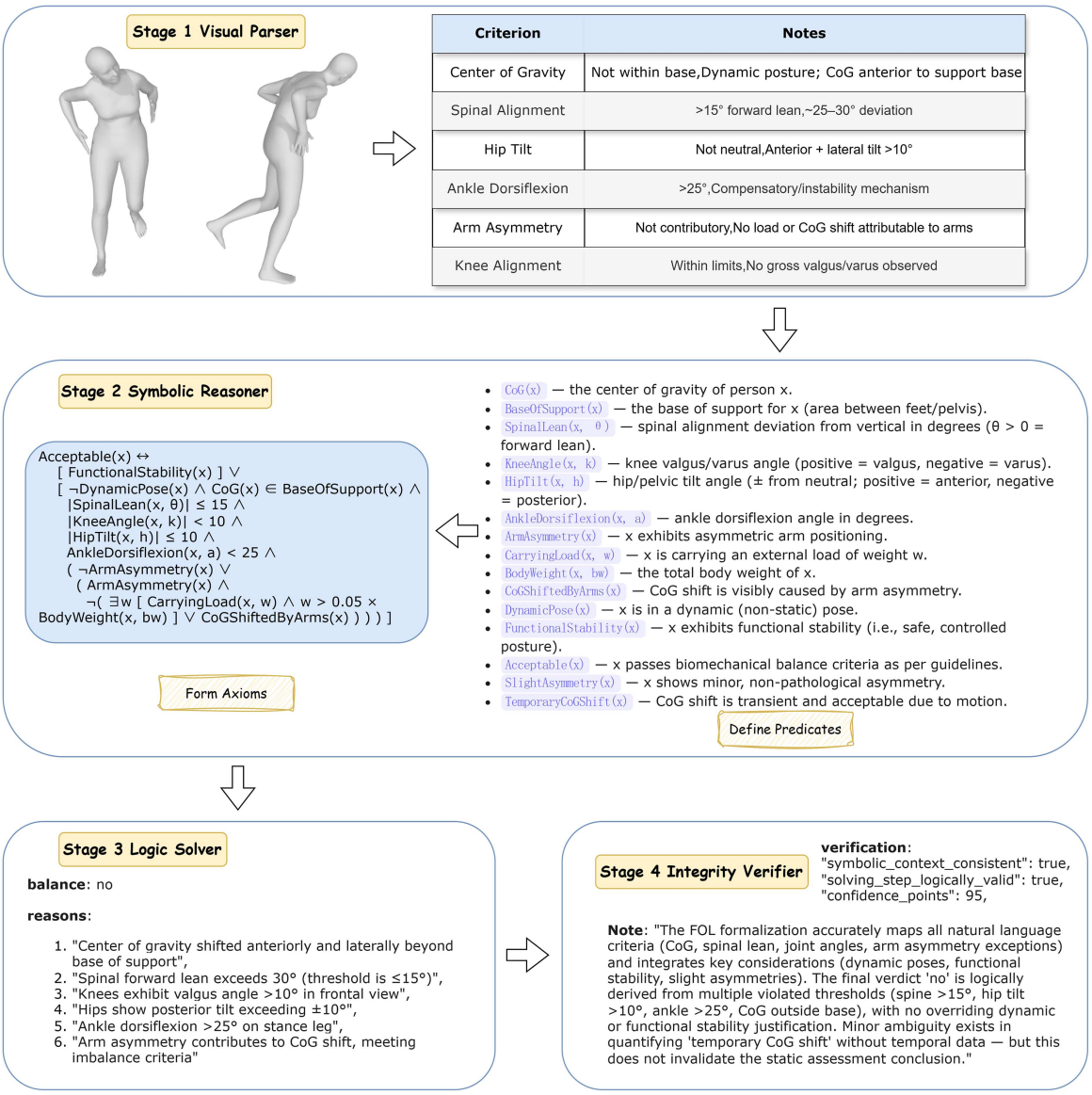

Figure 4: Detailed workflow of the four-stage Neuro-Symbolic Reasoning framework. (Stage 1) Visual Parser extracts biomechanical findings from multi-view mesh data; (Stage 2) Symbolic Reasoner defines predicates and axioms for formal logic mapping; (Stage 3) Logic Solver executes deductive reasoning to determine balance status and generate causal explanations; (Stage 4) Integrity Verifier validates logical consistency and calculates confidence scores.

