CCC 2025

Proceedings of the Creative Construction Conference (2025)
Edited by: Miroslaw J. Skibniewski & Miklós Hajdu & Žiga Turk
https://doi.org/10.22260/CCC2025/0060

FEASIBILITY OF AN INTEGRATED MULTI-MODEL APPROACH FOR DYNAMIC MUSCULOSKELETAL DISORDER RISK ASSESSMENT

Kangrui Ren[1], Eren (M.) Shahrokhi[1], Ali Golabchi[1,2], Gaang Lee[1]
1 University of Alberta, Edmonton, Canada
2 EWI Works International Inc., Edmonton, Canada

Abstract

High-load, repetitive construction tasks pose significant risks for the development of musculoskeletal disorders (MSDs), which adversely affect workers' health, productivity, and quality of life, making accurate risk quantification crucial for timely prevention. In recent years, computer vision (CV) has demonstrated significant potential in assessing MSD risks by offering an automated, contactless, and adaptable monitoring approach in complex environments. Existing approaches that rely on traditional biomechanical frameworks (e.g., REBA)—which assess MSD risks based on individual static frames by averaging the results—fail to capture the nonlinearly accumulating nature of MSDs risks in continuous motion and overlook individual variability in dynamic indicators such as amplitude, frequency, and smoothness. These limitations introduce bias, underscoring the need for an adaptive, personalized dynamic MSD risk scoring technique. To address these limitations, this study investigated the feasibility of a multi-model framework that integrates an enhanced Spatiotemporal Graph Convolutional Network (ST-GCN), a lightweight Transformer, and a diffusion model. It first employed an ST-GCN with dynamic adjacency and adaptive hypergraphs to model short-term dependencies; next, a Transformer refined long-term motion patterns; and finally, a diffusion model generated personalized risk score distributions to track MSDs risk evolution. The framework was trained on the HMR 2.0 dataset and evaluated on both HMR2.0 and MoYo dataset. Joint information was extracted for comprehensive MSDs risk assessment, and the resulting risk scores were compared with those from traditional biomechanical static single-frame methods. The results demonstrated that the proposed approach captured individual variability and nonlinearly accumulating MSDs risks more effectively, confirming the superiority of the proposed dynamic MSDs risk assessment. The findings underscore the significant potential of the proposed model for video-based MSDs risk assessment to enhance the accuracy of automated, real-time MSDs risk monitoring in high-load and dynamic environments such as construction sites.

Keywords: Dynamic Modelling, Multi-Model Framework, Musculoskeletal Disorder Risk Assessment, Intelligent Construction Management, Personalized Decision Support.

1. Introduction

Musculoskeletal disorders (MSDs) are among the most prevalent occupational health problems in construction, particularly affecting the back, neck, and upper limbs [1][2]. Construction workers experience a markedly higher risk of MSDs than workers in most other industries, largely because of prolonged heavy labor, irregular load handling, and frequent awkward postures. Economically, MSDs drive up medical expenditures and compensation claims, causing project delays and aggravating workforce shortages. Therefore, systematically assessing MSD risks is essential for designing targeted interventions that support effective prevention and management [3].

Traditionally, MSD risk has been assessed with checklist-based ergonomic risk assessment (ERA) tools. Classic examples—including the Rapid Upper Limb Assessment (RULA) [4] and Rapid Entire Body Assessment (REBA) [5]—assign risk scores from visual inspections of static postures (e.g., arm and trunk angles) without interrupting the workflow [6]. Although these methods can be applied quickly, they are labour-intensive, subjective, and susceptible to inter-rater variation. Moreover, they do not scale to continuous monitoring, which limits their usefulness in modern, dynamic workplaces [7].

Corresponding author email address: kangrui1@ualberta.ca

Recent computer-vision approaches automate ERA by extracting 2D/3D human poses from individual RGB frames and then computing ergonomic indices (e.g., RULA, REBA) from the resulting joint angles [8]. These methods either (i) calculate the scores directly [9], or (ii) use convolutional neural networks trained on labelled posture data to regress risk [10]. Although they reduce manual effort and permit simultaneous monitoring of multiple workers, their frame-level perspective ignores temporal posture dynamics and can underestimate the cumulative MSD risk that accrues over time [11].

To mitigate the limitations of frame-level analysis, recent AI-based ERA research integrates temporal modeling into skeletal data streams. Graph Convolutional Networks (GCNs), recurrent architectures such as Long Short-Term Memory (LSTM) networks, and Transformer models capture inter-joint dependencies over time and yield more accurate risk estimates [12]. Nonetheless, two challenges remain in construction environments. First, these models generally assume well-defined, repetitive tasks, whereas on-site construction work is fluid and context-rich; skeletal data alone may therefore overlook critical environmental cues. Second, they rarely personalize risk estimates to worker-specific attributes such as body composition or physical fitness, even though individuals performing identical tasks can face markedly different risk levels.

To address these gaps, this study investigates the feasibility of a multi-model AI pipeline for dynamic MSD-risk assessment. The pipeline adopts a two-stream architecture. The skeleton stream extracts 3D joint sequences from monocular video, refines them with spatiotemporal graph-neural-network modelling, and augments them with rule-based ergonomic cues. In parallel, the video stream captures scene context through a Vision Transformer and encodes worker-specific attributes. The two streams are fused via cross-modal attention, and the resulting joint representation is passed to a diffusion-based head that outputs a full probability distribution of MSD risk. By unifying spatial skeletal cues, temporal motion history, visual context, and personal factors, the framework simultaneously addresses task complexity, risk accumulation, and individual variability in a single end-to-end system.

2. Methods

2.1. Pipeline Overview

In dynamic construction environments, ERA must handle articulated postures, evolving visual contexts, and inter-individual variability. To meet these demands, we propose a multimodal AI pipeline that ingests monocular RGB video and produces a calibrated probability distribution of MSD risk. The video is processed in parallel by two complementary streams—one focused on skeletal kinematics and the other on visual context—whose features are subsequently fused to yield a holistic risk estimate (

Figure 1).

The Skeleton Stream first employs SKEL-HSMR [13] to regress 3D joint coordinates for every frame. These coordinates are fed into a temporal Channel-wise Topology-Refinement GCN (CTR-GCN) [14], enhanced with dynamic graph and hypergraph convolutions to model high-order spatiotemporal dependencies. In parallel, a rule-based branch extracts interpretable ergonomic metrics, including REBA scores, joint-excursion angles, and L5/S1 lumbar-load estimates [15]. The Video Stream applies a Video Swin Transformer [16] to the same frames and concatenates the resulting scene features with a fixed personal-attribute embedding that captures worker-specific susceptibility. A cross-modal attention fusion branch then aligns the two feature sets, producing a unified latent representation that integrates skeletal dynamics, visual context, and personal factors. Finally, a conditional denoising diffusion probabilistic model (DDPM) [17] leverages this fused representation to sample a calibrated probability distribution over MSD-risk levels. This architecture delivers dynamic, personalised ERA while preserving interpretability through its rule-based branch and explicit uncertainty quantification.

