AI-Assisted Work-Related Musculoskeletal Disorder Risk Identification & Intervention Recommendation

by

Leyang Wen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Civil Engineering)
in the University of Michigan
2026

Doctoral Committee:

Professor SangHyun Lee, Chair
Professor Vineet R. Kamat
Professor Carol C. Menassa
Assistant Professor Oshin Tyagi

Leyang Wen

wenleyan@umich.edu

ORCID iD: 0000-0001-7769-5582

© Leyang Wen 2026

Dedication

To my wife Xinyan

Acknowledgements

Firstly, I would like to express my deepest gratitude to my advisor, Prof. SangHyun Lee, for his continuous support and insightful guidance throughout my PhD journey. Prof. Lee's commitment to excellence, dedication to his students, and vision for applying technology to benefit the people in the construction industry have profoundly influenced my growth both as a researcher and as an individual. I feel incredibly fortunate to have him as my mentor.

I would also like to sincerely thank the members of my dissertation committee, Prof. Vineet R. Kamat, Prof. Carol C. Menassa, and Prof. Oshin Tyagi, for their guidance, constructive feedback, and encouragement throughout this process. Their expertise and perspectives have enriched my research and helped me grow as a researcher.

I am deeply grateful to the scholars who have directly shaped my research: Dr. Meiyin Liu, Dr. Daeho Kim, Dr. Jinwoo Kim, Dr. Francis Baek, and Dr. Dongmin Lee. I also extend special thanks to my undergraduate advisor, Dr. JoonOh Seo, who first introduced me to research and set me on this path.

I am thankful for the friends and colleagues I met through this journey, who made the experience not only productive but truly enjoyable: Dr. Juhyeon Bae, Dr. Gaang Lee, Dr. Hoonyong Lee, Dr. Kwonsik Song, Dr. Xiaowei Wang, Jiu Sohn, Gunwoo Yong, Minseo Park, Seongju Jang, Jinsik Park, Haytham Tang, Bongil Gu, and Sze Yan How.

I also wish to thank our industry partners at VelocityEHS, Dr. Julia Penfield, Dr. Veeru Talreja, Blake McGowan, and Rick Barker, for their close collaboration and continued support

throughout this research. Seeing our work applied in the field and helping real workers has been one of the most rewarding aspects of my PhD.

Finally, I would like to thank my family, Xinyan, Zeqing, Shiqiang, and Football, for their unconditional love, patience, and support. Their belief in me, even during the most challenging times, gave me the strength to persevere. This dissertation would not have been possible without them.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables ix

List of Figures xi

Abstract xiii

Chapter 1 Introduction 1

1.1 Background 1

1.2 AI-assisted WMSD Risk Identification via Monocular Motion Capture 3

1.3 AI-assisted WMSD Intervention Recommendation via Physics-based Simulation 5

1.4 Knowledge Gaps 6

1.5 Research Objectives and Approaches 8

1.6 The Structure of the Dissertation 11

Chapter 2 Ergonomics-Oriented 3D Human Motion Dataset 13

2.1 Introduction 13

2.2 Literature Review on Vision-based WMSD Risk Assessments and Datasets 14

2.3 3D Human Pose Keypoints and Joint Angle Calculations for Vision-based Pose Estimation 19

2.3.1 66 3D body keypoint positions 19

2.3.2 Joint angle calculations using the 66-keypoint 3D pose 22

2.4 3D Human Pose Dataset Collection 26

2.4.1 Motion capture equipment setup 27

2.4.2 Participant characteristics 28

2.4.3 MMH and assembly task motions 29
2.4.4 Data post-processing 32
2.4.5 Characteristics of the collected 3D human pose dataset 33
2.5 Evaluation of the Dataset's Effectiveness in Training a 3D Pose Estimation Model for Joint Angle Calculations 35
2.5.1 3D pose evaluation result 36
2.5.2 Joint angle evaluation result 37
2.6 Discussion 40
2.7 Conclusion 41
Chapter 3 Vision-Based WMSD Risk Assessment Computational Pipeline 43
3.1 Introduction 43
3.2 Stage 1: Image-to-2D Model Training Adapted to Real-World Conditions 46
3.2.1 Transfer Learning and Backbone Freezing 46
3.2.2 Merge Training with Open-Source 2D Human Pose Datasets 47
3.3 Stage 2: 2D-to-3D Model Training Adapted to Real-World Conditions 48
3.3.1 Camera Pitch Correction 49
3.3.2 Angle Loss Function Design 49
3.4 Evaluation of the End-To-End Computational Pipeline 51
3.4.1 Quantitative Evaluation of 2D Pose Estimation on Test Set Videos 51
3.4.2 Quantitative Evaluation of 3D Pose Estimation on Test Set Videos 53
3.4.3 Quantitative Evaluation of Joint Angle Estimation on Test Set Videos 54
3.4.4 Qualitative Evaluation on Real-World Videos 58
3.5 Discussion 61
3.6 Conclusion 62
Chapter 4 Intervention Recommendation Applying Scene-Aware Motion Synthesis within Physics-Based Simulations 63

4.1 Introduction 63

4.2 Literature Review On Current Scene-Aware Motion Synthesis Methods 65

4.3 Synthesizing Risk-reduced Motions with Custom Ergonomics Rewards 66

4.3.1 Learning from Published Ergonomics Guidelines 68

4.3.2 Learning from Reference Motions via Imitation Learning 70

4.3.3 Learning from Expert Interactions for Task-Specific Tacit Knowledge 72

4.4 Computational Framework: Generating Task and Site-specific Ergonomics Intervention Applying Motion Synthesis 73

4.5 Evaluation of the Proposed Intervention Recommendation Framework Under Various Task and Site Constraints 76

4.5.1 Verification of the Computational Framework 76

4.5.2 Test Subjects and Data Collection 77

4.5.3 Evaluation metric 78

4.5.4 Evaluation of Synthesized Motions on Unconstrained MMH Task Scenarios 79

4.5.5 Evaluation of Synthesized Motions on Site-constrained MMH Task Scenarios 85

4.6 Discussion 90

4.7 Conclusion 93

Chapter 5 Quantitative Evaluation Of Human Pose Representations for WMSD Risk Assessments 94

5.1 Introduction 94

5.2 Quantitative Comparison between Keypoint-Based and Mesh-Based Pose Estimation Models 95

5.2.1 Mesh to Body Landmark Mapping 96

5.2.2 Mesh Dataset Preparation 98

5.2.3 Model Architecture and Training 98

5.3 Evaluation of Pose Representations on Joint Angle Accuracy 99

5.4 Discussion 103

5.5 Conclusion 104
Chapter 6 Conclusions and Recommendations..... 106
6.1 Summary of research 106
6.2 Contributions 109
6.3 Future research..... 110
Bibliography 112

List of Tables

Table 2.1. Overview of human pose datasets that provide 3D pose, 2D pose, and 2D images. .. 18

Table 2.2. List of calculated joint angles and their application in WMSD risk assessments..... 19

Table 2.3. 47 body surface keypoint acronyms and full names 21

Table 2.4. Height, weight, and BMI of participants sorted by gender 29

Table 2.5. List of available task parameters in each task category 32

Table 2.6. Number of captured train and test frames broken down by task..... 35

Table 2.7. Joint angle estimation errors (in degrees) 37

Table 2.8. Number of distinct frames captured in the VEHS-6.7M dataset 41

Table 3.1. Mean error (px) and PCK of 2D pose estimation (keypoint abbreviations in Table 2.3). 52

Table 3.2. MPJPE and PA-MPJPE of 3D pose estimation (keypoint abbreviations in Table 2.3). 54

Table 3.3. Joint angle estimation errors (in degrees) 55

Table 4.1 Height, weight, and BMI of test subjects. 78

Table 4.2. NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in unconstrained task scenarios. 81

Table 4.3. 3D SSPP L4/L5 compression force and joint strength percent capable scores of synthesized and captured motions in unconstrained task scenarios..... 84

Table 4.4. NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in site-constrained task scenarios. 87

Table 4.5. 3D SSPP L4/L5 compression force and joint strength percent capable scores of synthesized and captured motions in constrained task scenarios..... 89

Table 5.1. Mapping between 47 body surface keypoints and SMPL vertex indices. 97

Table 5.2. Joint angle MAE comparison between keypoint-based and mesh-based models..... 100

List of Figures

Figure 1.1. Overall research framework. 11

Figure 2.1. Illustration of 3D human pose representation formats: (a) joint center representation from Human3.6M (Ionescu et al. 2014), visualized by Kudo et al. (2018); (b) mesh-based representation in 3DPW (Von Marcard et al. 2018); (c) our 66 body keypoints..... 17

Figure 2.2. Visualization of a) 47 body surface keypoint positions (acronym explanation listed in Table 2.3); and b) 19 calculated joint center keypoint positions. 22

Figure 2.3. Diagrams illustrating segment coordinate system definitions for joint angle calculations. 23

Figure 2.4. Overview of the 3D human pose dataset collection process. 27

Figure 2.5. a) Vicon motion capture and RGB video capture setup; b) example camera views. 28

Figure 2.6. Images and descriptions of the selected task categories. 30

Figure 2.7. Visualization of: a) raw motion capture marker trajectories; b) auto-labeled markers; c) gap-filling process of the occluded LASIS marker using temporal and spatial information of surrounding markers. 33

Figure 2.8. VEHS-6.7M dataset includes a) 2D RGB videos; b) 2D human pose annotations; c) 3D human pose annotations; and d) SMPL annotations. 34

Figure 2.9. Histograms of angle estimation errors with GT2D input, overlaid normal distribution curves computed from the corresponding means and standard deviations. 39

Figure 3.1. Computational pipeline for joint angle estimation from monocular video. 45

Figure 3.2. Overlapping keypoints between the three open-source 2D human pose datasets and the ergonomics-oriented keypoint set used in merge training. 48

Figure 3.3. Histograms of angle estimation errors with RGB video input, overlaid normal distribution curves computed from the corresponding means and standard deviations. 57

Figure 3.4. Example frames from industry videos with corresponding estimated 3D human poses (part1). 59

Figure 3.5. Example frames from industry videos with corresponding estimated 3D human poses (part 2). 60

Figure 4.1. Reference motion capture pipeline for imitation learning. 71

Figure 4.2. Tacit knowledge capture pipeline through experts' natural language feedback. 73

Figure 4.3. Computational framework for motion synthesis–based ergonomics interventions... 74

Figure 4.4. Task parameters for the 12 unconstrained MMH task scenarios across four task categories. 79

Figure 4.5. Visual comparison of captured task motions and synthesized motions across the four MMH task categories. 80

Figure 4.6. Box plots comparing the NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in unconstrained task scenarios. 82

Figure 4.7. Box plots comparing the 3D SSPP L4/L5 compression force and joint strength percent capable scores of captured and synthesized motions in unconstrained task scenarios. ... 85

Figure 4.8. Visual comparison of captured task motions and synthesized motions across the four MMH task categories under site constraints. 86

Figure 4.9. Box plots comparing the NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in constrained task scenarios. 88

Figure 4.10. Box plots comparing the 3D SSPP L4/L5 compression force and joint strength percent capable scores of captured and synthesized motions in constrained task scenarios. 90

Figure 5.1. Model architecture for keypoint-based and mesh-based 3D pose estimation models. 96

Figure 5.2. Histograms of angle estimation errors with RGB video input, for both the keypoint-based and mesh-based models. 101

Abstract

Work-related musculoskeletal disorders (WMSDs) are a major cause of nonfatal injuries in construction, impacting workers' health and productivity. Mitigating WMSDs typically requires substantial resources to identify risks and design corresponding interventions, a process that is impractical for the construction industry due to its diverse tasks, dynamic site conditions, and high labor turnover. As a result, construction workers exposed to WMSD risks are often left without the means or resources to assess their risk levels or design appropriate interventions to mitigate those risks. Monocular motion capture, which estimates human motion from ordinary videos, has been increasingly applied as a less intrusive and accessible method for WMSD risk assessment. However, the underlying machine learning models are built from generic pose datasets, lacking critical 3D information for calculating high-degree-of-freedom (DoF) joint angles needed in WMSD risk assessment. Beyond risk identification, intervention design remains a largely manual process that demands substantial resources, with limited research on accessible alternatives suitable for the unique characteristics of the construction industry.

To address these gaps, this dissertation proposes vision-based WMSD risk identification and simulation-driven intervention recommendation in construction, aiming to (1) identify WMSD risks from ordinary videos by estimating high-DoF joint angles needed for WMSD risk assessments and (2) recommend task- and site-specific interventions to provide construction site personnel with accessible and resource-efficient ergonomics guidance. To achieve the first aim, a large-scale 3D human motion dataset containing 6.7 million frames was collected, and a

computational pipeline was developed to estimate high-DoF joint angles from monocular video. To achieve the second aim, a simulation-driven intervention recommendation framework with a custom ergonomics reward was developed. Evaluation of the vision-based risk identification pipeline on the collected test set demonstrated a mean absolute error of 5.9° in joint angle estimation. Evaluation of the intervention recommendation framework across 12 unconstrained and 4 site-constrained manual material handling task scenarios demonstrated an average 39% reduction in NIOSH Lifting Index and 11% reduction in L4/L5 back compression force compared to captured task motions. The findings of this dissertation provide foundational steps toward accessible WMSD risk identification and intervention recommendations in construction, using devices such as smartphones.

Chapter 1 Introduction

1.1 Background

Work-related musculoskeletal disorders (WMSDs) are the leading cause of non-fatal injuries in the U.S., accounting for 29.6% of all injury cases requiring days away from work, job restriction, or transfer (DART) between 2021 and 2022 (Bureau of Labor Statistics 2023). During these two years, WMSDs led to 521,350 incidents necessitating days away from work, with a median absence duration of 14 days (Bureau of Labor Statistics 2023), and an estimated $34.65 billion in workers' compensation costs (Liberty Mutual 2021, 2022).

WMSDs are particularly prevalent in the construction industry because construction work frequently exposes workers to heavy lifting, forceful exertions, repetitive motions, and awkward postures such as bending, twisting, kneeling, and overhead reaching (Seo and Lee 2021). Among the 923,980 total nonfatal injury cases reported in the U.S. construction industry from 2011 to 2022, 26.8% were attributed to overexertion and bodily reaction (CPWR 2025). The incidence rate of such injuries for construction workers was 39.3 cases per 10,000 full-time equivalent workers, approximately 40% higher than the average across all occupations (BLS 2022). Beyond their direct impact on workers' health and associated healthcare and compensation costs, WMSDs also impose high indirect costs on construction projects. Affected workers often experience reduced work capacity and productivity, as well as elevated absenteeism rates, contributing to project delays and increased labor costs (Huang et al. 2020; Kee et al. 2011; Seo and Lee 2021; Vijayakumar and Choi 2022).

One common way to manage WMSD risks is to identify them through observation-based methods, such as the Rapid Entire Body Assessment (REBA), Rapid Upper Limb Assessment (RULA), Ovako Working Posture Assessment System (OWAS), and the National Institute for Occupational Safety and Health (NIOSH) Lifting Equation (Hignett and McAtamney 2000; McAtamney and Corlett 1993; Karhu et al. 1977; Waters et al. 2021). Such observation-based WMSD risk assessment methods require trained ergonomists to systematically record and evaluate workers' postures and movements during task execution, typically through direct on-site observation or video recording. Key posture characteristics, such as knee flexion angle in REBA, upper arm flexion in RULA, trunk posture category in OWAS, or torso asymmetry angle in the NIOSH Lifting Equation, are captured and evaluated against reference thresholds to derive body-part risk levels. These body-part risk levels are then combined through method-specific scoring procedures to produce an overall risk score indicating the urgency of ergonomics intervention, such as whether immediate action is required.

Based on the assessment results, the identified WMSD risks can be mitigated through corresponding interventions. Interventions can include helping workers improve their techniques by demonstrating safer ways to perform their tasks, such as maintaining a proper lifting posture during heavy lifting or keeping the wrist neutral during tool use. In cases where the task demands or workstation configuration are the root cause of the identified risk, they can also recommend task and workstation modifications, such as substituting lighter-weight materials, adjusting workbench height, or introducing mechanical aids to reduce biomechanical demand.