Table 5: Stability reasoning performance on SMPL-based poses. Bold text indicates the best performance for each metric. The downward arrow (↓) indicates that lower values correspond to better calibration performance.

Model	Accuracy	Precision	Recall	F1 Score	ECE ↓
Neuro-Symbolic Reasoning	**0.875**	**0.800**	**0.942**	**0.865**	**0.111**
Qwen2.5-VL-72B [52]	0.757	0.644	0.832	0.726	0.191
Kimi-k2.5 [59]	0.742	0.633	0.805	0.709	0.198
Qwen3-VL-8B [60]	0.678	0.558	0.735	0.634	0.295
Ministral-3-8B [61]	0.605	0.718	0.586	0.645	0.330

- **Detection Sensitivity and Recall:** The framework achieved a high recall of **0.942**. This suggests that the Logic Solver (Stage 3) is particularly effective at capturing instability in complex SMPL mesh data. By grounding the reasoning in symbolic constraints, the system identifies subtle threshold violations that standalone LVLMs—which rely on heuristic visual patterns—frequently overlook.
- **Precision in Biomechanical Constraints:** While high-parameter models struggle with precise geometric reasoning, our framework's Integrity Verifier (Stage 4) ensures that the output remains logically valid and symbolically consistent. The performance gap is even more pronounced when compared to smaller-scale models, which show a marked inability to handle the high dimensionality of SMPL-based poses.
- **Reliability and Calibration:** Notably, the framework achieves the lowest Expected Calibration Error (ECE) of **0.111**. This indicates that the neuro-symbolic approach not only improves accuracy but also enhances the reliability of the model's confidence scores. The combination of a high F1-score (0.865) and low ECE confirms the system's robustness in handling compensatory mechanisms and dynamic postures (Stage 2) across varied data formats.

The success of the framework on the HumanEva dataset confirms that the symbolic bottleneck provided by the Visual Parser serves as an effective filter for visual noise. This allows the Symbolic Reasoner to bypass the hallucination tendencies of standard LVLMs and focus on the underlying biomechanical truths of the pose.

6 Conclusion and Future Work

This paper presented a novel neuro-symbolic reasoning framework designed to bridge the gap between raw geometric data and high-level biomechanical analysis for 3D human pose balance assessment. By integrating a four-stage pipeline—comprising a Visual Parser, Symbolic Reasoner, Logic Solver, and Integrity Verifier—the proposed framework effectively translates visual skeletal and mesh data into a formal logic framework. Our experimental results, conducted on a specialized evaluation dataset constructed from Human3.6M and HumanEva source data, demonstrate that the framework achieves superior accuracy and recall compared to standalone large vision-language models. Specifically, the inclusion of the Integrity Verifier was shown to significantly reduce calibration error, ensuring that the system's deductions remain logically consistent and biomechanically grounded.

The robustness of this neuro-symbolic approach, validated through our curated benchmarks for both skeleton and SMPL-mesh modalities, confirms its potential for high-stakes applications in clinical rehabilitation and safety monitoring. By grounding inference in established biomechanical principles, such as the Limits of Stability and anatomical range-of-motion constraints, the framework provides a transparent and interpretable method for postural stability analysis. Future research will focus on extending this symbolic bottleneck to temporal motion sequences, enabling the real-time detection of dynamic stability transitions and compensatory movement strategies in naturalistic environments.

6.1 Limitations

Despite the robustness of the Integrity Verifier in mitigating logical inconsistencies, our framework encounters specific challenges related to temporal context. The current model operates primarily on single-frame mesh or skeletal data. As noted in the Stage 4 output, minor ambiguity exists in quantifying transient CoG shifts, as the model cannot easily distinguish between a purposeful dynamic movement and a genuine loss of balance without temporal information.