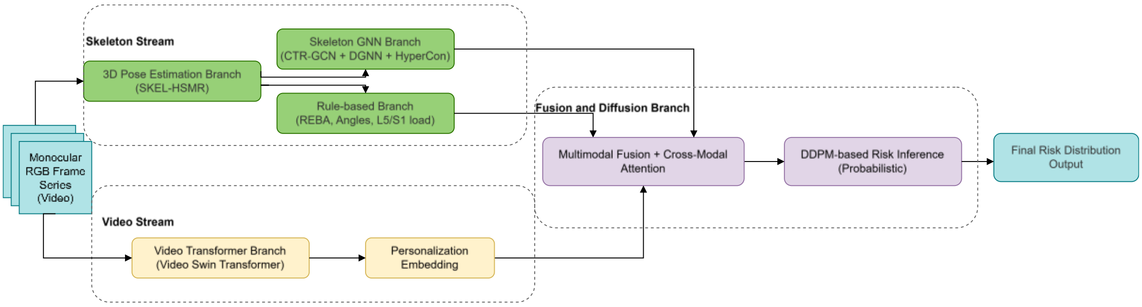
Skeleton Stream
Skeleton GNN Branch (CTR-GCN + DGNN + HyperCon)
3D Pose Estimation Branch (SKEL-HSMR)
Rule-based Branch (REBA, Angles, L5/S1 load)
Fusion and Diffusion Branch
Monocular RGB Frame Series (Video)
Multimodal Fusion + Cross-Modal Attention
DDPM-based Risk Inference (Probabilistic)
Final Risk Distribution Output
Video Stream
Video Transformer Branch (Video Swin Transformer)
Personalization Embedding

Figure 1. Overview of the proposed multimodal pipeline.

2.2. Skeleton Stream

2.2.1. 3D Pose Estimation Branch

Because downstream ergonomic metrics rely on accurate joint kinematics, we first reconstruct biomechanically plausible 3D skeletons from monocular video with SKEL-HSMR. The model couples a Vision Transformer (ViT) encoder with a lightweight decoder that regresses 3D joint coordinates directly. Unlike conventional SMPL pipelines, where every joint is treated as an unconstrained three-degree-of-freedom (3-DoF) rotation, often yielding non-physiological poses [18], SKEL-HSMR embeds hard-coded anatomical DoF and joint-limit priors (e.g., the knee hinge is restricted to $0°–135°$).

During inference, the ViT encoder produces visual tokens that the decoder maps to three parameter sets: (i) a 46-dimensional pose vector θ, (ii) a body shape vector S, and (iii) a global camera pose parameter p_m. A differentiable forward-kinematics layer then converts these parameters into fully articulated 3-D joint positions for each frame [19]. By construction, the resulting skeletons satisfy biomechanical constraints, eliminating post-processing and providing reliable input for the subsequent GNN, rule-based, and vision branches. An overview of the complete architecture appears in Figure 2.

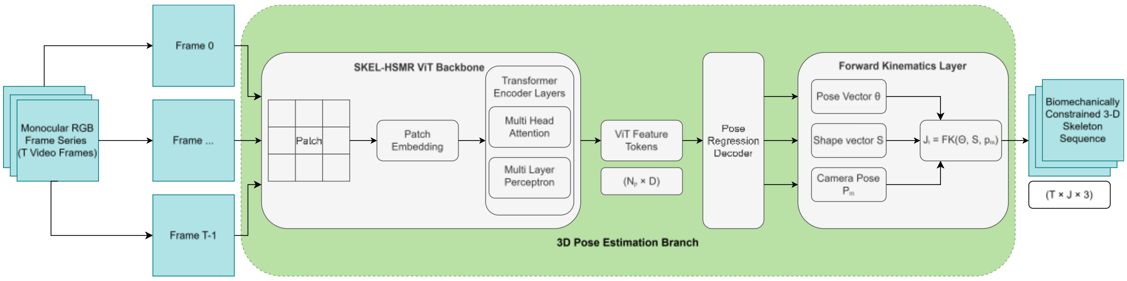

Figure 2. Architecture of the SKEL-HSMR-based 3D pose-estimation branch.

2.2.2. Skeleton Graph Neural Network Branch

Capturing fine-grained joint interactions over time is critical for ERA, so our skeleton branch (Figure 3) builds on the CTR-GCN. For each video clip, we obtain a 3D joint sequence of (T×J×3), reshape it to (C×T×V) and feed it to a two-stream encoder that processes joint coordinates together with bone vectors (differences between adjacent joints), thereby fusing absolute motion with relative-limb dynamics [20]. CTR-GCN first learns a coarse global adjacency prior and then refines it on a per-channel basis, providing far greater flexibility than fixed-topology GCNs.

To capture higher-order and time-varying dependency, we insert hypergraph-convolution layers [21][22] that link groups of joints via hyperedges, allowing the network to model the complex synergies involved in motions such as bending. In parallel, DGNN blocks [23] update adjacency weights frame by frame, so the graph continuously reflects changing joint relevance—offering finer-grained adaptation than earlier semantic GCN schemes [24]. After these spatiotemporal layers, global pooling produces a compact skeleton embedding that concentrates key risk cues (e.g., lumbar loading). This embedding can be fused with other modalities or fed directly to the prediction head. By integrating two-stream inputs, channel-wise topology refinement, hypergraph reasoning, and DGNN-driven dynamics, the skeleton branch provides sensitive, comprehensive detection of high-risk postures and movements.

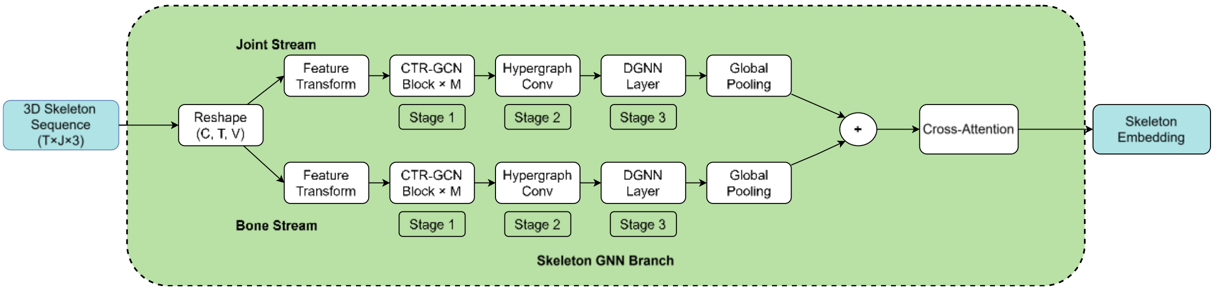

Figure 3. Architecture of the skeleton GNN branch.