However, both identifying WMSD risks and designing interventions remain resource-intensive processes that demand substantial domain expertise, time, and cost (Jeong & Kook, 2023; Li et al., 2019; Plantard et al., 2017; Ryu et al., 2021). This challenge is further exacerbated in

construction due to the diverse tasks (Choi 2012) and ever-changing site conditions (Choi et al. 2016). Moreover, the construction industry experiences significantly higher labor turnover than other WMSD-prone industries such as manufacturing, with annual separation rates approximately 44% higher (3.9% vs. 2.7%) and hire rates approximately 62% higher (4.2% vs. 2.6%) in 2024 (BLS 2025a; b), meaning that identified risks and designed interventions can quickly become outdated as the task, site, or workforce changes. As a result, construction workers exposed to WMSD risks are often left without the means or resources to assess their risk levels and design appropriate interventions to mitigate those risks. There is a need for accessible and resource-efficient methods that can both identify WMSD risks and recommend interventions. Sections 1.2 and 1.3 discuss the potential of monocular motion capture and physics-based simulation in addressing this need and identify the limitations that motivate this dissertation.

1.2 AI-assisted WMSD Risk Identification via Monocular Motion Capture

Accurately capturing and quantifying workers' body motion is fundamental to evaluating their exposure to WMSD risks. Recent advancements in monocular motion capture, the technique of estimating human body motion from a single-camera video, have enabled more accessible methods for WMSD risk identification. Such methods have been increasingly adopted in both academic research (Chu et al. 2020; Fan et al. 2024; Jeong and Kook 2023; Kwon et al. 2022; Paudel et al. 2022) and industry applications (ErgoPlus 2024; Sundareshan 2024; VelocityEHS 2024a). Unlike traditional motion capture approaches, such as marker-based systems that require workers to wear reflective markers and be surrounded by calibrated infrared cameras (Vicon 2023), inertial measurement unit (IMU)-based systems that require intrusive body-worn sensors (Sers et al. 2020), or multi-camera markerless systems that require extensive calibrated camera setup

(Salehi et al. 2026; Varcin and Boocock 2026), monocular motion capture operates from a regular single-camera video that can be captured by commonly available devices, such as a smartphone.

These methods utilize deep learning-based human pose estimation models to detect and localize 2D keypoints on the human body from each 2D video frame, such as the shoulder, elbow, wrist, hip, knee, and ankle joint centers, and lift these 2D detections into a 3D human pose. From estimated body keypoints in the 3D pose, joint angles at relevant body segments are calculated, such as back and knee flexion, which serve as essential inputs for subsequent WMSD risk assessment methods such as REBA, RULA, OWAS, and the NIOSH Lifting Equation.

Despite its strong potential in enabling accessible and less intrusive WMSD risk identification for workers and managers on site, current monocular motion capture methods remain limited in their ability to support WMSD risk assessment. Most existing methods use deep learning models trained on datasets developed for general computer vision tasks, where human pose is represented using a sparse set of 3D joint centers (Bonakdar et al. 2025; Chen et al. 2024; Fan et al. 2024). While such representations are sufficient for estimating angles in simple hinge-like joints such as the elbow and knee, they lack the 3D information needed to resolve individual rotation components in high-DoF joints, including the neck, shoulder, and wrist, which exhibit multiple rotational components such as wrist flexion/extension, ulnar/radial deviation, and pronation/supination. As a result, 3D human pose representations estimated by models trained on joint-center-based datasets lack the 3D information required to compute high-DoF joint angles used in WMSD risk assessment, leading to incomplete or oversimplified joint angle calculations that limit the applicability of monocular motion capture for WMSD risk evaluation.

Beyond the limitation of existing human pose representations in high-DoF joint angle calculation, real-world deployment conditions introduce additional sources of estimation error.

Occlusion, varying camera viewpoints, inconsistent lighting, and cluttered backgrounds are common in field settings but largely absent from the controlled laboratory environments on which existing 3D human pose estimation models are predominantly trained and validated. The robustness of these models under such conditions remains underexplored, and computational pipelines that account for these real-world factors in their design and evaluation remain limited.

1.3 AI-assisted WMSD Intervention Recommendation via Physics-based Simulation

Unlike risk identification, which has increasingly been made accessible to construction site personnel through monocular motion capture, the subsequent intervention design step has received comparatively limited research attention or industry adoption. While recent work has explored generative AI to identify root causes and produce text-based intervention suggestions from site images and videos (Yong et al. 2024, 2026), these approaches remain constrained in their ability to account for task- and site-specific conditions, and fall short of the actionable, context-sensitive guidance that typical interventions would provide, such as demonstrating correct techniques for specific tasks or redesigning tasks and workstations to reduce biomechanical demand. As a result, ergonomics intervention designs remain a resource-intensive process that demands substantial domain expertise, time, and cost, making them challenging for the construction industry to adopt in practice, given its unique characteristics of diverse tasks (Choi 2012), dynamic site conditions (Choi et al. 2016), and high labor turnover (BLS 2025a; b).

Motion synthesis methods within physics-based simulations, which can synthesize and visualize physically plausible human motions under specified task and site constraints, offer a promising path toward enabling accessible and resource-efficient intervention designs (Pan et al. 2025; Peng et al. 2021, 2022). The synthesized human motions can serve as visual demonstrations of correct techniques for performing workers' tasks, providing construction workers with task- and

site-specific posture guidance. Additionally, such synthesized motions can enable automated "what-if" analysis of different task or workstation modifications, allowing site personnel, such as workers and managers, to evaluate and adjust their tasks and workstations toward safer configurations.

However, for motion synthesis to serve as a viable intervention recommendation tool, the synthesized motions must be safe to perform and reflect the real-world task and site conditions under which WMSD risk was identified. Current motion synthesis methods are primarily designed for general applications such as animation and game character control, with limited consideration of WMSD risks (Pan et al. 2025; Peng et al. 2021, 2022). As a result, there is no guarantee that synthesized motions reflect proper task techniques or reduced biomechanical demand, making them unsuitable for intervention applications such as posture demonstrations or simulation-based task and workstation redesign. Moreover, generating task- and site-specific interventions using motion synthesis requires users to recreate real-world task and site conditions within the physics-based simulation environment. This process demands both technical expertise in simulation setup and substantial manual effort to accurately capture and translate site geometry, object properties, and task parameters into the simulation. Without a supporting framework that automates or simplifies the capture of real-world task and site information, this process remains impractical for construction site personnel and limits the scalability of simulation-driven intervention recommendations beyond controlled research settings.

1.4 Knowledge Gaps

To fully leverage the potential of AI-assisted systems for vision-based WMSD risk identification and simulation-driven intervention recommendation, several knowledge gaps need

to be addressed. Building on the limitations of existing methods identified in Sections 1.2 and 1.3, this section summarizes the key gaps that motivate this dissertation.

First, existing monocular motion capture methods lack the capability to estimate high-DoF joint angles required for comprehensive WMSD risk assessment. This limitation stems from the predominant use of deep learning models trained on generic 3D human pose datasets, where human pose is represented using sparse sets of 3D joint centers. Such representations are insufficient for resolving the individual rotation components of high-DoF joints such as the shoulder, neck, and wrist. There is therefore a lack of an ergonomics-oriented 3D human motion dataset that captures richer keypoint representations beyond joint centers to support high-DoF joint angle estimation for WMSD risk assessment.

Second, computational pipelines for monocular motion capture have not been specifically adapted for the high-DoF angle calculation needed in WMSD risk assessment, nor for the visual challenges common in field environments such as occlusion, varying camera viewpoints, inconsistent lighting, and cluttered backgrounds. The development and evaluation of such adaptations for enabling practical vision-based WMSD risk assessment in real-world construction settings remains limited.

Third, despite its potential, existing motion synthesis methods lack the capability to reduce WMSD risks in synthesized motions, as they were not designed with ergonomic safety as an objective. Without dedicated mechanisms to constrain synthesized motions toward safer biomechanical outcomes, these methods are unsuitable for ergonomics intervention applications in construction.

Fourth, there is no computational framework that integrates motion synthesis methods for task- and site-specific ergonomics intervention recommendation grounded in identified WMSD

risks. Such a framework should allow construction site personnel to capture real-world task and site conditions and translate them into the physics-based simulation environment using accessible devices such as smartphones, without requiring substantial technical expertise. Without such a framework, simulation-driven ergonomics intervention remains largely confined to controlled research settings.

Beyond the knowledge gaps identified above, an additional question arises from the proposed keypoint-based pose representation for WMSD risk assessment. While the keypoint-based representation proposed in this dissertation supports accurate high-DoF joint angle calculations, it is bounded by the initial keypoint set defined during data collection and model training. Mesh-based pose representations have emerged as a versatile alternative that can recover the full body surface geometry, enabling it to adapt to different downstream tasks, such as joint angle calculation under different anatomical definitions. However, the impact of mesh-based representations on joint angle estimation accuracy compared to keypoint-based representations has not been systematically investigated, and this comparison is critical for WMSD risk assessment, where joint angle accuracy directly determines the accuracy of risk scores.

1.5 Research Objectives and Approaches

Building on the identified knowledge gaps, this dissertation is organized around two research thrusts (Figure 1.1). The first focuses on vision-based WMSD risk assessment, developing an ergonomics-oriented human motion dataset and a computational pipeline for accurate high-DoF joint angle estimation from monocular video. The second focuses on simulation-driven intervention recommendation, developing a computational framework for generating task- and site-specific ergonomics interventions through risk-reduced motion synthesis. These two thrusts are addressed through four specific research objectives.

To construct and evaluate an ergonomics-oriented 3D human motion dataset that captures richer keypoint representations for high-DoF joint angle estimation: Existing 3D human pose datasets represent the body using sparse joint center keypoints, which lack the 3D spatial information needed to resolve individual rotation components at high-DoF joints such as the shoulder, neck, and wrist. Accurately estimating these rotation components is essential for comprehensive WMSD risk assessment using methods such as REBA and RULA. To address this gap, this dissertation designs body surface landmark keypoint representations and corresponding joint angle calculation methods that provide the spatial detail necessary for high-DoF joint angle estimation, and uses marker-based motion capture to collect a large-scale 3D human motion dataset based on these designs. Once developed, this dataset can support training 3D human pose estimation models tailored for ergonomics applications.

To develop and evaluate a computational pipeline that derives accurate high-DoF joint angles from monocular video under real-world deployment conditions: Even with an ergonomics-oriented dataset that captures rich 3D spatial information, translating 2D videos into accurate high-DoF 3D joint angle estimates for WMSD risk assessment requires a computational pipeline that accounts for real-world challenges. This computational pipeline integrates pose estimation, keypoint processing, and joint angle calculation into a cohesive workflow, with specific adaptations incorporated to address real-world deployment conditions. The computational pipeline is validated under controlled laboratory conditions and through qualitative evaluation on real-world industry footage to assess its practical applicability.

To develop and evaluate a motion synthesis method that reduces WMSD risks in synthesized human motions: Existing physics-based motion synthesis methods are designed for general applications and do not account for ergonomic safety. To enable their use in intervention

recommendation, this dissertation incorporates a custom ergonomics reward as an optimization objective into reinforcement learning-based motion synthesis, so that synthesized motions reflect reduced biomechanical demand compared to the captured task motions. The resulting motions are evaluated under varying task and site conditions by comparing their WMSD risks against captured task motions to quantify the risk reduction achieved.

To develop a computational framework that enables task- and site-specific ergonomics intervention recommendation using accessible devices: For simulation-driven intervention to be practical, construction site personnel need to be able to recreate real-world task and site conditions within the simulation environment using accessible devices without substantial technical expertise or manual effort. This dissertation develops a computational framework that leverages accessible devices, such as smartphone-based LiDAR, to capture site geometry and natural language processing to extract task parameters, translating both into the physics-based simulation environment for intervention generation.

To compare keypoint-based and mesh-based pose representations for vision-based WMSD risk assessment: Mesh-based representations have emerged as an alternative to the keypoint-based approach proposed in this dissertation, with more versatility to adapt to different downstream applications, such as computing joint angles under alternative anatomical definitions. However, their impact on joint angle estimation accuracy has not been systematically investigated. This dissertation conducts a controlled comparison by training both a keypoint-based and a mesh-based model on the same dataset with the same backbone architecture, and evaluating their joint angle estimation accuracy against motion capture ground truth on the joint angles used in WMSD risk assessments.

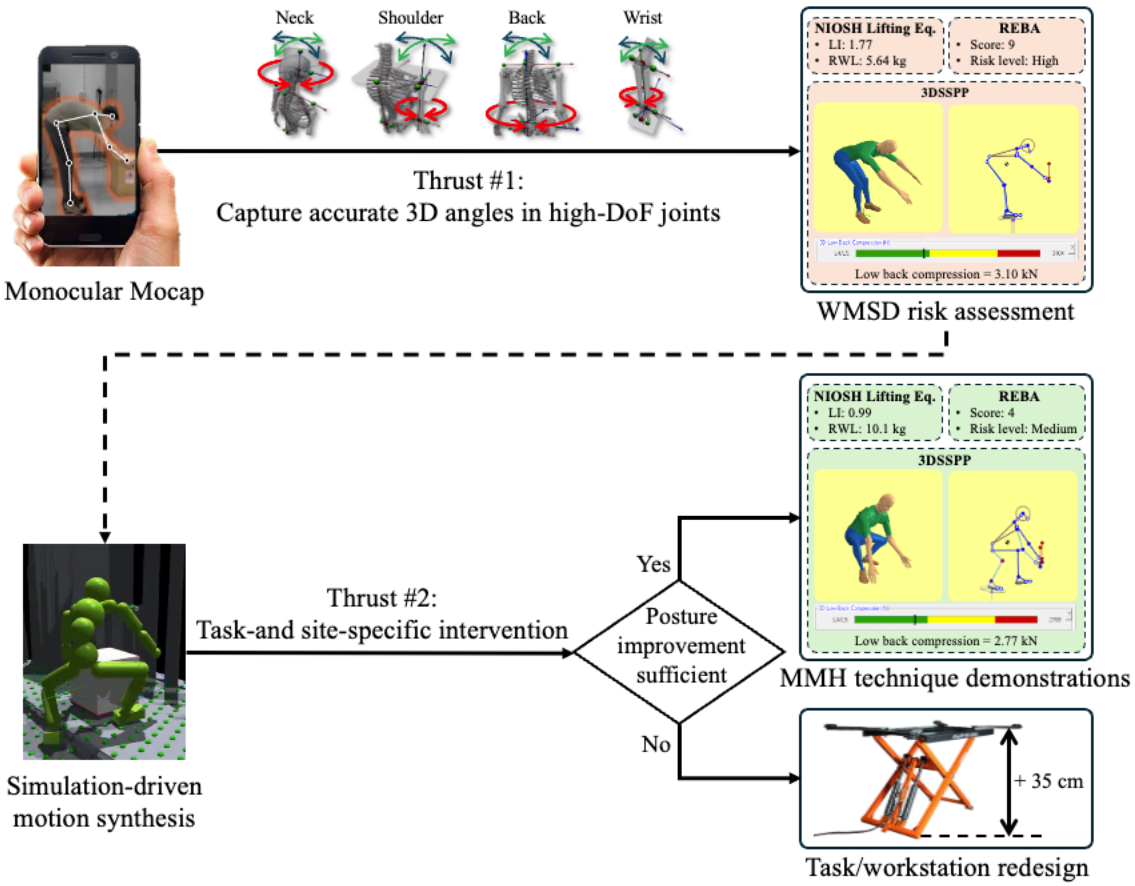


Figure 1.1. Overall research framework.

1.6 The Structure of the Dissertation

In accomplishing these research objectives, the dissertation is structured into 6 chapters. Chapters 1 and 6 provide the introduction and conclusion of this work. Chapters 2 and 3 present the studies on vision-based WMSD risk assessment. Chapter 4 presents the study on simulation-driven ergonomics intervention recommendations. Chapter 5 presents a quantitative comparison between the proposed approach and an emerging alternative for vision-based WMSD risk assessment. The following is the list of chapters.

Chapter 1: Introduction. This chapter presents the background, identifies knowledge gaps in AI-assisted WMSD risk assessment and intervention recommendation, outlines the research objectives, and describes the overall research approaches.