Another limitation concerns visual ambiguity in the multi-view projections used by the Visual Parser. The framework accepts either skeleton or mesh renderings. In skeleton images, joints and kinematic connections remain explicitly visible, making self-occlusion uncommon. In mesh images, however, overlapping

body surfaces or extreme viewpoints may obscure anatomical landmarks and affect the extraction of spatial predicates, particularly for subtle features such as knee valgus or axial spinal rotation. Although the orthogonal front and side views reduce this ambiguity, robustness under controlled occlusion and viewpoint perturbations remains to be systematically evaluated.

Furthermore, the symbolic axioms rely on fixed biomechanical thresholds, such as a 15-degree spinal lean or specific pelvic tilt angles. In real-world clinical scenarios, these borderline cases often require personalized adjustments based on an individual's age, height, and center-of-mass distribution. The current static axiom set does not yet accommodate these person-specific anatomical variations.

6.2 Future Work

A key direction for future research involves extending the neuro-symbolic framework to a broader spectrum of 3D pose analysis tasks beyond balance assessment. Our modular architecture is fundamentally task-agnostic: the Visual Parser, Symbolic Reasoner, Logic Solver, and Integrity Verifier remain unchanged, requiring only task-specific adaptation of the predicate set and rule base. Potential applications include action recognition, gesture understanding, ergonomic assessment, sports analysis, rehabilitation monitoring, and fall prediction.

Future work will also extend the current static axiom set into a subject-adaptive symbolic framework. Physiological attributes, including age, height, and center-of-mass distribution, can be encoded as additional symbolic facts, while biomechanical thresholds can be parameterized using validated population-specific reference ranges or calibrated functions conditioned on these attributes. For example, a fixed spinal-angle threshold may be replaced with a subject-specific threshold $\theta_{\text{spine}} = g(\text{age}, \text{height}, \text{CoM})$. This design would preserve the explicit and interpretable structure of the logical rules while adapting the reasoning criteria to individual anatomical and physiological characteristics.

Finally, we intend to expand our framework toward video-based extensions and temporal reasoning to address the limitations of single-frame analysis. By transitioning from static First-Order Logic to temporal logic formulations, future iterations of the Symbolic Reasoner will be capable of analyzing movement sequences over time, allowing the system to effectively differentiate between intentional dynamic transitions and accidental instability. Furthermore, this video-based extension will enable continuous motion tracking across full gait cycles and facilitate the use of recursive feedback loops to dynamically calibrate stability confidence scores in naturalistic environments.

Acknowledgement: None.

Funding Statement: This work was supported in part by Hong Kong Reasearch Grants Council (Project 11201185), City University of Hong Kong (Project 9610034 and 9610460).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Yucheng Huang; methodology, Yucheng Huang; software, Yucheng Huang; validation, Jianze Wei; formal analysis, Yucheng Huang; investigation, Yucheng Huang; resources, Xingyu Gao; data curation, Yucheng Huang; writing—original draft preparation, Yucheng Huang; writing—review and editing, Jianze Wei; visualization, Yucheng Huang; supervision, Xingyu Gao and Hong Yan; project administration, Hong Yan; funding acquisition, Xingyu Gao and Hong Yan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The original source datasets utilized in this study are Human3.6M [53] and HumanEva [57]. Specifically, for HumanEva, we use the SMPL mesh representation provided through the AMASS project [62], which is publicly available for download at https://amass.is.tue.mpg.de/download.php under the AMASS/SMPL licensing terms. The original Human3.6M dataset can be accessed directly from its official repository