2.2.3. Rule-Based Branch

Purely data-driven models often lack interpretability—a crucial prerequisite for real-world deployment. To bridge this gap, we integrate a rule-based feature module that applies expert-defined ergonomic rules to every frame, injecting official safety standards directly into the pipeline. The module produces transparent metrics that map cleanly onto established guidelines, thereby strengthening user trust [25].

A primary function of this module is the automatic generation of REBA scores. Following the CREBAS protocol [26], it discretizes trunk, neck, and limb angles extracted from the 3-D skeleton, adds twist- and load-related penalties, and merges the corresponding lookup tables to produce a deterministic risk rating that matches expert assessment. It simultaneously reports the underlying joint angles—trunk flexion, shoulder abduction, elbow/knee flexion, and neck inclination—allowing practitioners to reference values routinely used in ergonomic checklists [27]. To capture load effects that posture alone cannot, a simplified biomechanical model estimates L5/S1 compressive force from anthropometrics, trunk flexion, and external load mass. Despite its quasi-static formulation, this estimate agrees within 90% of full motion-capture analyses. Together, the REBA score and compressive-force metric ground the assessment in established rules while quantifying both positional and load-bearing risk.

Rule-based features are computed for every frame, optionally aggregated across the entire task, and concatenated with the deep-learning embeddings before the final classifier. This hybrid input grounds the predictions in established ergonomic standards and makes high-risk outputs more transparent (e.g., linking excessive trunk flexion to an elevated REBA score). Recent studies show that combining deterministic metrics with learned representations enhances both accuracy and interpretability, making it easier for safety practitioners to understand and trust the system [28].

2.3. Vision Stream

Skeleton sequences describe joint motion but miss visual appearance and context—lighting, workspace layout—that also shape ergonomic risk. To fill this gap, we add a parallel Video Vision Transformer (ViT) branch that extracts spatiotemporal and contextual cues directly from RGB frames [29] (Figure 4). The raw video stream reveals subtle muscle tremors, object-handling patterns, and environmental hazards such as uneven flooring, providing safety signals unavailable to the posture-only branch.

We use the Video Swin Transformer as the backbone. Its shifting 3D window–attention mechanism restricts computation to local regions and gradually enlarges the receptive field across successive layers, achieving a favorable speed–accuracy trade-off. Pre-trained Kinetics weights are fine-tuned on our dataset, enabling the model to learn both motion cues—such as velocity and acceleration—and contextual features, including tools and obstacles, directly from the raw frames.

Ergonomic risk varies widely among individuals, and attributes such as body-mass index (BMI) have been shown to affect injury likelihood markedly [30]. To capture this variability, we insert a learnable attribute token to encode worker metadata (height, weight, BMI) at the start of the video-patch sequence. Processed alongside the patches, this token allows the self-attention mechanism to fuse personal and visual cues; as a result, the ViT branch can, for example, assign greater importance to subtle joint deviations in a worker with a high BMI.

The last Video Swin layer outputs a pooled vector that integrates motion, appearance, environmental cues, and the personal information carried by the attribute token. We concatenate this vector with the skeleton-GNN features and the rule-based metrics and pass the fused representation to the fusion-and-diffusion branch for risk prediction. By combining scene context with subject-specific data, the model delivers more robust and personalised estimates than skeleton cues alone.

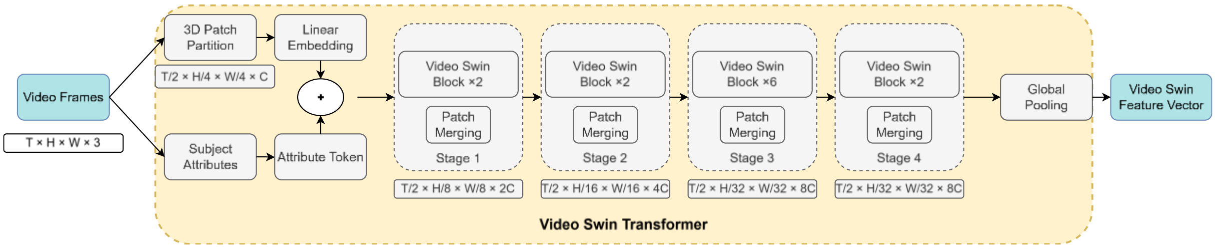

Figure 4. Architecture of the Video Swin Transformer branch.

2.4. Multimodal Fusion and Diffusion

After the skeleton, video, and rule-based streams extract their complementary cues, the resulting feature tensors are concatenated and passed to a cross-modal Transformer that performs bidirectional attention across modalities. Skeletal queries attend to visual and rule tokens, allowing composite patterns—such as an awkward arm posture aligned with an abnormal joint-angle reading and an elevated REBA score—to emerge within a shared latent space. The resulting embedding unifies motion dynamics, scene context, and ergonomic indices into a single, holistic representation of musculoskeletal risk. This multimodal embedding is then fed to a DDPM, which iteratively drives latent noise toward the learned risk manifold and outputs a calibrated probability distribution over risk levels. The distribution's mean serves as a point estimate, while its variance quantifies predictive uncertainty—crucial in safety-critical settings where even low-probability, high-impact events demand attention. Working in tandem, the cross-modal Transformer and DDPM convert heterogeneous cues into a principled, uncertainty-aware assessment that links data-driven insights to established ergonomic standards.

3. Training Implementation

3.1. Training Overview

To minimize manual labelling while still capturing the full complexity of ergonomic risk, we adopt a three-phase curriculum that progressively enriches supervision using publicly available 3D pose resources. Phase 1 pre-trains the two geometry-centred backbones—a Vision Transformer (ViT) that lifts single RGB frames to SKEL-HSMR parameters, and a spatiotemporal skeleton GNN—on the large, diverse HMR 2.0 corpus [31] specifically its four largest subsets: Human3.6M [32], MPI-INF 3DHP, COCO, and MPII. Phase 2 freezes those weights and learns a compact 16-dimensional latent code that fuses the pretrained features with rule-based ergonomic cues, remaining kinematically faithful while aligning with REBA scores. Phase 3 attaches a conditional DDPM that converts each frame-level latent into a calibrated probability distribution over risk outcomes, thereby modelling predictive uncertainty. Training relies solely on public data—HMR 2.0 for supervision—and evaluation is conducted on held-out Human3.6M and MOYO [33] test sets.