Chapter 2: Ergonomics-Oriented 3D Human Motion Dataset. This chapter introduces the development of a 3D human motion dataset tailored for WMSD risk assessment applications. It presents the 66-keypoint pose representation and corresponding high-DoF joint angle calculation methods, details the data collection process, and evaluates the dataset's effectiveness in training 3D pose estimation models for joint angle calculations.

Chapter 3: Vision-Based WMSD Risk Assessment Computational Pipeline. This chapter develops and evaluates an end-to-end computational pipeline for vision-based WMSD risk assessment. It describes the two-stage model training process: from image-to-2D and from 2D-to-3D pose estimation adapted to real-world conditions, and presents both quantitative and qualitative evaluation results on lab and real-world datasets.

Chapter 4: Intervention Recommendation Applying Scene-Aware Motion Synthesis within Physics-Based Simulations. This chapter proposes a computational framework for generating task- and site-specific ergonomics interventions using motion synthesis within physics-based simulations. It details the reward design for synthesizing risk-reduced motions and evaluates the proposed intervention recommendation framework in diverse task and site scenarios.

Chapter 5: Quantitative Evaluation of Human Pose Representations for WMSD Risk Assessments. This chapter investigates the effectiveness of mesh-based versus keypoint-based human pose representations for vision-based WMSD risk assessment, quantitatively evaluating whether the richer spatial information provided by mesh-based representations offers a viable and more accurate alternative for computing high-DoF joint angles required by WMSD risk assessment methods.

Chapter 6: Conclusions and Recommendations. This chapter summarizes the research conclusions and proposes recommended directions for future work based on this research.

Chapter 2 Ergonomics-Oriented 3D Human Motion Dataset

2.1 Introduction

Monocular motion capture has emerged as a promising approach for vision-based WMSD risk assessment, enabling worker posture monitoring and joint angle estimation from ordinary video cameras (ErgoPlus, 2024; Jeong & Kook, 2023; Kwon et al., 2022; Li et al., 2020; Sundareshan, 2024). Such methods apply machine learning models to infer 3D joint center keypoints from videos. Subsequently, they calculate joint angles from the inferred 3D keypoints for WMSD risk assessments.

While these methods work well for simpler joints like the elbow and knee, they lack sufficient 3D body keypoint information to calculate angles in high-DoF joints, such as the neck, shoulder, or wrist, resulting in oversimplified joint angle calculations. This oversimplification is due to the human pose estimation models in these methods being trained on generic 3D human pose datasets with insufficient 3D keypoints, making them unsuitable for calculating angles in high-DoF joints.

To address this limitation, we defined 66 3D body keypoints that provide sufficient 3D information for joint angle calculations needed in WMSD risk assessments while also containing distinct visual features suitable for vision-based pose estimation. Additionally, we derived joint angle calculation steps that can utilize the estimated 66 body keypoint positions from the vision-based pose estimation model to calculate joint angles needed in WMSD risk assessments. To demonstrate the effectiveness of the proposed keypoint set and joint angle calculation steps, we

collected a large-scale 3D human pose dataset, VelocityEHS-6.7Million (VEHS-6.7M), containing 6.7 million frames of 66-keypoint 3D poses featuring nine categories of typical manual material handling (MMH) and assembly tasks, including lifting/lowering, carrying, pushing/pulling, unboxing, and assembling. We further evaluated the angle estimation capability of a baseline 3D pose estimation model trained on our collected dataset. The proposed keypoint set and joint angle calculation steps serve as the foundation work for training 3D pose estimation models specialized for WMSD risk assessments in industrial settings. Researchers and trained ergonomists can use them to create their own 3D pose dataset tailored for their target task motions. Vision-based pose estimation models trained on such datasets offer sufficient 3D keypoint outputs, enabling joint angle calculations—especially for high-DoF joints—which are critical for assessing WMSD risks.

2.2 Literature Review on Vision-based WMSD Risk Assessments and Datasets

Several existing studies utilize 3D human pose estimation as a foundation for vision-based WMSD risk assessments, typically following a three-step process. First, 2D human poses are extracted from images using 2D pose estimation models. Second, the 2D pose is lifted into 3D using 2D-to-3D pose estimation models. Finally, joint angles are calculated from the estimated 3D pose for WMSD risk assessments. While the first two steps—2D and 3D pose estimation—are well-studied tasks in the computer vision domain, the third step, converting the estimated 3D pose into joint angles for WMSD risk assessments, remains a challenge. The simplified human pose representation used in computer vision domain datasets often only features joint center keypoints and lacks key 3D information, such as rotation information, needed for joint angle calculations (Ionescu et al. 2014).

Consequently, pose estimation models trained on these datasets are limited by the number of keypoints defined in computer vision domain datasets. Chen et al. (2024) fused multiple existing datasets to train a 3D pose estimation model with an expanded 32-keypoint pose. They calculated simplified joint angles as the angles between two 3D vectors (e.g., neck-head and neck-pelvis vectors for the neck angle). This method yields a single combined angle, unable to separate high-DoF joint angle components (e.g., neck flexion, side bending, and rotation, and is restricted to 0°-180°, making it difficult to distinguish between positive and negative angles, such as wrist radial and ulnar deviations. Many existing studies worked around this limitation by making assumptions and approximations when converting 3D poses into joint angles. Chu et al. (2020) used an off-the-shelf 3D pose estimation model (Bogo et al. 2016) to extract 3D body meshes from single images for WMSD risk assessments, but their method only calculated horizontal and vertical body segment projection angles, ignoring all rotation components. The angles also required additional simplified conversions (Li et al., 2019) to fit into definitions commonly used by ergonomists in risk assessments. Similarly, Yu et al. (2019) proposed a projection-based method that projects the 3D pose onto the frontal and sagittal body planes to distinguish the angle components in high-DoF body joints in the trunk, shoulder, and neck. However, due to the lack of sufficient hand, shoulder, and head keypoints in their simplified human pose representation, their method could not estimate wrist angles or capture rotational movements of the neck and upper arm. Jeong & Kook (2023) utilized MediaPipe (Lugaresi et al. 2019) to estimate simplified joint-center 3D human poses from videos. They then applied inverse kinematics to approximately fit a humanoid model to the 3D keypoints for joint angle calculations. Fan et al. (2024) directly addressed the insufficient keypoints in existing datasets by collecting a 3D motion dataset with additional hand keypoints. However, they still adopted a simplified 18-keypoint joint center human pose representation,

requiring approximations in joint angle calculations. For instance, they defined the nose keypoint as the midpoint between the head and thorax. This poor approximation can lead to inaccuracies in neck twisting angles, which were calculated using head-nose and left-right shoulder vector projections on the frontal plane.

Existing 3D human pose datasets in the computer vision domain, built for generic pose estimation tasks (Table 2.1), face two key limitations that restrict their suitability for training pose estimation models for vision-based WMSD risk assessments. First, most datasets only contain human pose annotations in simplified joint center representation (Figure 2.1a), which limits their ability to calculate joint angles for high-DoF joints, such as the neck, shoulder, or wrist. 3D human poses in the simplified joint center representation do not contain enough 3D information to distinguish between different angles in high-DoF joints, such as wrist flexion and deviation, or to calculate intricate joint angles needed in WMSD risk assessments, such as wrist rotations (Kwon et al. 2022; Van Crombrugge et al. 2022). While some datasets using the skinned multi-person linear model (SMPL) (Figure 2.1b) offer 3D angles for each body joint (Lassner et al. 2017; Von Marcard et al. 2018), their definitions do not always align with the angle definitions used in WMSD risk assessments (Wu et al., 2002, 2005), requiring additional simplifications and assumptions during conversion (Chu et al. 2020).



(a) (b) (c)

Figure 2.1. Illustration of 3D human pose representation formats: (a) joint center representation from Human3.6M (Ionescu et al. 2014), visualized by Kudo et al. (2018); (b) mesh-based representation in 3DPW (Von Marcard et al. 2018); (c) our 66 body keypoints.

Second, current datasets primarily focus on generic everyday activities such as sitting, walking, and exercising. (Table 2.1), which deviate from the typical task motions observed in labor-intensive industries where WMSDs are most prevalent. Awkward postures associated with MMH tasks, such as back twisting and backward leaning, are not well-represented in existing datasets (Table 2.1). A recent dataset by Fan et al. (2024) captures 14 construction-related activities, including some MMH tasks such as carrying, dragging, and pushing. However, its focus on construction tasks limits its coverage of the broader range of MMH tasks across industries. Furthermore, with only 0.42 million frames, the dataset is relatively small for training deep learning-based pose estimation models, which can contain tens of millions of trainable parameters (Zhu et al. 2023)

Table 2.1. Overview of human pose datasets that provide 3D pose, 2D pose, and 2D images.

Dataset	Reference	Environment	MoCap Method	Joint #	Frame #	Subject #	Camera #	Activity
HumanEva	Sigal et al. (2010)	MoCap lab	Vicon	15	80k	4	3, 4	Walk, jog, throw/catch, gesture, box, combo
Human3.6 M	Ionescu et al. (2014)	MoCap lab	Vicon	17, 32	3.6M	11	4	Directions, discussion, eating, greeting, phone talk, posing, buying, sitting, smoking, taking photos, waiting, walking
CMU Panoptic	Joo et al. (2015)	MoCap lab	Multi-view RGB	15, 19	1.5M	-	511	ROM, role-play games, dance, musical instrument performance
MPI-INF-3DHP	Mehta et al. (2017)	MoCap lab w/ chroma masks	Multi-view RGB	17	1.3M	8	14	Standing/walking, sitting/reclining, exercise, sports, on the floor, dancing
Unite the people	Lassner et al. (2017)	In the wild	Single view RGB w/ pseudo-GT	14, 91, SMPL	6.8k	-	1	Sports, posing, unspecified subcollection of MPII activities (Andriluka et al. 2014)
TotalCaptu re	Trumble et al. (2017)	MoCap lab	Vicon + wearable IMU	26	1.89M	5	8	ROM, walking, acting, running, freestyle
3DPW	Von Marcard et al. (2018)	In the wild	Single view RGB + wearable IMU	SMPL	51k	7	1	Shopping, sports, hugging, discussing, selfies, riding the bus, playing guitar, relaxing
PROX	Hassan et al. (2019)	Indoor scenes	Single view RGBD w/ pseudo-GT	SMPL	100k	20	1	Interacting w/ objects in bedrooms, living rooms, sitting booths and offices
EgoBody	Zhang et al. (2022)	Indoor scenes	Multi-view RGBD	SMPL	339k	36	4-6	Sitting, standing, walking, dancing, exercising, bending, lying, squatting, interacting w/ objects
BEHAVE	Bhatnagar et al. (2022)	Indoor scenes	Multi-view RGBD	SMPL	15.2k	8	4	Interacting w/ everyday objects like chairs, backpacks, tables, basketballs, yoga balls, etc.
RICH	Huang et al. (2022)	Indoor & outdoor scenes	Multi-view RGB	SMPL	557k	22	6-8	Walking, exercising, climbing over obstacles, etc.
EMDB	Kaufmann et al. (2023)	In the wild & MoCap lab	Wearable EM sensors + iPhone's self-localization	SMPL	105k	10	1	ROM, walking, sitting, jogging, exercising, dancing, boxing, jumping, etc.
CP3D	Fan et al. (2024)	MoCap lab	Vicon	18	421k	7	1	Kneeling, nailing, carrying, dragging, pushing, drilling, grabbing, reaching, installing windows, sitting
VEHS-6.7M	**(Ours)**	MoCap lab	Vicon	19, 66, SMPL	6.7M	10	4	Range of motion movements, poke, pull/push, carry, lift/lower, unbox/pack, assemble w/ back/knee bending, assemble overhead

2.3 3D Human Pose Keypoints and Joint Angle Calculations for Vision-based Pose Estimation

We defined a comprehensive set of 66 3D body keypoints that provide enough 3D information to calculate joint angles in high-DoF joints and are suitable for vision-based pose estimation. Using the 66-keypoint 3D human pose, we derived corresponding calculation steps for 22 joint angles across 6 body joints for WMSD risk assessments (Table 2.2). These include four high-DoF body joints, namely the neck (flexion, lateral flexion, and rotation), shoulder (flexion, abduction, and rotation), back (flexion, lateral flexion, and rotation), and wrist (flexion, deviation, and pronation), which are challenging to calculate using the limited 3D keypoints in the traditional joint center human pose used in prior vision-based WMSD risk assessment methods.

Table 2.2. List of calculated joint angles and their application in WMSD risk assessments.

Body Joints	**Joint Angles**	**REBA** (Hignett & McAtamney, 2000)	**RULA** (McAtamney & Corlett, 1993)
Neck	Flexion/extension	Neck score	Neck score
	Lateral bend	Neck change score	Neck score increase
	Rotation	Neck change score	Neck score increase
Shoulder	Flexion/extension	Upper arms score	Upper arms score
	Abduction/adduction	Upper arms change score	Upper arms score increase
	Rotation	Upper arms change score	-
Elbow	Flexion	Lower arms score	Lower arms score
Wrist	Flexion/extension	Wrists score	Wrists score
	Deviation	Wrists change score	Wrists score increase
	Pronation/supination	Wrists change score	Wrists score increase
Back	Flexion/extension	Trunk score	Trunk score
	Lateral flexion/extension	Trunk change score	Trunk score increase
	Rotation	Trunk change score	Trunk score increase
Knee	Flexion	Legs score	-

2.3.1 66 3D body keypoint positions

We selected these body keypoints to satisfy two criteria. First, the keypoints should provide sufficient 3D spatial information to calculate the required joint angles in common WMSD risk

assessments, such as wrist flexion, deviation, and pronation angles for evaluating the wrist posture score in RULA (McAtamney and Corlett 1993), knee flexion angle for the leg posture score in REBA (Hignett and McAtamney 2000), and back flexion and rotation angle for evaluating the back posture code in OWAS (Karhu et al. 1977; Diego-Mas 2023). Second, the keypoints should contain enough visual features for vision-based pose estimation models to accurately track from videos. To facilitate joint angle calculation, we based our 3D body keypoint positions on Plug-In Gait (Bertaux et al. 2022; Vicon 2021), an established optical motion capture marker layout for biomechanical analysis (Paterson et al. 2017). Plug-In Gait uses reflective markers on bony body landmarks to minimize soft-tissue artifacts, an ideal setup for marker-based motion capture systems. However, some marker positions, such as "left back of head" or "right upper arm," lack prominent visual features, making them unsuitable for vision-based pose estimation. Vision-based pose estimation models need to consistently identify and track the same 3D keypoint positions from videos without the help of reflective markers. Keypoints lacking visual features can compromise the accuracy of 3D keypoint estimation and, consequently, angle estimations. For example, without reflective markers, even human observers may struggle to consistently determine the precise location of the "left back of head" across video frames. For this reason, we replaced keypoints with weak visual features (e.g., "left back of head") with more visually prominent alternatives (e.g., the ear). Furthermore, we introduced additional keypoints to capture joint rotations, including keypoints on both the back and front of the shoulder for shoulder rotation and hand knuckles for wrist pronation. In total, 47 body surface keypoints were captured in each motion capture session (Table 2.3, Figure 2.2a).

Table 2.3. 47 body surface keypoint acronyms and full names.

Acronyms	Keypoint names	Acronyms	Keypoint names
HDTP	Head top	L/RUS	Left/right ulnar styloid process
MDFH	Middle forehead	L/RMCP2	Left/right 2nd metacarpophalangeal
L/REAR	Left/right ear	L/RMCP5	Left/right 5th metacarpophalangeal
SS	Suprasternal notch	L/RPSIS	Left/right posterior superior iliac spine
XP	Xiphoid process	L/RASIS	Left/right anterior superior iliac spine
C7	Spine of the 7th cervical vertebra	L/RGT	Left/right greater trochanter
C7_d	Thoracic vertebra horizontal to SS	L/RMFC	Left/right medial femoral condyle
T8	Spine of the 8th thoracic vertebra	L/RLFC	Left/right lateral femoral condyle
L/RAP	Left/right acromion process	L/RMM	Left/right medial malleolus
L/RAP_f	Front of left/right shoulder joint	L/RLM	Left/right lateral malleolus
L/RAP_b	Back of left/right shoulder joint	L/RMTP1	Left/right 1st metatarsophalangeal
L/RLE	Left/right lateral humeral epicondyle	L/RMTP5	Left/right 5th metatarsophalangeal
L/RME	Left/right medial humeral epicondyle	L/RHEEL	Left/right heel
L/RRS	Left/right radial styloid process		

Compared to the simplified joint center human pose used in existing datasets, the expanded surface keypoint set encompasses a more comprehensive human pose representation, containing key information for calculating intricate joint angles. We first calculated 19 joint center keypoints from the surface keypoints (Figure 2.2b). These keypoints align with the commonly used joint center keypoint locations, making them compatible for training alongside existing datasets. In total, 66 body keypoints, including the 47 body surface keypoints and 19 joint center keypoints, are reported in the collected dataset.