at http://vision.imar.ro/human3.6m/. The synthetically generated perturbation benchmark (506 samples) derived from the above datasets, along with the inference prompts and evaluation scripts used in our neuro-symbolic reasoning pipeline, are available upon reasonable request from the corresponding author. These materials are provided for non-commercial research purposes to facilitate reproducibility and further investigation.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824–37. doi:10.52202/068431-1800.
2. Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. Adv Neural Inf Process Syst. 2023;36:11809–22.
3. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. Int J Comput Vis. 2022;130(9):2337–48. doi:10.1007/s11263-022-01653-1.
4. Lin X, Zhu M, Dang R, Zhou G, Shu S, Lin F, et al. Clipose: category-level object pose estimation with pre-trained vision-language knowledge. IEEE Trans Circuits Syst Video Technol. 2024;34(10):9125–38.
5. Behzad M. FACET-VLM: facial emotion learning with text-guided multiview fusion via vision-language model for 3D/4D facial expression recognition. Neurocomputing. 2025;657(8):131621. doi:10.1016/j.neucom.2025.131621.
6. Delmas G, Weinzaepfel P, Moreno-Noguer F, Rogez G. PoseEmbroider: towards a 3D, visual, semantic-aware human pose representation. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2024. p. 55–73.
7. Subramanian S, Ng E, Müller L, Klein D, Ginosar S, Darrell T. Pose priors from language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 11–15; Nashville, TN, USA. p. 7125–35.
8. Wang Y, Sun Y, Patel P, Daniilidis K, Black MJ, Kocabas M. PromptHMR: promptable human mesh recovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 11–15; Nashville, TN, USA. p. 1148–59.
9. Qiao Y, Duan H, Fang X, Yang J, Chen L, Zhang S, et al. Prism: a framework for decoupling and assessing the capabilities of VLMs. Adv Neural Inf Process Syst. 2024;37:111863–98.
10. Xu G, Jin P, Wu Z, Li H, Song Y, Sun L, et al. LLaVA-CoT: let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025 Oct 19–23; Honolulu, HI, USA. p. 2087–98.
11. Zhao Q, Lu Y, Kim MJ, Fu Z, Zhang Z, Wu Y, et al. CoT-VLA: visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 11–15; Nashville, TN, USA. p. 1702–13.
12. Zhou X, Sun X, Zhang W, Liang S, Wei Y. Deep kinematic pose regression. In: European Conference on Computer Vision; 2016 Oct 8–16; Amsterdam, The Netherlands. Cham, Switzerland: Springer; 2016. p. 186–201.
13. Zeng A, Yang L, Ju X, Li J, Wang J, Xu Q. SmoothNet: a plug-and-play network for refining human poses in videos. In: European Conference on Computer Vision; 2022 Oct 23–27; Tel Aviv, Israel. Cham, Switzerland: Springer; 2022. p. 625–42.
14. Kang J, Fan W, Li Y, Liu R, Zhou D. 3D human pose estimation using two-stream architecture with joint training. Comput Model Eng Sci. 2023;137(1):607–29. doi:10.32604/cmes.2023.024420.
15. Li M, Hu H, Xiong J, Zhao X, Yan H. TSwinPose: enhanced monocular 3D human pose estimation with JointFlow. Expert Syst Appl. 2024;249:123545.
16. Li P, Wang R, Zhang W, Liu Y, Xu C. Lightweight multi-resolution network for human pose estimation. Comput Model Eng Sci. 2024;138(3):2239. doi:10.32604/cmes.2023.030677.
17. Niu Z, Lu K, Xue J, Qin X, Wang J, Shao L. From methods to applications: a review of deep 3D human motion capture. IEEE Trans Circuits Syst Video Technol. 2024;34(11):11340–59.