3.2. Phase 1: Pretraining Pose Estimation and Skeleton GNN

We train a Vision-Transformer-based 3D pose module that regresses a full human model from a single RGB image. The ViT encodes the image and outputs the parametric skeleton q (joint rotations) and shape β We minimize a composite loss $\mathcal{L}_{\text{pose}}$ (1) consisting of: A 2D reprojection loss $\mathcal{L}_{2D}$(2) that penalizing differences between projected 3D joints $\Pi(J_j)$ and ground-truth 2D key points u_j; A 3D parameter loss $\mathcal{L}_{3D}$(3) that matches the pseudo ground-truth pose q^* and shape β^*; and A Regularizer $\mathcal{L}_{\text{reg}}$ (4) enforces joint-angle limits and bone-length consistency for physical plausibility.

$$\mathcal{L}_{pose} = \mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{1}$$

$$\mathcal{L}_{2D} = \sum_{j} \| \Pi(J_j) - u_j \|^2 \tag{2}$$

$$\mathcal{L}_{3D} = \| q - q^* \|^2 + \| \beta - \beta^* \|^2 \tag{3}$$

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{lim}} \sum_{i} \left[\max(0, \theta_i - \theta_i^{\max})^2 + \max\left(0, \theta_i^{\min} - \theta_i\right)^2\right] + \lambda_{\text{bone}} \sum_{(i,j)\in\mathcal{E}} \left(\| \mathrm{J}_i - \mathrm{J}_j \|_2 - \ell_{ij}\right)^2 \tag{4}$$

To further improve accuracy, we employ an iterative refinement scheme. At the end of each epoch a lightweight inverse-kinematics (IK) solver updates the pseudo-ground-truth pose q^* so that the projected joints align more closely with the 2D key-points. The refined pair (q^*, β^*) is then passed through forward kinematics, J_j = FK (q^*, β^*) and these joints supervise the next epoch.

In parallel a spatio-temporal CTR-GCN is pre-trained on the same sequences. Its objective combines three terms: In parallel we pre-train a spatio-temporal CTR-GCN on the same poses. Its training objective combines three terms: a reconstruction loss $\mathcal{L}_{\text{AE}}$ (5) that forces the network to reproduce the input joints; a one-step prediction loss $\mathcal{L}_{\text{pred}}$ (6) that encourages temporal-dynamics modelling; and a smooth loss $\mathcal{L}_{\text{smooth}}$ (7) that penalises large joint velocities.

$$\mathcal{L}_{\text{AE}} = \sum_{j} \| \hat{J}_j - J_j \|^2 \tag{5}$$

$$\mathcal{L}_{\text{pred}} = \sum_{j} \| \hat{J}_j(t+1) - J_j(t+1) \|^2 \tag{6}$$

$$\mathcal{L}_{\text{smooth}} = \sum_{j} \| \hat{\mathrm{J}}_j^{t+1} - \hat{\mathrm{J}}_j^{t} \|^2 \tag{7}$$

The total loss $\mathcal{L}_{\text{GNN}}$ (8) is comprised of these three terms. Because the GNN is trained on joints that are continuously refined by IK, its latent features encode both accurate kinematic structure and short-term dynamics—assets that later strengthen the multimodal ergonomic-risk fusion.

$$\mathcal{L}_{\text{GNN}} = \mathcal{L}_{\text{AE}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}. \tag{8}$$

3.3. Phase 2: Fusion Training for Interpretable Risk Representation

Phase 2 learns a 16-D, single-frame risk embedding by fusing three frozen feature streams: visual tokens v_t from the ViT's penultimate layer, kinematic features g_t from the skeleton GNN, and heuristic scores r_t from the rule-based module. The concatenated vector is fed through a two-layer perceptron to yield the latent code z_t (9). The fusion network is trained end-to-end with the composite loss $\mathcal{L}_{fusion}$ (10), which balances four terms: a pose-reconstruction loss $\mathcal{L}_{\text{pose-rec}}$ (11) decodes z_t back to 3D joint angles and penalises deviations from the original pose, anchoring the latent to valid human kinematics; a risk-alignment loss $\mathcal{L}_{risk}$ (12) maps z_t to a predicted score $\hat{y}_t$ and penalises its discrepancy from the reference ergonomic metric (e.g., REBA), thereby distilling expert heuristics into the embedding; a temporal-smoothness loss T_s (13) discourages abrupt changes between consecutive embeddings, reflecting the ergonomic principle that micro-adjustments—not jerks—reduce musculoskeletal load; and a Accumulation loss $\mathcal{L}_{accum}$ (14) raises the target risk whenever a posture persists: if a pose lasts T frames, a duration-dependent increment Δ_T is added to the reference score, encoding the well-known fatigue effect of sustained static postures.

$$z_t = \mathrm{FC}_2\left(\mathrm{ReLU}\big(\mathrm{FC}_1([v_t; g_t; r_t])\big)\right) \tag{9}$$

$$\mathcal{L}_{fusion} = \lambda_{pose}\mathcal{L}_{pose\text{-rec}} + \lambda_{risk}\mathcal{L}_{risk} + \lambda_{accum}\mathcal{L}_{accum} \tag{10}$$

$$\mathcal{L}_{\text{pose-rec}} = \sum_{j} \| \hat{J}_j - J_j \|^2 \tag{11}$$

$$\mathcal{L}_{risk} = \| \hat{y}_t - y_t^{REBA} \|^2 \tag{12}$$

$$T_s = \| z_{t+1} - z_t \|^2 \tag{13}$$

$$\mathcal{L}_{accum} = \| \hat{y}_{t+T} - (y_t^{REBA} + \Delta_T) \|^2 \tag{14}$$

Together, Equations (9) – (14) ensure that z_t preserves precise kinematics, aligns with conventional REBA scoring, remains temporally coherent, and explicitly captures risk accumulation over time.

3.4. Phase 3: Diffusion-Based Probabilistic Risk Prediction

To capture the variability inherent in ergonomic risk, we append a conditional Denoising Diffusion Probabilistic Model (DDPM) to the fused embedding z_t. Rather than outputting a single value, the DDPM learns a full conditional distribution $p(y_t \mid z_t)$ over possible risk outcomes y_t (e.g., injury risk levels), thereby reflecting differences in individual tolerance, environment, and task context.

The diffusion head is a U-Net conditioned on z_t. During training we corrupt an initial risk label y_0 with Gaussian noise over T=100 steps, using a linear schedule $\beta_1, \dots, \beta_T$ from 10^{-4} to 0.02. At each iteration, we sample *t* from $\{1, \dots, T\}$, add noise to obtain y_t, and train the network conditional $\mathrm{U} - \mathrm{Net}\epsilon_{\phi}(y_t, t \mid z_t)$ to predict the added noise ε. The objective is the standard DDPM loss (15), which minimises the mean-squared error between the true and predicted noise.

$$\mathcal{L}_{diff} = \mathbb{E}_{y_0,\epsilon,t}\big[\| \epsilon - \epsilon_{\phi}(y_t, t \mid z_t) \|^2\big] \tag{15}$$

After convergence, reverse diffusion yields multiple samples $\hat{y}_0$ from $p(y_0 \mid z_t)$, providing a calibrated risk distribution. For instance, an awkward posture embedding may return a high probability of "moderate risk" and a smaller tail of "extreme risk," faithfully reflecting real-world uncertainty.