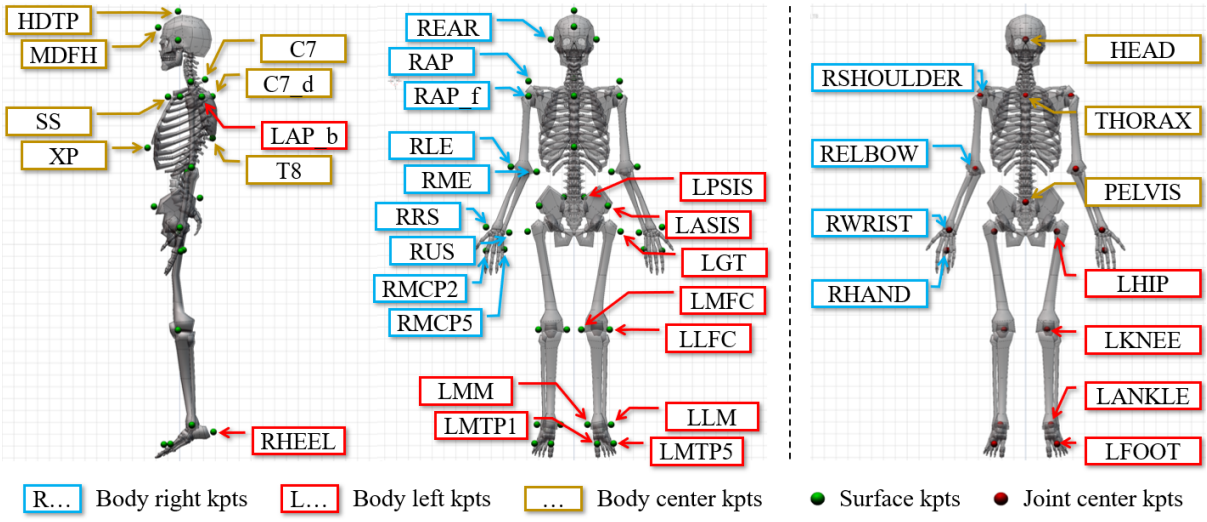
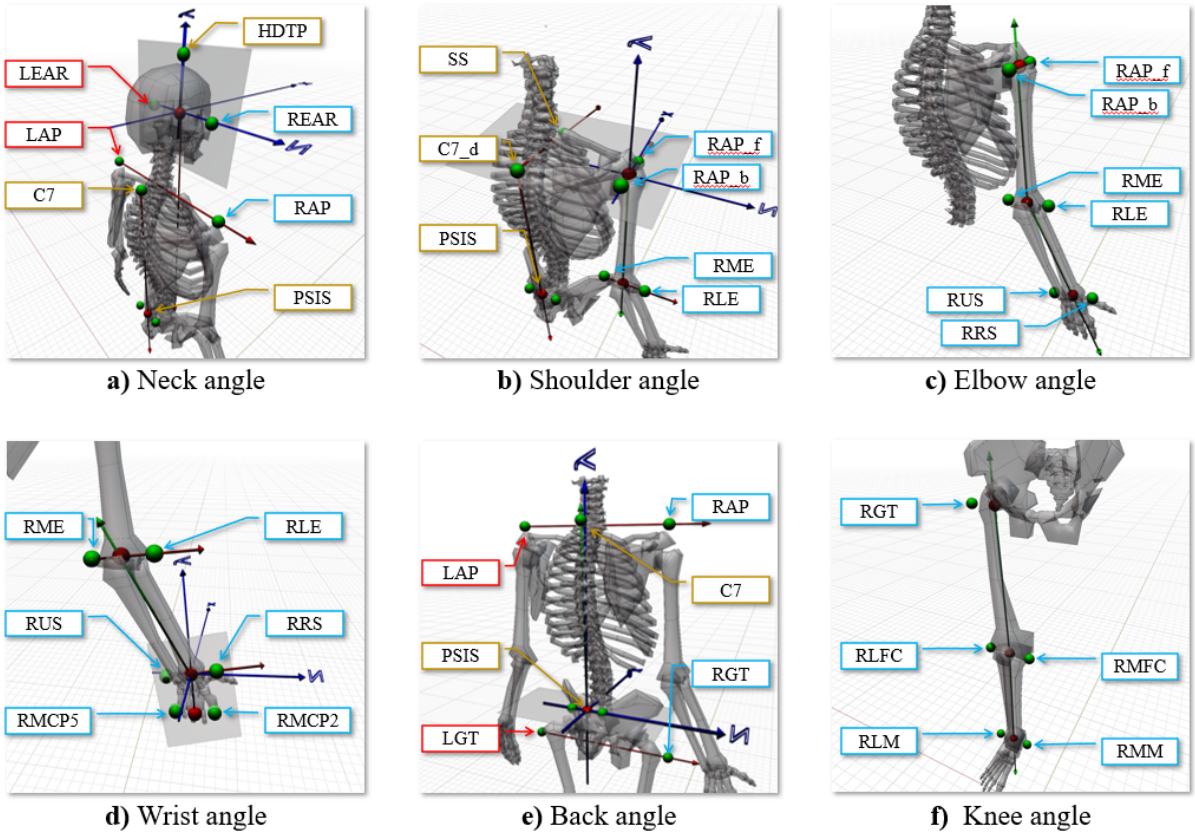


Figure 2.2. Visualization of a) 47 body surface keypoint positions (acronym explanation listed in Table 2.3); and b) 19 calculated joint center keypoint positions.

2.3.2 Joint angle calculations using the 66-keypoint 3D pose

From our set of 66 3D keypoints, we provided detailed angle calculation steps for each joint. In order to calculate the angle components in the high-DoF joints like the neck, shoulder, wrist, and back (Figure 2.3a, b, d, e), we first defined segment coordinate systems following Wu et al. (2002, 2005) with the X-axis pointing towards the anterior direction, the Y-axis pointing towards the superior direction, and the Z-axis in the lateral direction to the right-hand side. The principal joint angles (e.g., neck flexion and lateral flexion) were calculated by projecting 3D segment vectors (e.g., back vector) onto the sagittal (X-Y) and coronal (Y-Z) planes and comparing them with the coordinate axes. Rotation angles were calculated as the dihedral angle between two planes defined using keypoints from the corresponding body segment undergoing rotation. For example, in the case of neck rotation, we calculated the dihedral angle between one plane defined by the ears and the center of the shoulders, and a second plane defined by the shoulders and the center of the ear. For low-DoF joints such as the elbow and knee (Figure 2.3c, 2f), the joint angle was calculated as the 3D angle between the proximal and distal segment vectors. Detailed

calculation steps for each joint angle are provided in the sections below and illustrated in Figure 2.3.



a) Neck angle **b)** Shoulder angle **c)** Elbow angle

d) Wrist angle **e)** Back angle **f)** Knee angle

Figure 2.3. Diagrams illustrating segment coordinate system definitions for joint angle calculations.

2.3.2.1 Neck angle calculations:

Neck flexion/extension, lateral flexion, and rotation are calculated from 7 keypoints (i.e., HDTP, LEAR, REAR, C7, LSHOULDER, RSHOULDER, and PELVIS, as illustrated in Figure 2.3a). The center of the head, calculated as the center point between LEAR and REAR, is defined as the origin of the head coordinate system. A coronal head plane is created using the left and right ear as well as the HDTP keypoint. The X-axis of the head coordinate system is defined as the vector orthogonal to the coronal plane pointing in the anterior direction. The Y-axis is defined as the vector from the head center to the HDTP keypoint. The Z-axis is defined using the cross-product of the other two axes. The back vector is defined from C7 to the PELVIS keypoint, using

keypoints in similar positions as the C7-L5 torso vector defined in Baer & Cohen (2023). For angle calculation, the neck flexion/extension is calculated by projecting the back vector onto the X-Y plane of the head coordinate system. Similarly, the lateral flexion is calculated as the Y-Z plane projection. The neck rotation angle is calculated as the dihedral angle between the plane defined by LEAR, REAR, and the shoulder center, and the plane defined by LSHOULDER, RSHOULDER, and the ear center.

2.3.2.2 Shoulder angle calculations:

Using the right shoulder as an example, the shoulder flexion/extension, abduction/adduction, and rotation are calculated from 7 keypoints (i.e., RAP_b, RAP_f, SS, C7_d, RME, RLE, and PELVIS as illustrated in Figure 2.3b). The shoulder joint center, defined as the center of RAP_b and RAP_f, is the origin of the shoulder coordinate system. The Y-axis of the shoulder coordinate system is defined as the superior vector PELVIS-C7_d. A transverse shoulder plane is created orthogonal to the PELVIS-C7_d vector and passes through the shoulder origin. The X-axis is defined as the C7_d-SS vector's projection on the transverse shoulder plane. The Z-axis is defined using the cross-product of the other two axes. The humerus vector is defined from the shoulder origin to the center of the elbow keypoints (i.e., RME, RLE). For angle calculation, the shoulder flexion/extension is calculated by projecting the humerus vector onto the X-Y plane of the shoulder coordinate system. Similarly, the abduction/adduction is calculated as the Y-Z plane projection. The shoulder rotation angle is calculated as the dihedral angle between the plane defined by RAP_b, RAP_f, and the elbow center and the plane defined by RME, RLE, and the shoulder center.

2.3.2.3 Elbow angle calculations:

Using the right elbow as an example, the elbow joint center is calculated as the center point of RLE and RME (Figure 2.3c). The wrist joint center is calculated as the center point between RRS and RUS. The forearm vector is defined from the elbow to the wrist joint center. Elbow flexion is calculated as the angle between the forearm and humerus vectors.

2.3.2.4 Wrist angle calculations:

Using the right wrist as an example, the wrist flexion/extension, deviation, and pronation/supination are calculated from 6 keypoints (i.e., RLE, RME, RRS, RUS, RMCP2, and RMCP5, as illustrated in Figure 2.3d). The wrist joint center is defined as the origin of the hand coordinate system. A coronal hand segment plane is created using the knuckle keypoints RMCP2 and RMCP5, as well as the wrist joint center. The X-axis of the hand segment is defined as the vector orthogonal to the coronal plane pointing in the anterior direction. The Y-axis is defined as the vector from the center of the knuckle keypoints to the wrist joint center. The Z-axis is defined using the cross-product of the other two axes. For angle calculation, the wrist flexion/extension is calculated by projecting the forearm vector onto the X-Y plane of the hand coordinate system. Similarly, the deviation is calculated as the Y-Z plane projection. The wrist pronation/supination angle is calculated as the dihedral angle between the plane defined by RLE, RME, and the wrist center and the plane defined by RRS, RUS, and the elbow center.

2.3.2.5 Back angle calculations:

Back flexion/extension, lateral flexion/extension, and rotation are calculated from 6 keypoints (i.e., C7, PELVIS, LGT, RGT, LSHOULDER, and RSHOULDER, as illustrated in Figure 2.3e), and a reference up direction in the world coordinate frame. The origin of the back coordinate system is defined as the PELVIS keypoints. The Y-axis of the back coordinate system

is defined as the reference up direction. A transverse plane is created orthogonal to the Y-axis and passes through the PELVIS center. The X-axis is defined as the PELVIS-RGT vector's projection on the transverse plane. The Z-axis is defined using the cross-product of the other two axes. For angle calculation, the back flexion/extension is calculated by projecting the back vector onto the X-Y plane of the back coordinate system. Similarly, the lateral flexion/extension is calculated as the Y-Z plane projection. The back rotation angle is calculated as the dihedral angle between the plane defined by LGT, RGT, and the shoulder center, and the plane defined by LSHOULDER, RSHOULDER, and the hip center.

2.3.2.6 Knee angle calculations:

Using the right knee as an example, the right hip joint center is calculated by translating the RGT keypoint 2 inches in the RGT-LGT vector direction (Figure 2.3f). The knee joint center is calculated as the center point between the RLFC and RMFC keypoints. The ankle joint center is calculated as the center point between RMM and RLM keypoints. The femur vector is defined from the knee to the hip joint center. The tibia vector is defined from the knee to the ankle joint center. Elbow flexion is calculated as the angle between the femur and tibia vectors.

2.4 3D Human Pose Dataset Collection

To test the proposed 66-keypoint set's capability in enabling vision-based joint angle estimation needed for WMSD risk assessments, we collected VEHS-6.7M, a large-scale human pose dataset featuring 20 participants performing 9 categories of MMH and assembly tasks, using a marker-based motion capture system. The following sections detail our motion capture equipment setup, participant characteristics, collected task motions, and data post-processing details, with an overview visualized in Figure 2.4.

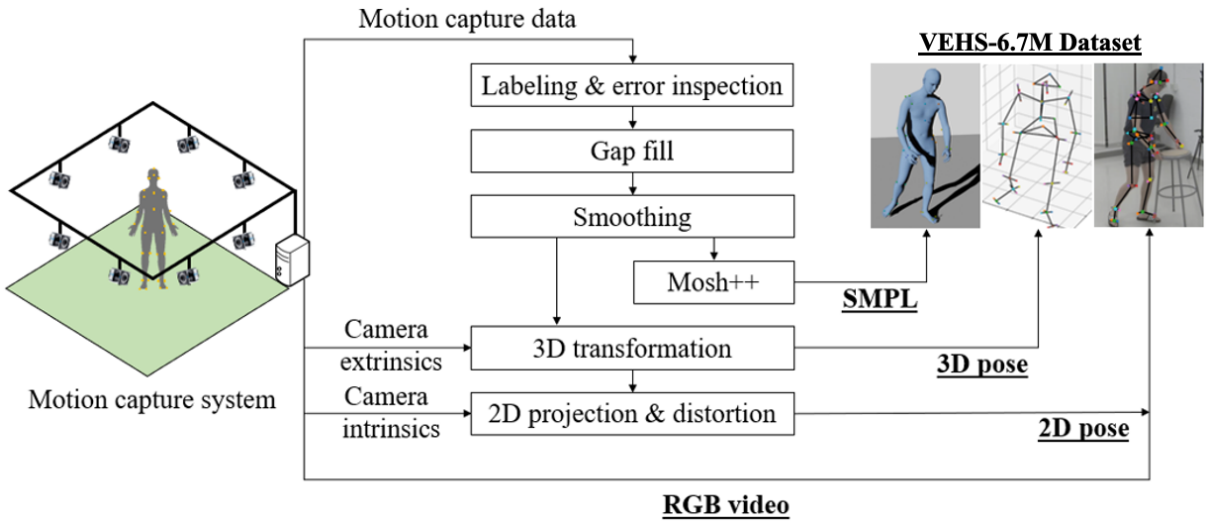


Figure 2.4. Overview of the 3D human pose dataset collection process.

2.4.1 Motion capture equipment setup

Our motion capture laboratory is equipped with eight Vicon® Vero v1.3 motion capture cameras and four FLIR® RGB video cameras (Vicon, 2022, 2023). The eight motion capture cameras were evenly spaced on wall railings, while the four RGB cameras were positioned on two of the four sides of the room (Figure 2.5a). The four RGB cameras were set at different heights: one at chest level (~1.5m), one at eye level (~1.7m), one slightly above head height (~2.2m), and the last one at a surveillance camera's vantage point (~2.7m) (Figure 2.5b) to capture videos with diverse camera pitch angles that resemble common industrial settings.