18. Shahjahan ATM, Hamza AB. Flexible graph convolutional network for 3D human pose estimation. arXiv:2407.19077. 2024.
19. Islam Z, Hamza AB. Multi-hop graph transformer network for 3D human pose estimation. J Vis Commun Image Represent. 2024;101:104174. doi:10.1016/j.jvcir.2024.104174.
20. Rempe D, Guibas LJ, Hertzmann A, Russell B, Villegas R, Yang J. Contact and human dynamics from monocular video. In: European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. Cham, Switzerland: Springer; 2020. p. 71–87.
21. Zou Y, Yang J, Ceylan D, Zhang J, Perazzi F, Huang JB. Reducing footskate in human motion reconstruction with ground contact constraints. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2020 Mar 1–5; Snowmass Village, CO, USA. p. 459–68.
22. Martinez-Hernandez U, Awad MI, Dehghani-Sanij AA. Learning architecture for the recognition of walking and prediction of gait period using wearable sensors. Neurocomputing. 2022;470:1–10. doi:10.1016/j.neucom.2021.10.044.
23. Shimada S, Golyanik V, Xu W, Theobalt C. Physcap: physically plausible monocular 3D motion capture in real time. ACM Trans Graph. 2020;39(6):1–16. doi:10.48550/arxiv.2008.08880.
24. Shimada S, Golyanik V, Xu W, Pérez P, Theobalt C. Neural monocular 3D human motion capture with physical awareness. ACM Trans Graph. 2021;40(4):1–15. doi:10.1145/3450626.3459825.
25. Xie K, Wang T, Iqbal U, Guo Y, Fidler S, Shkurti F. Physics-based human motion estimation and synthesis from videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, QC, Canada. p. 11532–41.
26. Huang B, Pan L, Yang Y, Ju J, Wang Y. Neural MoCon: neural motion control for physically plausible human motion capture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 6417–26.
27. Du Y, Kips R, Pumarola A, Starke S, Thabet A, Sanakoyeu A. Avatars grow legs: generating smooth human motion from sparse tracking inputs with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 18–22; Vancouver, BC, Canada. p. 481–90.
28. Yuan Y, Wei SE, Simon T, Kitani K, Saragih SJ. SimPoE: simulated character control for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Nashville, TN, USA. p. 7159–69.
29. Luo Z, Iwase S, Yuan Y, Kitani K. Embodied scene-aware human pose estimation. Adv Neural Inf Process Syst. 2022;35:6815–28. doi:10.52202/068431-0494.
30. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171. 2022.
31. Wang X, Zhou D. Chain-of-thought reasoning without prompting. Adv Neural Inf Process Syst. 2024;37:66383–409. doi:10.52202/079017-2123.
32. Wang Y, Wu S, Zhang Y, Yan S, Liu Z, Luo J, et al. Multimodal chain-of-thought reasoning: a comprehensive survey. arXiv:2503.12605. 2025.
33. Wang W, Yang Y, Wu F. Towards data-and knowledge-driven AI: a survey on neuro-symbolic computing. IEEE Trans Pattern Anal Mach Intell. 2024;47(2):878–99.
34. Bao Y, Xing T, Chen X. Confidence-based interactable neural-symbolic visual question answering. Neurocomputing. 2024;564(8):126991. doi:10.1016/j.neucom.2023.126991.
35. Jena M, Khan N, Lee MY, Rho S. Neuro-symbolic graph learning for causal inference and continual learning in mental-health risk assessment. Comput Model Eng Sci. 2026;146(1):75119. doi:10.32604/cmes.2025.075119.
36. Pan L, Albalak A, Wang X, Wang W. Logic-LM: empowering large language models with symbolic solvers for faithful logical reasoning. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023 Dec 6–10; Singapore. p. 3806–24.
37. Olausson T, Gu A, Lipkin B, Zhang C, Solar-Lezama A, Tenenbaum J, et al. LINC: a neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 5153–76.