4. Model Evaluation and Discussion

Section 4 adopts a two-part, four-experiment evaluation protocol to highlight the strengths of the proposed framework. In subsection 4.1 we benchmark five low-complexity actions, showing that the predicted risk distributions align closely with conventional single-value REBA scores and therefore satisfy point-wise accuracy requirements. Subsection 4.2 comprises three targeted tests that (i) quantify the model's ability to capture risk accumulation during prolonged tasks, (ii) examine its sensitivity to inter-subject variability, and (iii) assess its robustness to camera-view changes and partial occlusions in complex motions. Taken together, these experiments demonstrate the feasibility of the proposed model.

4.1. Low complexity action risk evaluation

Five repeatable, low-complexity actions from Human3.6M—*Taking Photo, Sitting on Chair, Walking Dog, Walking,* and *Walking Together*—were selected because they span the static-to-mildly-dynamic spectrum while avoiding severe occlusions and extreme viewing angles. From three held-out subjects we sampled 150 consecutive frames per action, yielding a clean evaluation set of 2 250 frames. Each frame was evaluated in three ways: (i) the deterministic REBA score computed with CREBAS; (ii) the single-value estimate produced by the MTL-ERA-S baseline; and (iii) the full risk distribution generated by the proposed model, whose mean (μ) serves as the point estimate and whose standard deviation (σ) is retained for later analysis. Because Human3.6M contains no ergonomic annotations, we built a pseudo–ground-truth risk curve by running a conventional pipeline—3-D joint extraction followed by automatic REBA scoring—on every frame and smoothing the result with a Savitzky–Golay filter [34].

Figure 5 overlays this reference curve (red) on the three predicted densities. For every action the blue mode aligns with the red peak, indicating an unbiased centre, while the blue confidence band widens from 5 units for Sitting on Chair to 8 units for Walking Dog, showing that the model's uncertainty increases with range of motion. Figure 6 pools reliability diagrams across the five actions: predicted means are binned into ten equal-frequency intervals and compared with empirical frequencies. The blue curves trace the diagonal closely, confirming that the distributional predictions are well calibrated.

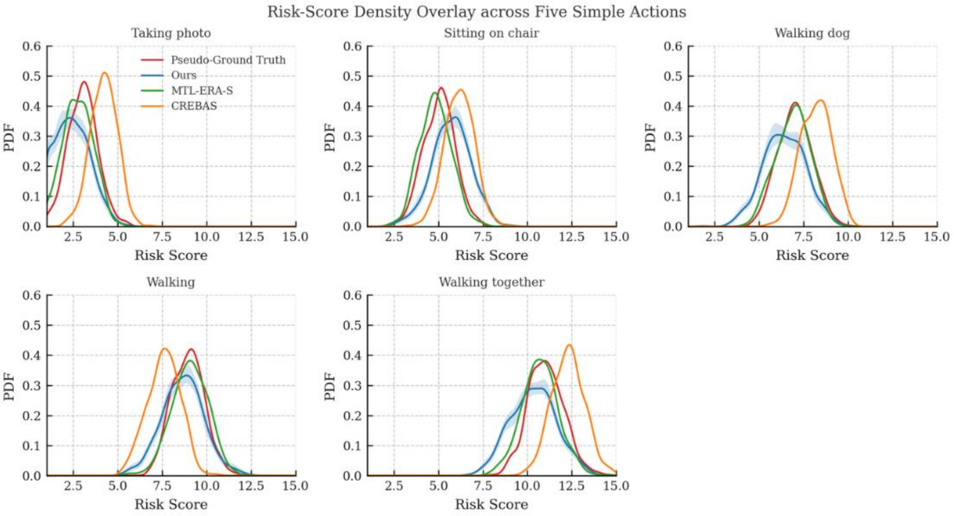

Figure 5. Risk Score Density Overlay across Five Simple Actions.

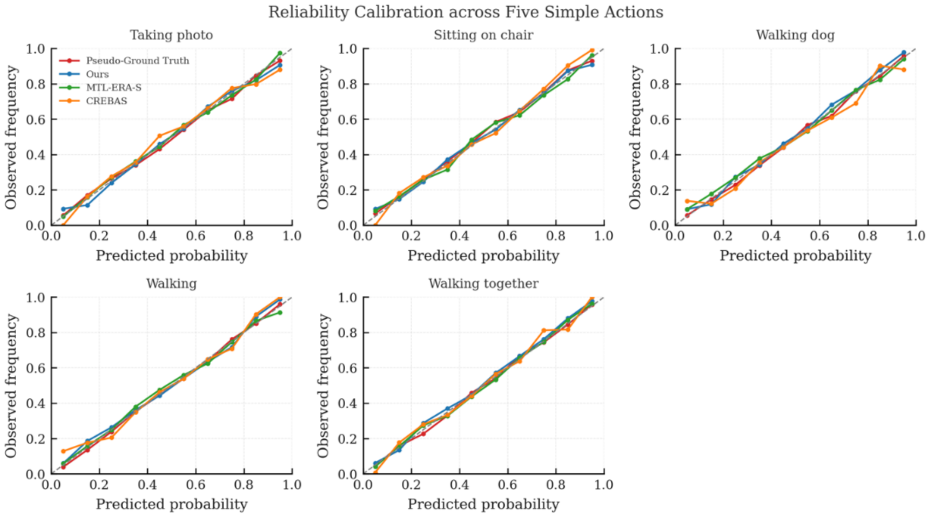

Figure 6. Reliability Calibration across Five Simple Actions.

4.2. Targeted tests

4.2.1. Temporal Modeling of Risk Accumulation

We extend the frame-level analysis to cumulative exposure. Each source clip is looped until its total duration reflects task intensity; the resulting clip lengths are listed in Figure 7. The visual content remains unchanged—only the exposure time increases. At every frame t we compute three cumulative measures from the clip start to frame t: (i) the running-mean REBA score produced by CREBAS, (ii) the running-mean estimate of the MTL-ERA-S baseline, and (iii) the risk distribution yielded by the proposed model.

Figure 7 shows that the predicted mean risk μ continues to rise in all tasks, whereas both baselines plateau after the first few seconds. For Sitting on Chair, CREBAS and MTL-ERA-S stabilise near 4.3–4.5 risk units, but our μ exceeds 5 and its standard deviation σ widens, indicating growing uncertainty as trunk flexion persists. In Walking Dog, successive gait cycles drive μ from 6.5 to 8.8, confirming that the model integrates load over time. Vertical dashed lines mark clip boundaries; the continuity of μ across these lines demonstrates that the network tracks cumulative exposure rather than clip position. Hence, the proposed distributional metric captures long-term musculoskeletal stress and provides calibrated uncertainty, whereas the two running-mean baselines quickly lose sensitivity.