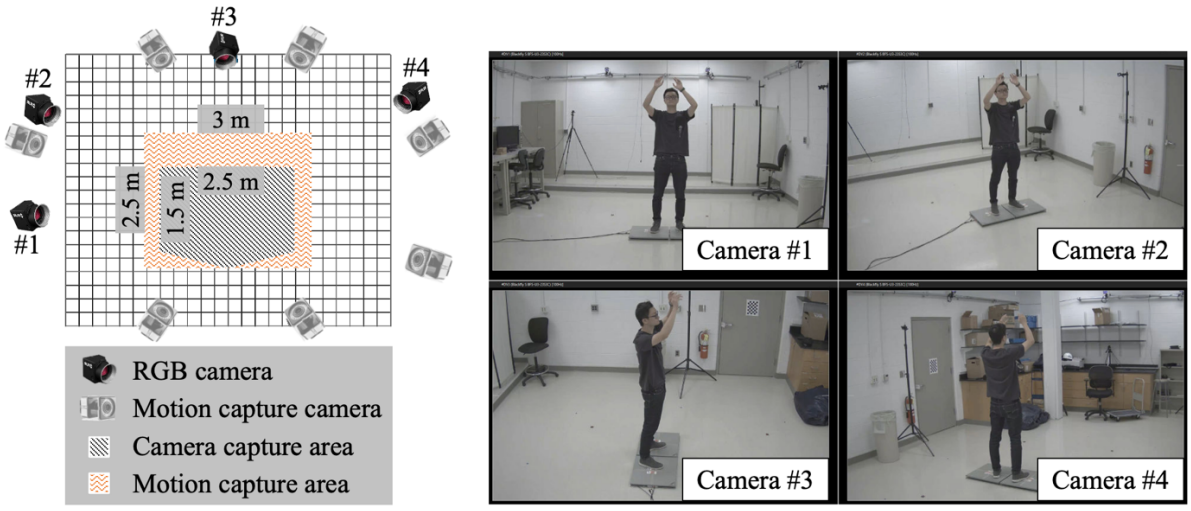


Figure 2.5. a) Vicon motion capture and RGB video capture setup; b) example camera views.

2.4.2 Participant characteristics

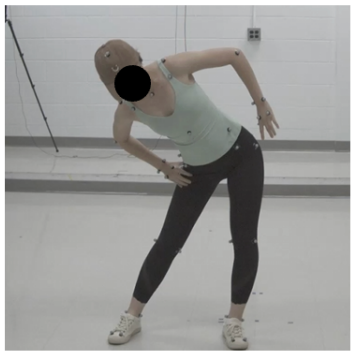
We recruited ten men and ten women with a body mass index (BMI) range of 17.53 to 32.23 (Table 2.4). The BMI range is comparable to previous datasets, while our dataset includes 82% more participants than the widely used Human3.6M dataset (Ionescu et al. 2014). This research was approved by the University of Michigan Institutional Review Board (HUM00209880). Written informed consent was obtained from each participant before their participation.

Table 2.4. Height, weight, and BMI of participants sorted by gender

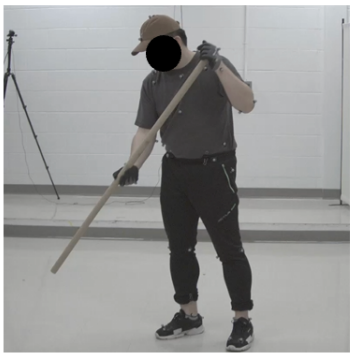
Participant #	Gender	Height (m)	Weight (Kg)	BMI
#1	Man	1.76	54.3	17.5
#2	Man	1.73	54.0	18.0
#3	Man	1.82	70.3	21.2
#4	Man	1.82	73.4	22.2
#5	Man	1.81	76.7	23.4
#6	Man	1.75	74.0	24.2
#7	Man	1.72	73.0	24.7
#8	Man	1.62	65.0	24.8
#9	Man	1.93	114.5	30.7
#10	Man	1.75	98.7	32.2
#11	Women	1.62	46.7	17.8
#12	Women	1.66	52.0	18.9
#13	Women	1.61	52.0	20.1
#14	Women	1.64	55.7	20.7
#15	Women	1.65	56.4	20.7
#16	Women	1.66	59.0	21.4
#17	Women	1.65	61.7	22.7
#18	Women	1.72	70.0	23.7
#19	Women	1.69	72.5	25.4
#20	Woman	1.72	78.2	26.4

2.4.3 MMH and assembly task motions

We based our task selection on typical MMH tasks associated with high WMSD risks, including push/pull, lift/lower, and carry (Ciriello et al. 1999; Yang et al. 2020). We then collaborated with three certified professional ergonomists, each with over 15 years of experience, to expand the initial selections by incorporating common and high-risk motions not adequately represented in the selected task categories, including poke, unbox/pack, and assemble (Figure 2.6).



Task Category #1: General motions, including full ROM in standing and sitting positions



Task Category #2: Material manipulation using a pole



Task Category #3: Push, pull, and load hand trucks with objects of different weights.



Task Category #4: Carry objects horizontally between various start and end locations



Task Category #5: Lift and lower objects between various start and end locations



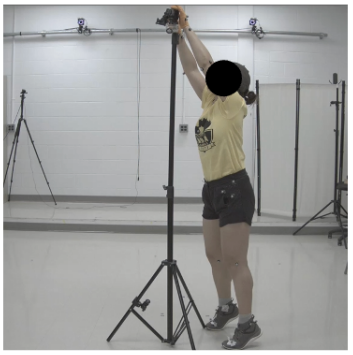
Task Category #6: Unbox and repack items of varying weight and shapes.



Task Category #7: Assemble at knee height while bending the back.



Task Category #8: Assemble at knee height while bending the knees.



Task Category #9: Assemble in overhead locations.

Figure 2.6. Images and descriptions of the selected task categories.

Capturing a dataset with diverse human motions is a key consideration when designing our dataset. Rather than following a rigid sequence of task variable combinations, which could result in capturing repetitive or similar motions across participants, we instructed participants to randomly select from available task parameters, such as different speeds, exerted force, movement patterns, and target locations, and combine them during execution. A comprehensive list of considered task parameters is provided in Table 2.5. On average, our dataset captures 29.8 unique combinations per participant per task category, resulting in approximately 5,364 task variations.

Using the carry task as an example, we provided four types of objects with varying weights, shapes, and handle types. We also specified four general start and end carrying locations, allowing participants to create diverse combinations. Additionally, participants were verbally instructed to vary their carrying posture throughout the session, such as carrying over the shoulder, at their side, or one-handed. In each of the task categories, participants were also instructed to simulate scenarios with constrained space, such as reaching into the interior of a car frame to assemble components on the roof and twisting or leaning around a structural beam to access roof fasteners.

To further improve motion diversity and minimize constraints imposed by predefined task descriptions, we added a general motion task category designed to capture anatomically feasible but rarely performed motions. During this task, participants were instructed to move their joints through their full range of motion in both standing and sitting positions, performing actions such as forward bends to touch their toes, stretches, and jumping jacks. These motions can help the pose estimation model learn the range of motion of each joint.

Table 2.5. List of available task parameters in each task category.

Task Categories	Average # of Combinations / Participant	Task Parameters
General motions	31.0	Bending, stretching, squatting, sitting, jogging, walking, jumping jack, etc.
Poke	39.0	Location, poke direction, facing direction, one/two hand, movement speed, applied force
Push/pull	28.8	Cart type, empty/weighted, start & end location, facing direction, cart size, one/two hand, movement speed, applied force
Carry	24.4	Start & end location, item type, item number, item weight, one/two hand, hand switching, movement speed, carrying method
Lift/lower	33.8	Start & end location, item type, item weight, facing direction, one/two hand, hand switching, movement speed, back/knee bending, lifting method
Unbox/pack	38.8	Start & end location, one/two hand, hand switching, movement speed
Assemble w/ back bending	26.0	Location, height, facing direction, one leg forward/both legs back, back twisting
Assemble w/ knee bending	22.6	Location, height, facing direction, one leg forward/both legs back
Assemble overhead	24.2	Location, height, facing direction, one leg forward/both legs back, back twisting

2.4.4 Data post-processing

After the motion capture session, the collected 3D marker trajectories (Figure 2.7a) were converted into human pose annotations using a five-step process (Figure 2.4). First, the markers were auto-labeled using our 66-keypoint labeling template (Figure 2.7b) and manually inspected frame-by-frame for errors. Second, occluded markers were filled using temporal and spatial information from surrounding markers (Figure 2.7c). Third, the marker trajectories were smoothed to reduce high-frequency noise. Fourth, the marker trajectories were transformed into the camera coordinate system to generate 3D poses (Eq. (1)). Lastly, the 3D poses were projected onto the 2D image plane (Eq. (2)) and applied corresponding camera distortion (Eq. (3)-(6)), resulting in 2D poses that align with the 2D RGB images in the pixel coordinate system. As an additional output,

we also estimated the body surface mesh in SMPL and SMPLX format from the body surface marker positions using Mosh++ (Mahmood et al. 2019). This data captures each participant's body shape, which can be used to train human body shape estimation models.

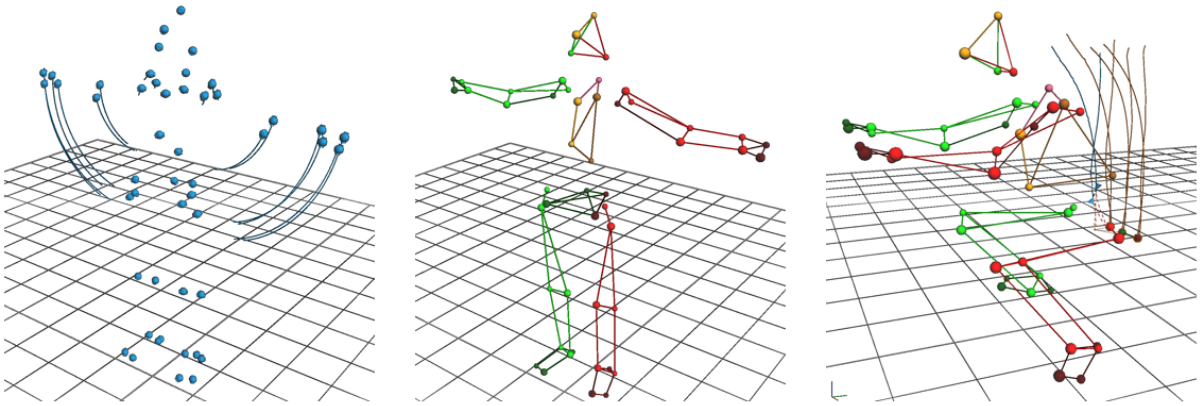


Figure 2.7. Visualization of: a) raw motion capture marker trajectories; b) auto-labeled markers; c) gap-filling process of the occluded LASIS marker using temporal and spatial information of surrounding markers.

$$3D\ Pose = \begin{pmatrix} X_{camera} \\ Y_{camera} \\ Z_{camera} \\ 1 \end{pmatrix} = \begin{pmatrix} R_{3\times3} & t_{3\times1} \\ 0_{3\times1} & 1_{1\times1} \end{pmatrix} \cdot \begin{pmatrix} X_{world} \\ Y_{world} \\ Z_{world} \\ 1 \end{pmatrix} \quad \text{Eq. (1)}$$

$$\begin{pmatrix} u_{camera} \\ v_{camera} \\ 1 \end{pmatrix} = \begin{pmatrix} f & k & u_{pp} \\ 0 & f/a & v_{pp} \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} X_{camera} \\ Y_{camera} \\ Z_{camera} \end{pmatrix} \quad \text{Eq. (2)}$$

$$dp = \begin{pmatrix} du \\ dv \end{pmatrix} = \begin{pmatrix} u_{camera} - u_{pp} \\ a \cdot v_{camera} - v_{pp} \end{pmatrix} \quad \text{Eq. (3)}$$

$$r = \|dp\| \quad \text{Eq. (4)}$$

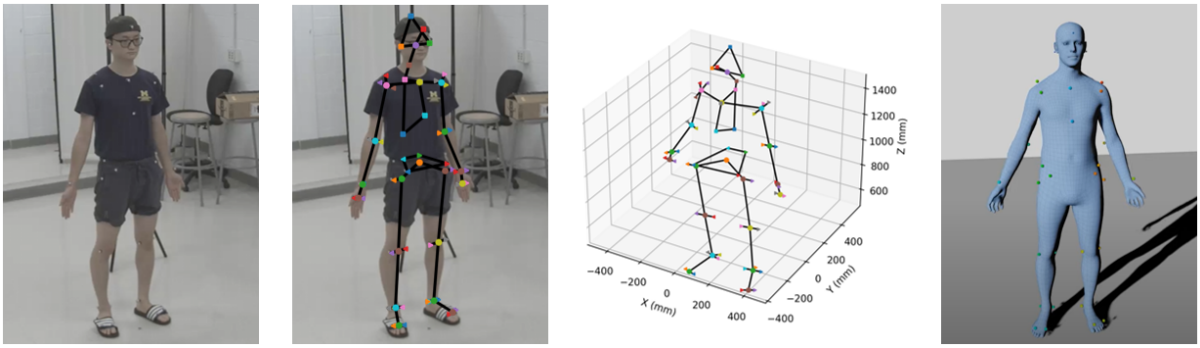
$$s = 1 + w_0 r^2 + w_1 r^4 + w_2 r^6 \quad \text{Eq. (5)}$$

$$2D\ Pose = \begin{pmatrix} u_{camera_distorted} \\ v_{camera_distorted} \end{pmatrix} = \begin{pmatrix} du/s + u_{pp} \\ dv/(s \cdot a) + v_{pp} \end{pmatrix} \quad \text{Eq. (6)}$$

Where $t_{3\times1}$ is the camera's 3D position, $R_{3\times3}$ is the camera rotation matrix, f is focal length, k is skew, a is pixel aspect ratio, w_0, w_1, w_2 are the radical distortion parameters, (u_{pp}, v_{pp}) is the distortion center and principal point.

2.4.5 Characteristics of the collected 3D human pose dataset

The resultant 3D human pose dataset that we collected contains four types of data: a) 2D RGB videos; b) 2D human pose annotations in pixels; c) 3D human pose annotations in millimeters; and d) SMPL body mesh (Figure 2.8).



(a) **(b)** **(c)** **(d)**

Figure 2.8. VEHS-6.7M dataset includes a) 2D RGB videos; b) 2D human pose annotations; c) 3D human pose annotations; and d) SMPL annotations.

The dataset features 20 participants performing 9 categories of MMH and assembly tasks simulating industrial activities, capturing an estimated 5,364 task variations. Compared to previous datasets that primarily focus on generic activities, our dataset is more closely aligned with motions typically observed in industrial environments, making it better suited for training models for industrial ergonomics applications. Compared to prior datasets that primarily capture joint center, our dataset includes 66-keypoint 3D annotations selected for joint angle calculations, especially for high-DoF joints. Each captured session lasts over three minutes and is recorded at 50 frames per second (fps) from four camera angles. In total, the dataset accumulates 37 hours of video footage, or 6.7 million frames—almost doubling the frames of the widely used Human3.6M dataset (Ionescu et al. 2014). The large number of frames is essential for training deep 3D pose estimation models, which contain large numbers of trainable parameters. Among the participants, 14 were used for training, and 6 for testing. A summary of the dataset is shown in Table 2.6.

Table 2.6. Number of captured train and test frames broken down by task.

#	Task Categories	Captured Frames	
		Train	Test
#1	General motions	525,077	238,905
#2	Poke	481,950	231,192
#3	Push/pull	521,104	233,340
#4	Carry	515,570	234,565
#5	Lift/lower	507,169	238,341
#6	Unbox/pack	511,751	237,017
#7	Assemble w/ back bending	512,492	232,957
#8	Assemble w/ knee bending	475,138	234,673
#9	Assemble overhead	505,578	222,666
	Total	**4,555,829**	**2,103,656**

2.5 Evaluation of the Dataset's Effectiveness in Training a 3D Pose Estimation Model for Joint Angle Calculations

Using the collected VEHS-6.7M training set, we trained a baseline 3D human pose estimation model to evaluate the dataset's effectiveness in training pose estimation models specialized for joint angle calculations. Current vision-based pose estimation methods with leading performance typically follow a two-step process: extracting the 2D pose from images and then estimating the 3D pose from the 2D pose (Zhu et al. 2023). While 2D pose estimation models can define new keypoints using manual annotation on the abundant existing 2D datasets (Duan et al., 2019; Jin et al., 2020; Wu et al., 2018), adding new keypoints in 3D pose estimation models is more complicated, requiring dedicated motion capture sessions with the keypoint set predetermined before data collection. The selected keypoints also need to be tested to ensure their physical marker locations are compatible with the target MMH and assembly task motions and are not constantly obstructed by body parts or task-related items (e.g., hand trucks and boxes). For the scope of this chapter, we focused on the more complicated process of adding our newly defined 66 3D keypoints to the 3D pose estimation model in the vision-based pose estimation process.