38. Ranaldi L, Valentino M, Freitas A. Improving chain-of-thought reasoning via quasi-symbolic abstractions. arXiv:2502.12616. 2025.
39. Liu C, Yuan Y, Yin Y, Xu Y, Xu X, Chen Z, et al. Safe: enhancing mathematical reasoning in Large Language Models via retrospective step-aware formal verification. arXiv:2506.04592. 2025.
40. Luo L, Zhao Z, Haffari G, Li YF, Gong C, Pan S. Graph-constrained reasoning: faithful reasoning on knowledge graphs with large language models. arXiv:2410.13080. 2024.
41. Thakkallapelly R, Bandla SL, Chava SC, Rathod K, Gudelli VR, Sadat QT. Neuro symbolic AI for context-aware decision making in real time systems. In: Proceedings of the 2026 IEEE International Conference for Convergence in Computing Technology (I3CTCON); 2026 Mar 14–15; Lonavala, India. New York, NY, USA: IEEE; 2026. p. 1–7.
42. Delmas G, Weinzaepfel P, Lucas T, Moreno-Noguer F, Rogez G. PoseScript: 3D human poses from natural language. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2022. p. 346–62.
43. Delmas G, Weinzaepfel P, Moreno-Noguer F, Rogez G. PoseFix: correcting 3D human poses with natural language. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2–6; Paris, France. p. 15018–28.
44. Wang R, Ma C, Li G, Xu H, Li Y, Wang Z. You think, You ACT: the new task of arbitrary text to motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025 Oct 19–23. Honolulu, HI, USA. p. 12012–22.
45. Feng Y, Lin J, Dwivedi SK, Sun Y, Patel P, Black MJ. ChatPose: chatting about 3D human pose. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. New York, NY, USA: IEEE; 2024. p. 2093–103.
46. Feng D, Guo P, Peng E, Zhu M, Yu W, Wang P. PoseLLaVA: pose centric multimodal LLM for fine-grained 3D pose manipulation. Proc AAAI Conf Artif Intell. 2025;39:2951–9.
47. Li Y, Hou R, Chang H, Shan S, Chen X. UniPose: a unified multimodal framework for human pose comprehension, generation and editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 11–15; Nashville, TN, USA. p. 27805–15.
48. Zhang J, Peng J, Wang K. Athlete posture estimation and analysis based on embodied artificial intelligence. Image Vis Comput. 2025;162(1):105598. doi:10.1016/j.imavis.2025.105598.
49. Xu J, Guo Y, Peng Y. FinePOSE: fine-grained prompt-driven 3D human pose estimation via diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 17–21; Seattle, WA, USA. p. 561–70.
50. Li J, Cao J, Zhang H, Rempe D, Kautz J, Iqbal U, et al. GENMO: a GENeralist model for human MOtion. arXiv:2505.01425. 2025.
51. Lin J, Chang J, Liu L, Li G, Lin L, Tian Q, et al. Being comes from not-being: open-vocabulary text-to-motion generation with wordless training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 18–22; Vancouver, BC, Canada. p. 23222–31.
52. Bai S, Chen K, Liu X, Wang J, Ge W, Song S, et al. Qwen2.5-VL technical report. arXiv:2502.13923. 2025.
53. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell. 2013;36(7):1325–39. doi:10.1109/tpami.2013.248.
54. Elizabeth B, Garima G, Sudhakar R, Karen Y. Anthropometry, biomechanics, and strength; 2023 [cited 2026 Jan 1]. Available from: https://www.nasa.gov/wp-content/uploads/2023/12/ochmo-hb-004-rev-a-dec2023.pdf.
55. Juras G, Słomka K, Fredyk A, Sobota G, Bacik B. Evaluation of the limits of stability (LOS) balance test. J Hum Kinet. 2008;19(1):39–52. doi:10.2478/v10078-008-0003-0.
56. Naeini MP, Cooper G, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. Proc AAAI Conf Artif Intell. 2015;29(1):2901–7. doi:10.1609/aaai.v29i1.9602.
57. Sigal L, Balan AO, Black MJ. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int J Comput Vis. 2010;87(1):4–27.
58. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: a skinned multi-person linear model. Semin Graph Pap Push Boundaries. 2023;2:851–66.

59. Team K, Bai T, Bai Y, Bao Y, Cai S, Cao Y, et al. Kimi K2.5: visual agentic intelligence. arXiv:2602.02276. 2026.
60. Bai S, Cai Y, Chen R, Chen K, Chen X, Cheng Z, et al. Qwen3-VL technical report. arXiv:2511.21631. 2025.
61. Liu AH, Khandelwal K, Subramanian S, Jouault V, Rastogi A, Sadé A, et al. Ministral 3. arXiv:2601.08584. 2026.
62. Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black MJ. AMASS: archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 5442–51.