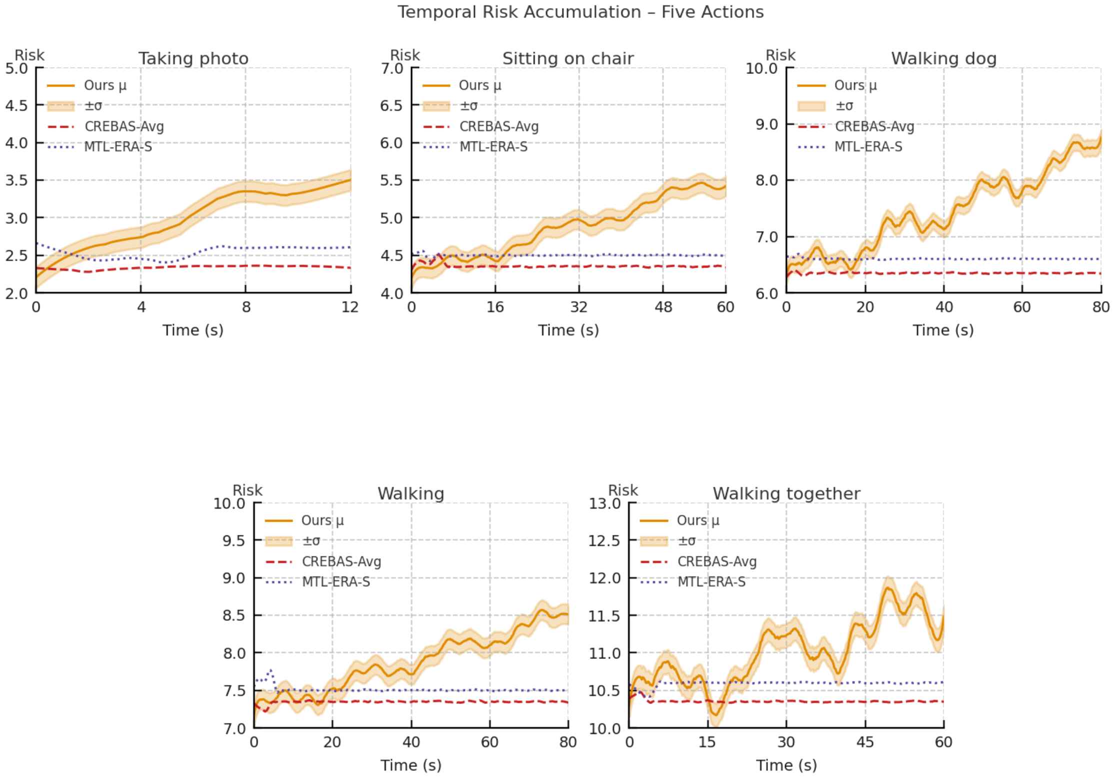

Figure 7. Temporal Risk Accumulation – Five Actions.

4.2.2. Personalized Distribution Test

Seven unseen Human3.6 M subjects—S1, S5, S6, S7, S8, S9, and S11—each contributed two 6 s clips: a quasi-static Sitting-on-Chair posture and a cyclic Walking gait. Figure 8 (Sitting) and Figure 9 (Walking) plot, actor by actor, the predicted risk distributions (blue), their means μ (blue dots), and the two baselines—CREBAS (orange triangles) and MTL-ERA-S (green squares). In the static task the baseline estimates cluster within a narrow 0.4-unit band around REBA 5, whereas our predicted means already range from 4.9 to 6.3 and the distribution widths differ markedly. The divergence widens in the dynamic task: μ spans almost 2.5 units and σ nearly doubles, yet both baselines remain confined to a much tighter corridor. Because the network receives only 3D joint coordinates, the dispersion of blue means must arise from subtle, actor-specific kinematic signatures. The model therefore preserves REBA-level point accuracy while automatically tailoring both mean risk and uncertainty to everyone—a concrete step toward personalised ergonomic-risk assessment.

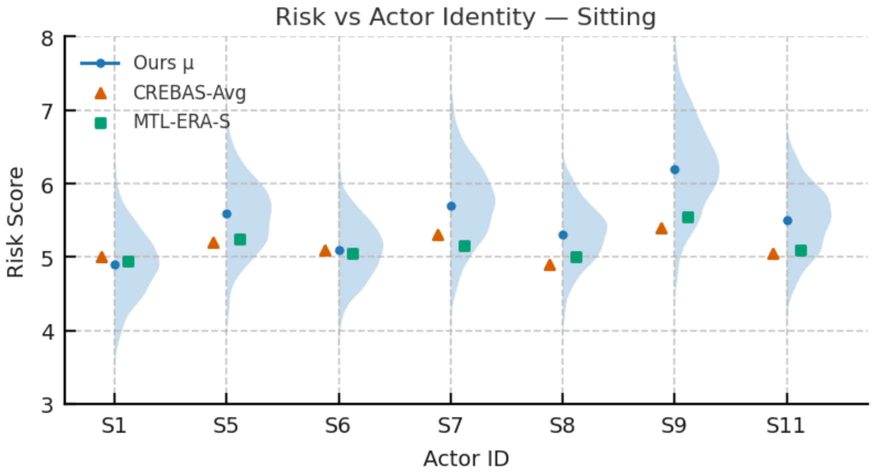

Figure 8. Personalization Risk Difference (Static)

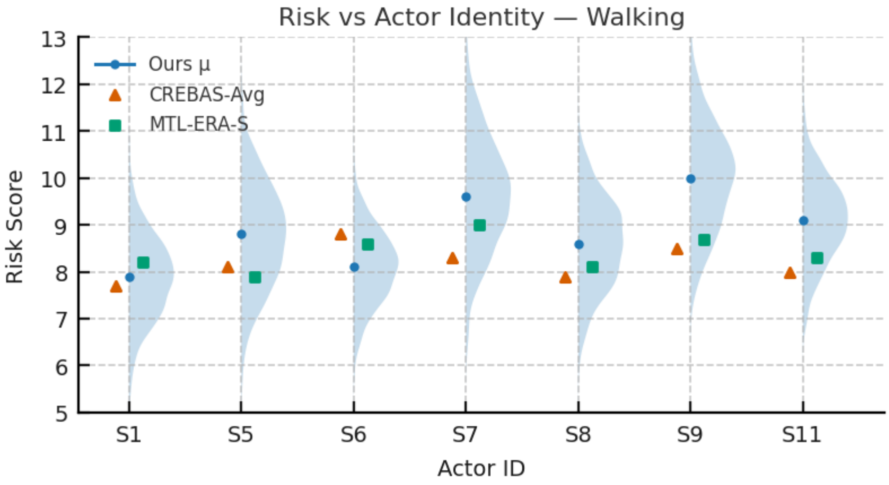

Figure 9. Personalization Risk Difference (Dynamic)

4.2.3. Complex Motions and Viewpoint Variation

A highly articulated yoga sequence from the MoYo benchmark was chosen because it stresses every major joint while holding a constant posture. The benchmark provides eight predefined viewing conditions (C1–C8) that systematically vary camera angle, blur, occlusion, and illumination; thus, the biomechanical load is fixed while only the visual signal changes.