For the 3D pose estimation model architecture, we selected MotionBert (Zhu et al. 2023) for its leading performance on the Human3.6M benchmark dataset. The MotionBert model applies a Dual-stream Spatio-temporal Transformer that utilizes both spatial and temporal information to estimate 3D pose from 2D pose input. We modified the first and last layer of the MotionBert model to allow for both 66-keypoint 2D input and 3D output. The model was trained for 100 epochs on five NVIDIA A40 GPUs with a batch size of 32, which took approximately 50 hours. We used the AdamW optimizer with a learning rate of 0.0002 and a decay factor of 0.99, following the training hyperparameters recommended in MotionBert.

2.5.1 3D pose evaluation result

Obtaining an accurate 3D human pose is a critical intermediate step for vision-based WMSD risk assessments, as it directly influences subsequent joint angle calculation accuracy. We evaluated the estimated 3D pose against the ground truth 3D pose collected using marker-based motion capture on the reserved test set.

The trained MotionBert model achieved a Mean Per Joint Position Error (MPJPE) of 15.4 mm and Procrustes-aligned MPJPE (PA-MPJPE) of 11.1 mm for the 66 keypoints, and an MPJPE of 10.3 mm and PA-MPJPE of 14.2 mm when only evaluated on a subset of 17 joint center keypoints that best match the Human3.6M 17-keypoint format. This indicates that, on average, the estimated 3D keypoints are within 2 cm of the ground truth position. Though not a strict comparison due to the difference in the evaluation dataset, the original MotionBert model trained on the Human3.6M dataset achieved a similar MPJPE of 16.9 mm when evaluated on the 17-keypoint pose of the Human3.6M validation set (Zhu et al. 2023). Additionally, the model runs at 737 frames per second on a single NVIDIA Tesla V100 GPU during inference, making it suitable for real-time WMSD risk assessments.

2.5.2 Joint angle evaluation result

We evaluated the angle estimation accuracy to demonstrate the effectiveness of our 66-keypoint set and angle calculation steps in enabling accurate joint angle estimation for WMSD risk assessments. Using the estimated 66-keypoint 3D pose, we calculated 22 joint angles from 6 body joints commonly used in WMSD risk assessments (Table 2.2). These include angles from 4 high-DoF body joints, namely the neck (flexion, lateral flexion, and rotation), shoulder (flexion, abduction, and rotation), back (flexion, lateral flexion, and rotation), and wrist (flexion, deviation, and pronation). These high-DoF joint angles, which are challenging to calculate using the limited 3D keypoints in the simplified joint center human pose used by previous methods, can now be calculated using the estimated 66-keypoint 3D human pose and proposed angle calculation steps. For accuracy evaluation, we compared the joint angles calculated from the estimated 3D pose with the ground truth angles collected using the marker-based motion capture system. For each angle, we reported the mean absolute error (MAE) and the median absolute error in Table 2.7. Averaging across all joint angles, our estimated angles achieved an MAE of 2.39º and a median absolute error of 1.43º.

Table 2.7. Joint angle estimation errors (in degrees)

	Angle Name	MAE	Median Abs. Err.	RMSE	Mean	SD	Skewness	Kurtosis	IQR
Neck	Flexion/extension	1.62	1.04	2.48	-0.10	2.48	-1.61	22.05	2.07
	Lateral flexion	1.54	1.15	2.12	0.41	2.08	-0.55	9.99	2.31
	Rotation	1.89	1.29	4.55	0.23	4.54	-0.95	385.15	2.57
Shoulder	Flexion/extension	2.64	1.62	5.42	-0.16	5.42	-0.91	184.48	3.19
	Abduction/adduction	2.67	1.34	6.38	-0.14	6.37	-0.04	167.60	2.67
	Rotation	3.52	2.63	5.08	1.42	4.87	0.49	33.55	5.02
Elbow	Flexion	1.78	1.19	2.76	0.12	2.76	1.70	38.18	2.38
Wrist	Flexion/extension	2.58	1.35	4.63	-0.05	4.63	-0.31	39.21	2.70
	Radial/ulnar deviation	2.51	1.53	4.08	0.20	4.08	-0.44	27.33	3.03

	Pronation/supination	3.44	2.28	5.51	0.05	5.51	-0.29	73.81	4.56
Back	Flexion/extension	1.18	0.72	1.84	0.17	1.83	0.94	10.06	1.42
	Lateral flexion	2.49	1.11	7.72	-0.57	7.70	-1.58	196.54	2.26
	Rotation	2.01	1.42	2.78	-0.83	2.65	-0.30	2.52	2.95
Knee	Flexion	1.80	1.24	2.69	0.48	2.64	0.88	44.21	2.45
	Average	**2.39**	**1.43**	**4.57**	**0.14**	**4.57**	**-0.60**	**205.74**	**2.86**

Additionally, we analyzed the variability in joint angle estimation errors. We reported the mean, standard deviation (SD), skewness, kurtosis, and interquartile range (IQR) of the error in Table 2.7. The overall mean error across the entire dataset was 0.14°, with a standard deviation of 4.57°, demonstrating an improvement over Yu et al. (2019), which reported a larger bias and spread with a -0.70° mean error and 8.21° standard deviation. We also provided histograms of the angle estimation error separated by joint angles in Figure 2.9, overlaid with the normal distribution curve calculated using the error mean and standard deviation. The histograms show a noticeable deviation from the expected normal distribution. We further conducted Kolmogorov-Smirnov tests, which rejected the null hypothesis of normality across all joint angles and confirmed that the error distributions deviate significantly from normal distributions. Given the non-normal distribution, standard deviation may not be the best indicator for characterizing the error distribution. Alternative metrics like median and IQR (Table 2.7), along with the histogram plots (Figure 2.9), offer a more comprehensive representation of the angle error distribution.

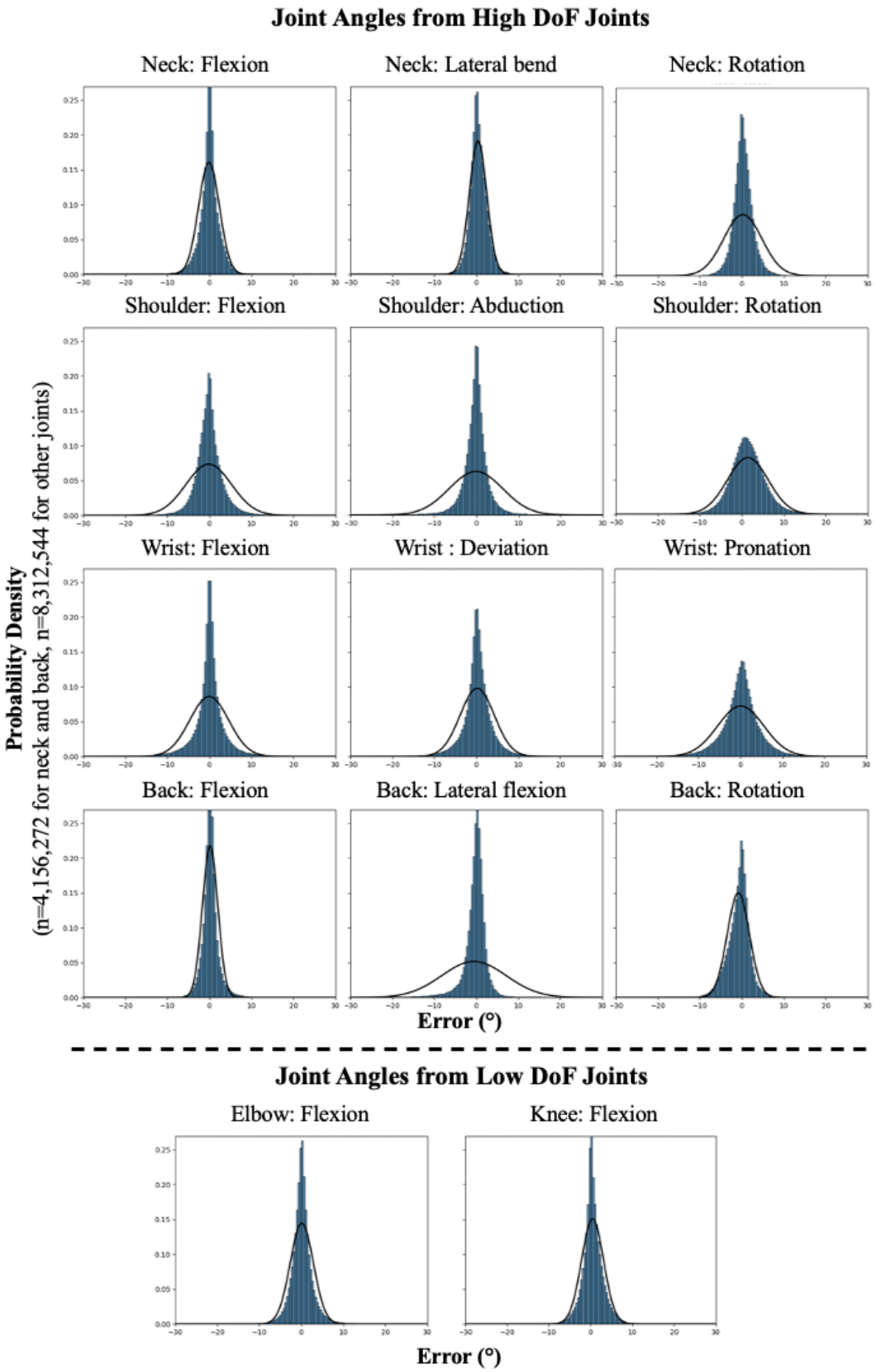


Figure 2.9. Histograms of angle estimation errors with GT2D input, overlaid with the normal distribution curves computed from the corresponding means and standard deviations.

2.6 Discussion

Capturing a dataset with diverse motions is essential for training a robust 3D human pose estimation model. Estimating 3D human pose from 2D inputs is inherently ambiguous, as the model must infer missing depth information by analyzing spatial relationships between body joints, understanding joint kinematics, and leveraging temporal correlations. Pose estimation models achieve this through a data-driven approach, leveraging large amounts of diverse example inputs (i.e., 2D pose) and outputs (i.e., 3D pose in the same frame) to train deep neural networks for accurate pose estimation. For instance, our MotionBert model contains 42,491,405 trainable parameters, requiring a large number of diverse 2D-3D pose pairs to train effectively. Our VEHS-6.7M dataset collected 6.7 million frames of 2D-3D pose pairs, with 4.6 million frames reserved for training. A common concern with video-based datasets, especially those captured at high fps, is that consecutive frames often exhibit minimal pose variation. While video-based datasets provide valuable temporal information, the limited variation between frames may restrict the number of distinct poses available for the model to learn. To address this concern, we reported the distinct frame count as a motion diversity metric, following Ionescu et al. (2014). Two poses are considered distinct if at least one joint differs from its corresponding joint in another frame beyond tolerance t, i.e., $\left\|p_{f1,j} - p_{f2,j}\right\|_2 > t$. Our dataset contains 1.8M distinct frames when using 50 mm tolerance and 907k with 100 mm tolerance (Table 2.8). Compared to the widely used Human3.6M dataset, our dataset captures more than twice the number of distinct frames and includes a higher percentage of them, indicating that our captured motions exhibit greater motion diversity.

Table 2.8. Number of distinct frames captured in the VEHS-6.7M dataset

Tolerance	Dataset	Distinct Frames	Frames @ 50 fps	
			Total	Distinct %
50 mm	VEHS-6.7M	1,805,345	6,659,485	27.1%
	Human3.6M	886,409	3,578,080	24%
100 mm	VEHS-6.7M	907,033	6,659,485	13.6%
	Human3.6M	438,654	3,578,080	12%

Our trained pose estimation model demonstrated accurate angle estimation results with an MAE of 2.4º across 22 joint angles, including 18 angles from 4 high-DoF joints (i.e., neck, shoulder, back, and wrist) with a marginally higher MAE of 2.7º. By utilizing the additional 3D positions in the 66-keypoint set, we can calculate distinct joint angle components along each axis for high-DoF joints instead of being limited to a single combined 3D angle. These detailed joint angle components serve as valuable inputs for WMSD risk assessments such as REBA, RULA, and OWAS, allowing vision-based methods to use these estimated angles for risk assessment rather than relying on simplified assumptions (e.g., assume no neck twist or side bend). Moreover, we trained the 3D pose estimation model using MMH and assembly task motions rather than relying on the generic task motions in standard computer vision datasets (e.g., sitting, walking, exercising). Models trained on such domain-specific motions are expected to perform better when deployed in industrial settings when monitoring workers performing similar tasks.

2.7 Conclusion

We defined a set of 66 3D body keypoints and derived steps to calculate joint angles needed for WMSD risk assessments from vision-based 3D human pose estimation. The proposed 66-keypoint set provides sufficient 3D information for joint angle calculations and contains distinct visual features suitable for vision-based human pose estimation. The proposed 66-keypoints, along with the angle calculation steps, enabled the calculation of 22 angles across 6 body joints, including angles in 4 high-DoF joints (i.e., neck, shoulder, back, and wrist) that are challenging to calculate

using the limited 3D keypoints in the traditional joint center human pose typically adopted by prior vision-based WMSD risk assessment methods. Evaluation test on a baseline 3D human pose estimation model trained on a 66-keypoint pose dataset with ground truth 2D pose input demonstrated accurate joint angle estimation capability, achieving an angle MAE of 2.4º.

Future researchers and trained ergonomists can use the proposed keypoint set to build their own datasets and train vision-based pose estimation models. They can further apply the proposed joint angle calculation steps on the estimated 3D pose output to obtain the necessary joint angles, especially for high-DoF joints, required for WMSD risk assessments.

Chapter 3 Vision-Based WMSD Risk Assessment Computational Pipeline

3.1 Introduction

The previous chapter presented an ergonomics-oriented 3D human motion dataset designed for WMSD risk assessment and evaluated its effectiveness in training 3D human pose estimation models for accurate joint angle estimation under the assumption of perfect 2D pose inputs. In practice, however, 2D pose estimation from images is itself subject to error. Since the 3D pose lifting model takes estimated 2D keypoint positions as input, errors in 2D detection can propagate through the pipeline, affecting the accuracy of downstream risk assessment applications. This is further exacerbated under real-world construction deployment conditions. Unlike controlled laboratory settings, where subjects wear form-fitting clothing suitable for motion capture and are clearly visible against uniform backgrounds with consistent lighting, construction sites present occlusion from objects, tools, or other workers, inconsistent lighting conditions ranging from direct sunlight to shadowed interiors, cluttered and visually complex backgrounds, and diverse worker appearances in loose or bulky protective clothing. A pose estimation model trained solely on laboratory-collected data would not have been exposed to such visual variability during training, creating a significant gap between the training data distribution and the conditions encountered during real-world deployment. This distribution gap can cause the model to produce inaccurate 2D keypoint detections when applied to real-world videos, which in turn degrades the accuracy of 3D pose estimates and downstream joint angle calculations for WMSD risk assessment.

In addition to challenges in the image-to-2D pose estimation stage, inconsistent camera viewpoints in the subsequent 2D-to-3D pose lifting stage introduce a separate source of error. Models trained on data captured from a fixed camera perspective may not generalize to the varying viewpoints encountered in field recordings, causing systematic bias in the estimated 3D poses. In particular, camera pitch angle, the vertical tilt of the camera relative to the horizontal plane, is common in real-world recordings, especially when the camera is positioned close to the target worker and angled downward to capture the full body. The pitch angle introduces systematic bias because the camera coordinate system, in which the 3D pose estimation model operates, no longer aligns with the world coordinate system. This misalignment particularly affects joint angles that are defined using the gravity direction, such as back flexion and lateral bending. These challenges need to be addressed in the pipeline design to ensure reliable joint angle estimation under field conditions.