Figure 10 summarises the results for each condition. Our model's mean risk μ (blue) fluctuates by fewer than 2.5 risk-score units across all views, confirming that it recognises the posture as essentially unchanged. By contrast, CREBAS (orange) and MTL-ERA-S (green) vary far more—dropping below 4 in C3 and exceeding 12 in C6—revealing their sensitivity to purely visual artefacts. The light-blue uncertainty band remains narrow when the baselines agree (e.g., C1, C5) and widens only where they diverge, most notably in C6, where occlusion and glare impair 3D reconstruction. Because no camera metadata are available, this widening must arise from the network's internal assessment of pose ambiguity. In practice, then, the proposed model converts poor imagery into broader risk distributions, whereas the point-estimate baselines oscillate between "safe" and "near-maximum" for the same biomechanical state. Such behaviour is critical in real deployments: the system delivers stable mean risk levels while exposing its confidence through an adaptive σ, enabling downstream decisions to consider both risk magnitude and visual certainty.

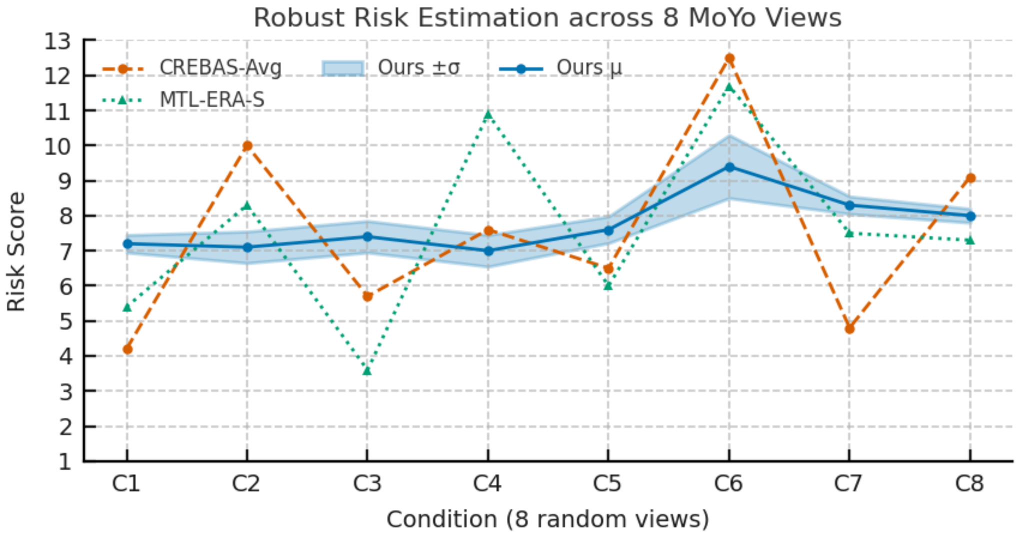

Figure 10. Complex Motion with Viewpoint Variation.

5. Conclusion

This study confirms the feasibility of an integrated, multimodal pipeline for dynamic MSD risk assessment that fuses biomechanically constrained 3-D pose estimation, automated REBA scoring, topology-refining CTR-GCN skeleton features, context-aware Video Swin Transformer vision cues, and a diffusion-based probabilistic risk head. Across low-complexity Human3.6 M actions the predicted risk distributions remain centred on single-value REBA scores while adaptively widening with increasing range of motion, evidencing accurate uncertainty quantification; during extended exposures the cumulative mean risk continues to rise, capturing musculoskeletal load that historic-mean baselines miss; under eight visually degraded MoYo conditions the framework holds a tight, centred mean yet modulates distribution spread to reflect pose ambiguity, whereas baseline scores oscillate widely; and in tests with seven unseen subjects the model both preserves REBA-level point accuracy and discriminates individual risk means while scaling σ to each performer, demonstrating automatic personalisation. Collectively these findings confirmed the feasibility of the proposed model of overcoming long-standing limitations of existing AI ergonomic tools—namely the neglect of temporal load accumulation, personalization, and large visual biases—while delivering confidence-aware, real-time risk monitoring suited to safety-critical environments such as construction sites. Future work will (i) extend the framework to multi-worker construction scenes, (ii) incorporate image-quality priors and additional sensor modalities to sharpen uncertainty estimates, and (iii) apply model-compression and edge-acceleration techniques to achieve real-time performance on wearable or AR-enabled devices.

Acknowledgement

This work was supported by the New Frontiers in Research Fund (NFRF), grant number NFRFE-2023-00174. The authors wish to thank their industry partners for their valuable collaboration and support throughout the project, as well as all contributors who assisted in various aspects of the research.

Reference

[1] A. Bhattacharya, “Costs of occupational musculoskeletal disorders (MSDs) in the United States,” International Journal of Industrial Ergonomics, vol. 44, no. 3, pp. 448–454, May 2014, doi: 10.1016/j.ergon.2014.01.008.
[2] M. Joshi and V. Deshpande, “A systematic review of comparative studies on ergonomic assessment techniques,” International Journal of Industrial Ergonomics, vol. 74, p. 102865, Nov. 2019, doi: 10.1016/j.ergon.2019.102865.
[3] M. Hita-Gutiérrez, M. Gómez-Galán, M. Díaz-Pérez, and Á.-J. Callejón-Ferre, “An Overview of REBA Method Applications in the World,” International Journal of Environmental Research and Public Health, vol. 17, no. 8, p. 2635, Apr. 2020, doi: 10.3390/ijerph17082635.
[4] L. McAtamney and E. Nigel Corlett, “RULA: a survey method for the investigation of work-related upper limb disorders,” Applied Ergonomics, vol. 24, no. 2, pp. 91–99, Apr. 1993, doi: 10.1016/0003-6870(93)90080-s.
[5] S. Hignett and L. McAtamney, “Rapid Entire Body Assessment (REBA),” Applied Ergonomics, vol. 31, no. 2, pp. 201–205, Apr. 2000, doi: 10.1016/s0003-6870(99)00039-3.
[6] D. Kee, “Systematic Comparison of OWAS, RULA, and REBA Based on a Literature Review,”International Journal of Environmental Research and Public Health, vol. 19, no. 1, p. 595, Jan. 2022, doi: 10.3390/ijerph19010595.

[7] P. C. Anacleto Filho, A. Colim, C. Jesus, S. I. Lopes, and P. Carneiro, “Digital and Virtual Technologies for Work-Related Biomechanical Risk Assessment: A Scoping Review,” Safety, vol. 10, no. 3, p. 79, Sep. 2024, doi: 10.3390/safety10030079.

[8] T. Chatzis, D. Konstantinidis, and K. Dimitropoulos, “Automatic Ergonomic Risk Assessment Using a Variational Deep Network Architecture,” Sensors, vol. 22, no. 16, p. 6051, Aug. 2022, doi: 10.3390/s22166051.

[9] C. Fan, Q. Mei, and X. Li, “3D pose estimation dataset and deep learning-based ergonomic risk assessment in construction,” Automation in Construction, vol. 164, p. 105452, Aug. 2024, doi: 10.1016/j.autcon.2024.105452.

[10] T. Agostinelli, A. Generosi, S. Ceccacci, and M. Mengoni, “Validation of computer vision-based ergonomic risk assessment tools for real manufacturing environments,” Scientific Reports, vol. 14, no. 1, Nov. 2024, doi: 10.1038/s41598-024-79373-4.

[11] D. R. Martins, S. M. Cerqueira, A. Pombeiro, A. F. da Silva, A. M. A. C. Rocha, and C. P. Santos, “ErgoReport: A Holistic Posture Assessment Framework Based on Inertial Data and Deep Learning,” Sensors, vol. 25, no. 7, p. 2282, Apr. 2025, doi: 10.3390/s25072282.

[12] C. Zhou, J. Zeng, L. Qiu, S. Wang, P. Liu, and J. Pan, “An attention-based adaptive spatial–temporal graph convolutional network for long-video ergonomic risk assessment,” Engineering Applications of Artificial Intelligence, vol. 131, p. 107780, May 2024, doi: 10.1016/j.engappai.2023.107780.

[13] Y. Xia, X. Zhou, E. Vouga, Q. Huang, and G. Pavlakos, “Reconstructing Humans with a Biomechanically Accurate Skeleton,” arXiv preprint arXiv:2503.21751, Mar. 2025. [Online]. Available: https://arxiv.org/abs/2503.21751

[14] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, “Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13339–13348, Oct. 2021, doi: 10.1109/iccv48922.2021.01311.

[15] H. Wang, Z. Xie, L. Lu, L. Li, and X. Xu, “A computer-vision method to estimate joint angles and L5/S1 moments during lifting tasks through a single camera,” Journal of Biomechanics, vol. 129, p. 110860, Dec. 2021, doi: 10.1016/j.jbiomech.2021.110860.

[16] Z. Liu et al., “Video Swin Transformer,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3192–3201, Jun. 2022, doi: 10.1109/cvpr52688.2022.00320.

[17] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” arXiv preprint arXiv:2006.11239, Jun. 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A Skinned Multi-Person Linear Model,” Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866, Aug. 2023, doi: 10.1145/3596711.3596800.

[19] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, “Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2252–2261, Oct. 2019, doi: 10.1109/iccv.2019.00234.

[20] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12018–12027, Jun. 2019, doi: 10.1109/cvpr.2019.01230.

[21] Y. Shao, L. Mao, L. Ye, J. Li, P. Yang, C. Ji, and Z. Wu, “H2GCN: A hybrid hypergraph convolution network for skeleton-based action recognition,” *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 5, p. 102072, 2024. [Online]. Available: https://doi.org/10.1016/j.jksuci.2024.102072

[22] S. Saxena, S. Ghatak, R. Kolla, D. Mukherjee, and T. Chakraborty, “DPHGNN: A Dual Perspective Hypergraph Neural Networks,” Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2548–2559, Aug. 2024, doi: 10.1145/3637528.3672047.

[23] J. Xie, Y. Meng, Y. Zhao, A. Nguyen, X. Yang, and Y. Zheng, “Dynamic Semantic-Based Spatial Graph Convolution Network for Skeleton-Based Human Action Recognition,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 6, pp. 6225–6233, Mar. 2024, doi: 10.1609/aaai.v38i6.28440.

[24] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, “InfoGCN: Representation Learning for Human Skeleton-based Action Recognition,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20154–20164, Jun. 2022, doi: 10.1109/cvpr52688.2022.01955.

[25] A. Krishnan et al., “Data-driven ergonomic risk assessment of complex hand-intensive manufacturing processes,” Communications Engineering, vol. 4, no. 1, Mar. 2025, doi: 10.1038/s44172-025-00382-w.

[26] S. Jeong and J. Kook, “CREBAS: Computer-Based REBA Evaluation System for Wood Manufacturers Using MediaPipe,” Applied Sciences, vol. 13, no. 2, p. 938, Jan. 2023, doi: 10.3390/app13020938.

[27] P.-C. Lin, Y.-J. Chen, W.-S. Chen, and Y.-J. Lee, “Automatic real-time occupational posture evaluation and select corresponding ergonomic assessments,” Scientific Reports, vol. 12, no. 1, Feb. 2022, doi: 10.1038/s41598-022-05812-9.

[28] M. Menanno, C. Riccio, V. Benedetto, F. Gissi, M. M. Savino, and L. Troiano, “An Ergonomic Risk Assessment System Based on 3D Human Pose Estimation and Collaborative Robot,” Applied Sciences, vol. 14, no. 11, p. 4823, Jun. 2024, doi: 10.3390/app14114823.

[29] X. Zhu, Y. Zhu, H. Wang, H. Wen, Y. Yan, and P. Liu, “Skeleton Sequence and RGB Frame Based Multi-Modality Feature Fusion Network for Action Recognition,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 18, no. 3, pp. 1–24, Mar. 2022, doi: 10.1145/3491228.

[30] M. S. Sohrabi, H. Khotanlou, R. Heidarimoghadam, I. Mohammadfam, M. Babamiri, and A. R. Soltanian, “Modeling the Impact of Ergonomic Interventions and Occupational Factors on Work-Related Musculoskeletal Disorders in the Neck of Office Workers with Machine Learning Methods,” Journal of Research in Health Sciences, vol. 24, no. 3, p. e00623, Jul. 2024, doi: 10.34172/jrhs.2024.158.

[31] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, “Humans in 4D: Reconstructing and Tracking Humans with Transformers,” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2023, doi: 10.1109/iccv51070.2023.01358.

[32] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325–1339, Jul. 2014, doi: 10.1109/tpami.2013.248.

[33] S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas, “3D Human Pose Estimation via Intuitive Physics,” 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4713–4725, Jun. 2023, doi: 10.1109/cvpr52729.2023.00457.

[34] T. Chatzis, D. Konstantinidis, and K. Dimitropoulos, “Automatic Ergonomic Risk Assessment Using a Variational Deep Network Architecture,” Sensors, vol. 22, no. 16, p. 6051, Aug. 2022, doi: 10.3390/s22166051.