This chapter addresses these challenges by developing an end-to-end computational pipeline that takes monocular video as input and estimates high-DoF joint angles required for WMSD risk assessment, with specific adaptations incorporated to account for real-world deployment conditions. Utilizing the large-scale dataset collected in the previous chapter as training data, the pipeline adopts a two-stage architecture (Figure 3.1), consisting of an image-to-2D pose estimation stage and a 2D-to-3D pose lifting stage, with adaptations at each stage detailed below. For the image-to-2D stage, transfer learning with backbone freezing is employed to leverage visual features learned from open-source general-purpose datasets, which contain hundreds of thousands of images spanning diverse individuals, appearances, clothing, and backgrounds. By freezing these pretrained features rather than training from scratch, the model retains the capacity to detect human keypoints across the visual variability encountered in field

environments. To further leverage the diversity of open-source general-purpose datasets and improve the model's generalization capability, the model is jointly trained using the merge training technique on our ergonomics-oriented dataset together with three open-source 2D human pose datasets that share overlapping keypoint definitions. This exposes the model to a substantially larger and more varied pool of real-world images during training, reducing its reliance on laboratory-collected data and improving its robustness to the visual conditions encountered in field deployment.

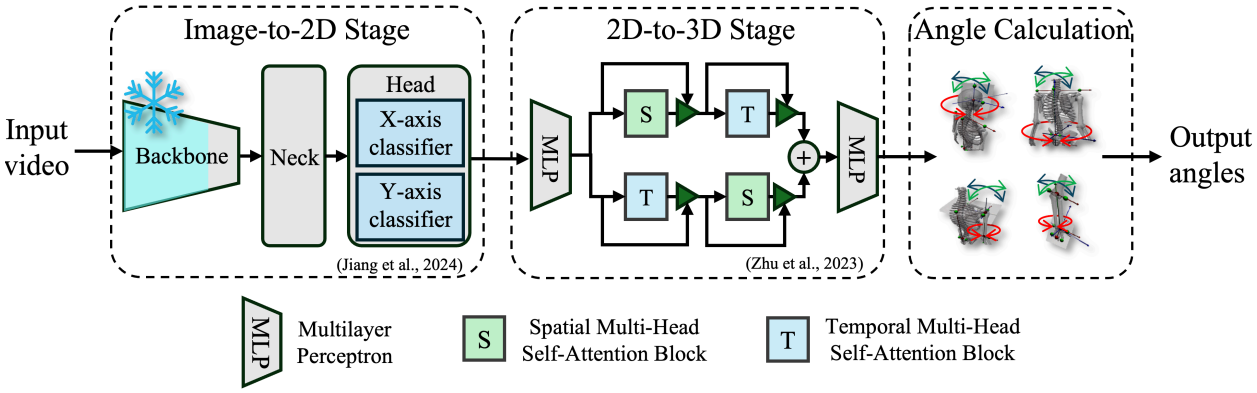


Figure 3.1. Computational pipeline for joint angle estimation from monocular video.

For the 2D-to-3D stage, in addition to the estimated 2D keypoint pixel locations, we also pass the 2D estimation confidence scores in as an additional input dimension to the 3D lifting model. This enables the model to learn to treat keypoints differently based on estimation quality. For example, the model can learn to rely less on a poorly estimated keypoint that is occluded and instead leverage surrounding higher-confidence keypoints to infer the occluded joint's 3D position. To address the systematic bias caused by variable camera pitch angles across different field recordings, we introduce a data-driven camera pitch angle correction method that ensures the estimated 3D pose is output with the correct gravity direction, regardless of the camera's pitch angle during recording. This enables accurate calculation of gravity-dependent joint angles such as back flexion and lateral bending.

The computational pipeline is evaluated both quantitatively on test set videos and qualitatively on 68 real-world footage captured from online repositories and industry partners to assess its practical applicability. Such a pipeline enables accurate joint angle estimation from regular videos captured by commonly available devices such as smartphones, making it practical for deployment on construction sites without specialized motion capture equipment.

3.2 Stage 1: Image-to-2D Model Training Adapted to Real-World Conditions

The image-to-2D pose estimation stage takes 2D RGB images as input and estimates the 2D pixel locations of the target keypoints. For this stage, we build our method upon RTMW (Jiang et al. 2024b), a 2D human pose estimation model with leading performance on the COCO-Wholebody benchmark (Jin et al. 2020) and real-time inference speed. RTMW features a specialized keypoint estimation head (Li et al. 2022), which formulates keypoint localization as a classification problem along the horizontal and vertical axes, rather than regressing the xy coordinates directly. This classification approach avoids the computationally expensive heatmap generation and post-processing required by regression-based methods, contributing to both inference speed and localization accuracy.

3.2.1 Transfer Learning and Backbone Freezing

As discussed in Section 3.1, a model trained solely on our laboratory-collected dataset would not generalize well to the diverse visual conditions encountered in real-world field environments. To address this distribution gap, we need to expose the model to a much wider range of visual conditions during training. Rather than collecting additional field data ourselves, we leverage the pretrained RTMW-x model, which was already trained on Cocktail14, a collection of 14 public datasets including AI Challenger, CrowdPose, MPII, sub-JHMDB, Halpe, PoseTrack18, COCO-Wholebody, UBody, Human-Art, WFLW, 300W, COFW, LaPa, and InterHand. These

datasets collectively contain hundreds of thousands of images spanning diverse individuals, appearances, clothing, and backgrounds. Through transfer learning, we initialize our model with these pretrained weights, allowing it to retain the robust visual feature representations already learned from this diverse data rather than learning from scratch on our laboratory data alone.

However, when fine-tuning the pretrained model on our ergonomics-oriented dataset to learn our newly defined keypoints, a trade-off arises in how much of the backbone to freeze. Freezing too many backbone layers preserves the pretrained generalization capability but limits the model's ability to learn visual features specific to our newly added keypoints. Freezing too few layers allows the model to fully adapt to the new keypoints but risks overfitting to the visual characteristics of our laboratory dataset, reducing the model's ability to generalize to real-world field conditions. To determine the optimal balance, we tested five configurations with different levels of backbone layer freezing and qualitatively evaluated each on real-world field videos, visually assessing the model's generalization to field conditions alongside its ability to detect the newly defined keypoints.

3.2.2 Merge Training with Open-Source 2D Human Pose Datasets

In addition to leveraging pretrained weights, we jointly train the model using the merge training technique, combining our ergonomics-oriented dataset with three open-source 2D human pose datasets: COCO-Wholebody (~200K images, 23 overlapping keypoints), AI Challenger (~300K images, 14 overlapping keypoints), and MPII-TRB (~25K images, 14 overlapping keypoints) (Figure 3.2). These datasets were selected for their overlapping keypoint definitions with our keypoint set and their large-scale coverage of diverse real-world imagery. During merge training, the model learns from the overlapping keypoints that are shared across datasets, while our newly added keypoints for joint angle calculation in WMSD risk assessment are only

supervised when training on our collected motion data. This allows the model to learn our newly added keypoints from the laboratory-collected data while simultaneously learning to generalize its visual feature extraction from the substantially larger and more visually diverse open-source datasets, further reducing the distribution gap between laboratory training conditions and real-world field deployment.

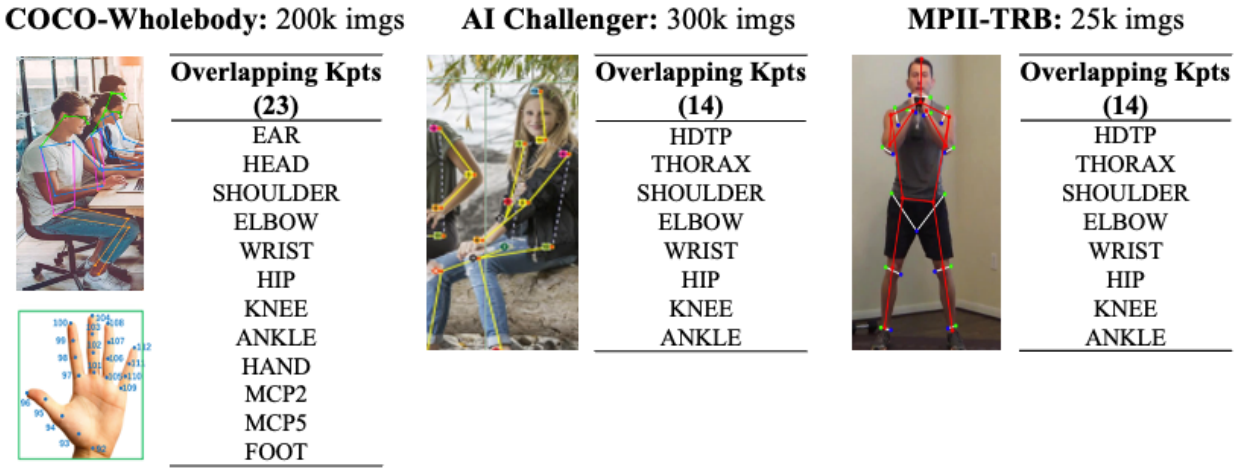


Figure 3.2. Overlapping keypoints between the three open-source 2D human pose datasets and the ergonomics-oriented keypoint set used in merge training.

For computational and training efficiency, we limited the keypoint output to only the keypoints directly used in our joint angle calculation steps, excluding keypoints such as heel and toe that are not currently used in angle calculations for the current WMSD risk assessment. The model was trained for 270 epochs on an 8-GPU H100 cluster on AWS, with a base learning rate of 5e-4, a training batch size of 112, and a validation batch size of 32.

3.3 Stage 2: 2D-to-3D Model Training Adapted to Real-World Conditions

The 2D-to-3D pose lifting stage takes the estimated 2D keypoint positions from the image-to-2D stage as input and estimates their corresponding 3D positions. For this stage, we adopt MotionBERT (Zhu et al. 2023) as detailed in Chapter 2. In addition to the estimated 2D keypoint pixel locations, we pass the 2D estimation confidence scores as an additional input dimension to

the 3D model. These confidence scores serve as an approximation of estimation quality at each keypoint, providing the model with information about which keypoints are reliably detected and which may be degraded by occlusion or other real-world factors. By incorporating this information, the 3D lifting model can learn to rely less on poorly estimated keypoints and instead leverage surrounding higher-confidence keypoints to infer occluded joints, improving the robustness of 3D pose estimation under challenging field conditions.

3.3.1 Camera Pitch Correction

As discussed in Section 3.1, camera pitch angle is common in field recordings and causes the camera coordinate system to misalign with the world coordinate system, introducing systematic bias in gravity-dependent joint angles such as back flexion and lateral bending. Since the pipeline takes only monocular video as input, without additional information about the camera's orientation in the world coordinate system (e.g., IMU data), we adopt a data-driven approach that enables the model to learn the correction directly. During training, we rotate the 3D ground truth poses so that the gravity direction is always aligned to the downward direction (+y axis) in the estimation camera coordinate system, regardless of the camera pitch angle at which the data was captured. Combined with the diverse camera viewpoints intentionally included in our dataset (Chapter 2), covering common viewpoints from handheld recording to elevated surveillance-style angles, the model learns to map 2D poses captured from varying pitch angles to 3D poses with the correct gravity orientation. At inference time, the model directly outputs 3D poses aligned with the world coordinate system without requiring knowledge of the camera pitch angle.

3.3.2 Angle Loss Function Design

To further align the model's optimization with the downstream task of estimating the 22 high-DoF joint angles used in WMSD risk assessment, an angle loss function is incorporated into

the training objective alongside the conventional MPJPE loss proposed in the original MotionBERT architecture. Since each joint angle is derived from the relative positions of multiple keypoints, small positional errors at individual keypoints can compound when computing the angle between body segments, resulting in larger errors in the derived joint angles. By including joint angle accuracy as an explicit training objective, the model directly optimizes for the quantities used in WMSD risk assessment rather than relying on positional accuracy alone.

The angle loss consists of two terms: an angle magnitude term ($L_{magnitude}$) and an angular velocity term ($L_{velocity}$). The angle magnitude term penalizes the difference between predicted and ground truth joint angles at each frame, directly supervising the accuracy of the estimated angles. The angular velocity term penalizes the difference in frame-to-frame angle changes between prediction and ground truth, encouraging temporally smooth angle estimates across consecutive frames. Notably, to avoid discontinuity errors at angular wrapping boundaries (e.g., the transition between 359° and 1°), both terms use a cosine similarity loss computed on unit vector representations of the angles, which naturally handles wrapping since angles such as 0° and 360° map to the same unit vector.

$$L_{angle} = \lambda_1 L_{\text{magnitude}} + \lambda_2 L_{velocity}$$

$$\mathrm{L_{magnitude}} = \frac{1}{\mathrm{F} \times \mathrm{A}} \sum_{\mathrm{f}=1}^{\mathrm{F}} \sum_{\mathrm{a}=1}^{\mathrm{A}} \left(1 - \cos \theta_{\mathrm{pred}}^{\mathrm{f,a}} \cos \theta_{\mathrm{gt}}^{\mathrm{f,a}} - \sin \theta_{\mathrm{pred}}^{\mathrm{f,a}} \sin \theta_{\mathrm{gt}}^{\mathrm{f,a}}\right)$$

$$L_{velocity} = \frac{1}{(F-1) \times A} \sum_{f=2}^{F} \sum_{a=1}^{A} \left(1 - \cos \theta_{pred}^{\dot{f,a}} \cos \theta_{gt}^{\dot{f,a}} - \sin \theta_{pred}^{\dot{f,a}} \sin \theta_{gt}^{\dot{f,a}}\right)$$

Where F is the number of frames, $A$is the number of joint angles (22 angles used in WMSD risk assessment), $\theta_{pred}^{f,a}$ and $\theta_{gt}^{f,a}$ are the predicted and ground truth values of the a-th joint angle

at frame f, $\dot{\theta}^{f,a} = \theta^{f,a} - \theta^{f-1,a}$ is the angular velocity, and λ_1 and λ_2 are the weighting coefficients.

3.4 Evaluation of the End-To-End Computational Pipeline

To evaluate the computational pipeline, we first conducted a quantitative evaluation on laboratory test set videos where ground truth human motion data was available. We assessed the accuracy of both the intermediate outputs from the image-to-2D stage and the 2D-to-3D stage, as well as the final joint angle estimates. We then conducted a qualitative evaluation on 68 real-world field videos sourced from online repositories and industry partners, where ground truth was unavailable, to assess the pipeline's practical applicability under real-world deployment conditions that differ from the controlled laboratory setting used for training and quantitative evaluation.

3.4.1 Quantitative Evaluation of 2D Pose Estimation on Test Set Videos

We evaluated the image-to-2D stage using mean error (in pixels) and Percentage of Correct Keypoints (PCK) (Table 3.1). Mean error measures the average Euclidean distance in pixels between estimated and ground truth keypoint locations. While this metric is sensitive to image resolution, all test set videos in our evaluation were captured at 1920x1080 resolution, providing a consistent and intuitive measure of accuracy. Across the test set, the 2D model achieved an average mean error of 9.4 pixels, meaning that estimated keypoints were on average fewer than 10 pixels from the ground truth on a 1920-pixel-wide frame.

Table 3.1. Mean error (px) and PCK of 2D pose estimation (keypoint abbreviations in Table 2.3).

Category	Keypoint	Mean Error (px)	PCK@0.2	PCK@0.05
Overlapping Keypoints	Pelvis	15.1	99.6%	57.4%
	Wrist	9.7	90.0%	46.3%
	Hip	12.4	99.6%	86.2%
	Knee	9.4	99.3%	96.9%
	Ankle	6.5	99.3%	98.1%
	Foot	9.7	99.1%	96.0%
	Hand	11.5	86.6%	39.7%
	Elbow	7.6	98.5%	90.1%
	Shoulder	8.7	99.6%	89.2%
	Head	6.8	99.4%	83.6%
	HDTP	9.9	97.9%	61.0%
	EAR	7.4	99.2%	73.5%
	C7	8.1	99.5%	98.8%
	MCP2	12.0	85.2%	36.5%
	MCP5	12.3	84.3%	33.1%
	Average	9.8	95.8%	72.4%
Unique Keypoints	LE	8.0	98.4%	88.5%
	ME	8.5	98.5%	87.6%
	Thorax	8.1	99.6%	97.0%
	SS	8.2	99.6%	96.8%
	C7_d	7.7	99.6%	97.5%
	AP_b	8.3	99.5%	88.0%
	AP_f	10.5	99.6%	73.6%
	Average	**8.6**	**99.6%**	**90.6%**
Overall Average		**9.4**	**96.4%**	**76.7%**

We also reported PCK, which measures the percentage of estimated keypoints that fall within a given threshold distance from the ground truth. To enable meaningful comparison across keypoints located on body regions of different scales, as keypoints on smaller regions such as the hand should have a tighter error tolerance than those on larger regions such as the torso, we grouped keypoints into body regions (e.g., head, torso, left arm, right hand) and normalized the PCK threshold by the diagonal length of each body region's bounding box. This allows accuracy to be interpreted relative to the spatial scale of each body region. The 2D model achieved an average PCK@0.2 of 96.4%, meaning that 96.4% of estimated keypoints fall within 20% of their respective body region's bounding box diagonal from the ground truth.

We further reported the 2D pose accuracy separated by overlapping keypoints (i.e., shared with the merge training datasets) and unique keypoints (present only in our ergonomics-oriented dataset). Overlapping keypoints achieved an average mean pixel error of 9.8 pixels with 95.8% PCK@0.2, while unique keypoints achieved 8.6 pixels with 99.6% PCK@0.2. The comparable or stronger 2D pose estimation performance on unique keypoints suggests that, despite these keypoints being supervised only by our ergonomics-oriented dataset during training, the large scale of our collected dataset provided sufficient data to train the parameters in the 2D model for estimating these newly added keypoints under laboratory visual conditions.

3.4.2 Quantitative Evaluation of 3D Pose Estimation on Test Set Videos

We evaluated the 2D-to-3D stage using two standard metrics: root-aligned Mean Per Joint Position Error (MPJPE) and Procrustes-Aligned MPJPE (PA-MPJPE). MPJPE measures the average Euclidean distance in millimeters between estimated and ground truth 3D keypoint positions after aligning the root joint (pelvis). PA-MPJPE further applies a rigid alignment (rotation, translation, and scaling) between the estimated and ground truth poses before computing the error, isolating the error in pose shape from errors in global orientation and scale.

The model achieved an overall average MPJPE of 16.7 mm and PA-MPJPE of 10.8 mm (Table 3.2), indicating that the estimated 3D keypoint positions were on average within approximately 1.7 cm of the ground truth, and within approximately 1.1 cm after removing global alignment differences. Consistent with the 2D evaluation, unique keypoints achieved comparable performance to overlapping keypoints, with an average MPJPE of 11.8 mm and PA-MPJPE of 8.6 mm, further supporting that the dataset provided sufficient data to train the 3D lifting model for estimating these newly added keypoints under laboratory conditions.

Table 3.2. MPJPE and PA-MPJPE of 3D pose estimation (keypoint abbreviations in Table 2.3).

Category	Keypoint	MPJPE (mm)	PA-MPJPE (mm)
Overlapping Keypoints	Pelvis	-	11.1
	Wrist	16.5	12.1
	Hip	14.6	13.8
	Knee	12.8	10.8
	Ankle	13.9	11.1
	Foot	16.0	13.2
	Hand	18.2	13.7
	Elbow	13.4	9.3
	Shoulder	11.2	8.0
	Head	10.4	7.0
	HDTP	13.4	10.5
	EAR	10.8	7.7
	C7	12.3	9.0
	MCP2	19.0	14.6
	MCP5	18.8	14.4
	Average	**14.4**	**11.1**
Unique Keypoints	LE	13.5	9.3
	ME	14.1	10.3
	Thorax	10.0	7.1
	SS	10.1	7.2
	C7_d	10.5	8.2
	AP_b	11.7	8.6
	AP_f	12.5	9.3
	Average	**11.8**	**8.6**
Overall Average		**16.7**	**10.8**

3.4.3 Quantitative Evaluation of Joint Angle Estimation on Test Set Videos

We evaluated the joint angle estimation accuracy of the end-to-end computational pipeline by comparing the angles calculated from the pipeline's estimated 3D poses with the ground truth angles derived from the marker-based motion capture system. Following the same evaluation protocol as Chapter 2, we reported angle MAE, median absolute error, mean, SD, skewness, kurtosis, and IQR for each of the 22 high-DoF joint angles used in WMSD risk assessment in Table 3.3.

Averaging across all joint angles, our estimated angles achieved an MAE of 5.93º and a median absolute error of 3.53º. While previous studies developing similar monocular vision-based WMSD risk assessment methods rarely reported their angle estimation accuracy due to the lack of ground truth 3D joint angles (Jeong and Kook 2023; Kwon et al. 2022; Paudel et al. 2022), Chu et al. (2020) reported an angle MAE of 11.7º when evaluating each human body segment's (i.e., clavicle, upper arm, lower arm, hand, upper leg, lower leg, and foot) 3D orientation relative to the world coordinate frame. Fan et al. (2024) reported an MAE of 9.6º for neck, trunk, leg, upper arm, lower arm, and wrist angles. Though not a strict comparison due to the differences in angle definitions and evaluation datasets, we achieved a notable improvement in joint angle estimation accuracy, contributing to more reliable results in subsequent WMSD risk assessments.

Table 3.3. Joint angle estimation errors (in degrees)

	Angle Name	MAE	Median Abs. Err.	RMSE	Mean	SD	Skewness	Kurtosis	IQR
Neck	Flexion/extension	4.71	3.34	7.40	1.30	7.29	-1.96	141.36	6.67
	Lateral flexion	3.10	2.28	5.48	0.64	5.44	-12.56	453.00	4.50
	Rotation	4.35	3.52	5.73	0.22	5.73	-0.64	10.20	7.02
Shoulder	Flexion/extension	5.23	3.38	8.77	-0.57	8.75	0.51	47.23	6.74
	Abduction/adduction	4.59	2.56	8.51	0.15	8.51	0.00	46.89	5.12
	Rotation	7.77	5.96	10.59	3.51	9.99	-0.32	10.07	11.03
Elbow	Flexion	4.45	3.04	6.71	-0.57	6.69	-0.20	14.40	5.95
Wrist	Flexion/extension	7.88	5.69	11.11	-0.76	11.08	-0.48	5.58	11.43
	Radial/ulnar deviation	6.37	4.50	9.00	-0.32	9.00	-0.03	4.63	8.99
	Pronation/supination	13.89	9.76	20.68	-1.77	20.61	0.45	11.77	19.14
Back	Flexion/extension	2.32	1.61	4.74	-0.41	4.72	17.02	798.77	3.38
	Lateral flexion	5.41	1.82	16.48	0.23	16.48	2.59	57.08	3.67
	Rotation	3.52	2.75	4.66	-0.26	4.65	-0.12	3.22	5.50
Knee	Flexion	3.39	2.47	5.06	0.51	5.03	2.42	50.10	4.98
	Average	**5.93**	**3.53**	**10.36**	**0.10**	**10.36**	**0.57**	**47.55**	**7.06**

We also provided histograms of the angle estimation errors for each joint angle in Figure 3.3, which offer an intuitive visualization of the estimation accuracy and allow direct comparison of performance across different joint angles. Comparing the joint angles, back angles showed the most concentrated error distributions around zero, which is expected since the keypoints used to calculate back angles (e.g., shoulders, hips) are located far apart on the body, meaning small errors in 3D keypoint positions have a relatively minor effect on the derived angles. Neck, shoulder, elbow, and knee angles showed moderately concentrated distributions. In contrast, wrist angles showed the most spread-out distributions, as the keypoints used for wrist angle calculation are closely spaced, fast-moving, and frequently occluded during manual tasks, making the derived angles more sensitive to small keypoint position errors.

Among the high-DoF shoulder and wrist joints, rotation angles exhibited more spread-out error distributions compared to their corresponding flexion and abduction components. This is consistent with Table 3.3, where wrist pronation/supination showed the highest MAE at 13.89°. It is worth noting, however, that not all joint angles are equally perceptible. Rotational angles such as wrist pronation/supination and shoulder rotation involve subtle movements that are inherently difficult to distinguish even for human observers, unlike larger and more visually apparent angles such as back flexion or knee flexion. This suggests that the higher estimation errors for rotation angles may partially reflect the inherent difficulty of capturing these subtle movements from monocular video, rather than a fundamental limitation of the pipeline.

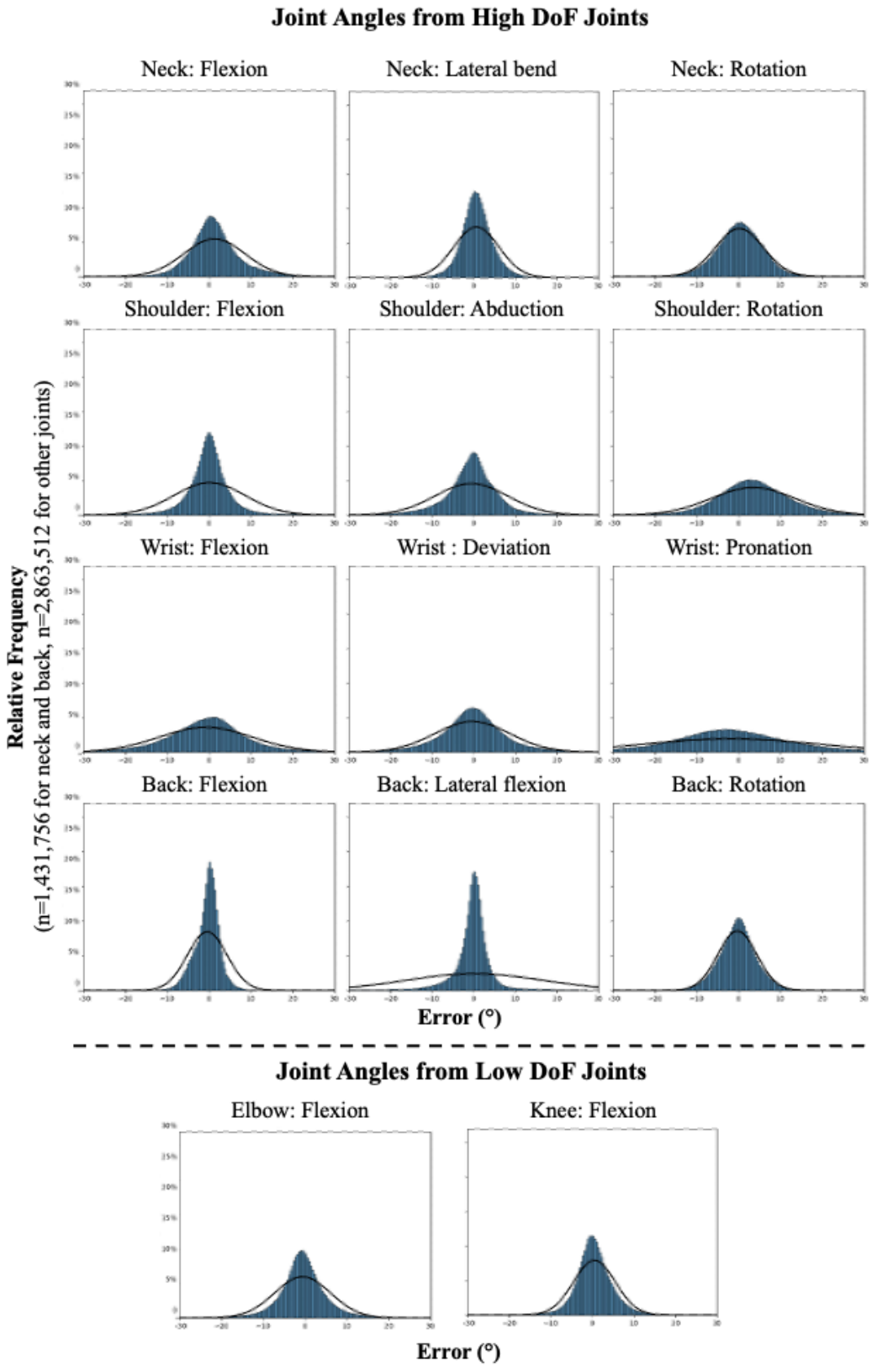


Figure 3.3. Histograms of angle estimation errors with RGB video input, overlaid with the normal distribution curves computed from the corresponding means and standard deviations.

3.4.4 Qualitative Evaluation on Real-World Videos

To evaluate the trained model's applicability in real-world settings, we applied it to 68 industry videos featuring workers performing MMH tasks such as installing drywall, lifting concrete blocks, and lifting lumber. These videos present additional challenges compared to our lab-collected test set, including occlusion, varying camera viewpoints, inconsistent lighting, and multi-person scenarios. Figure 3.4 and Figure 3.5. visualize representative frames from the industry videos with the estimated 3D pose results, demonstrating the model's capability to generalize beyond the lab environment.

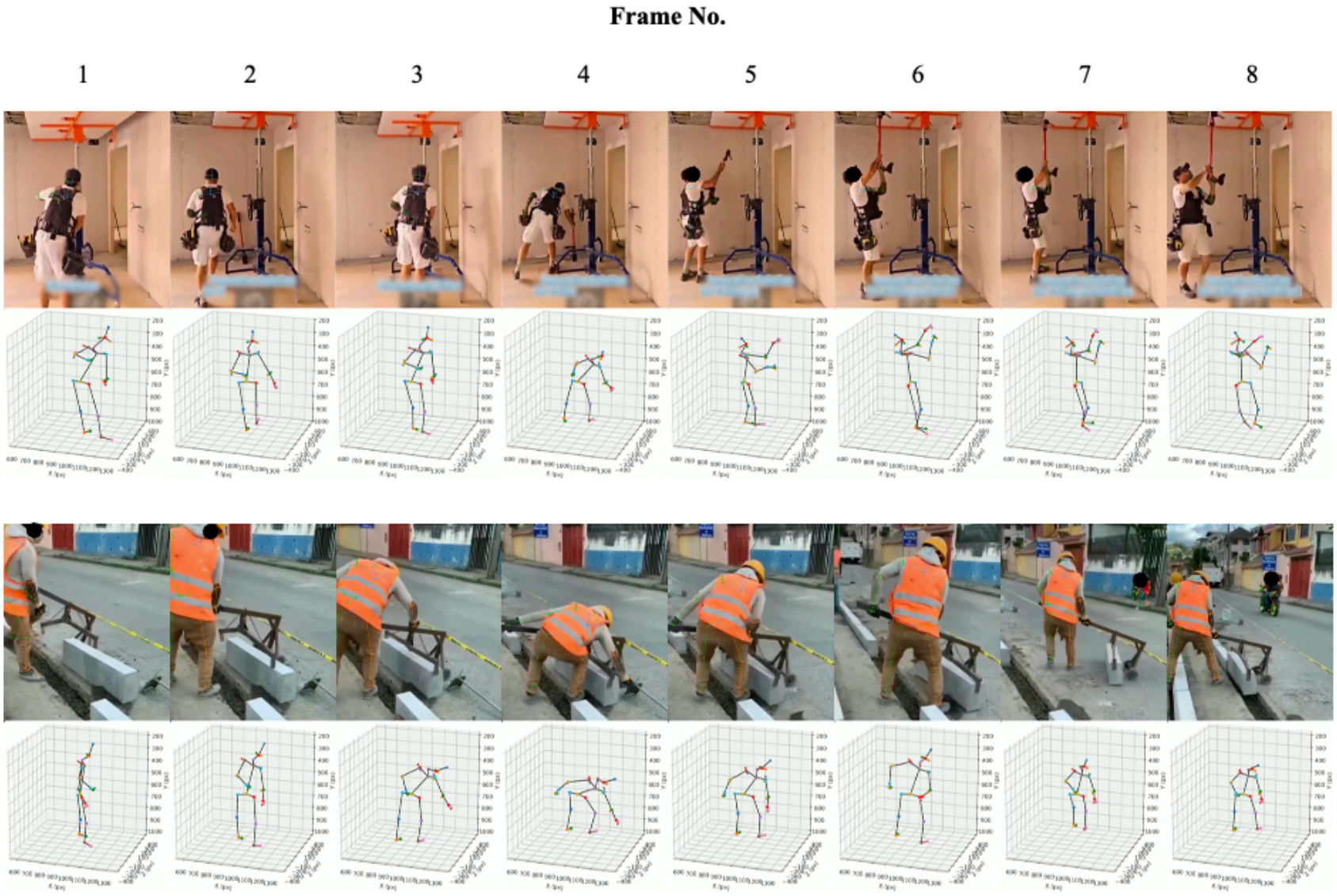


Figure 3.4. Example frames from industry videos with corresponding estimated 3D human poses (part 1).

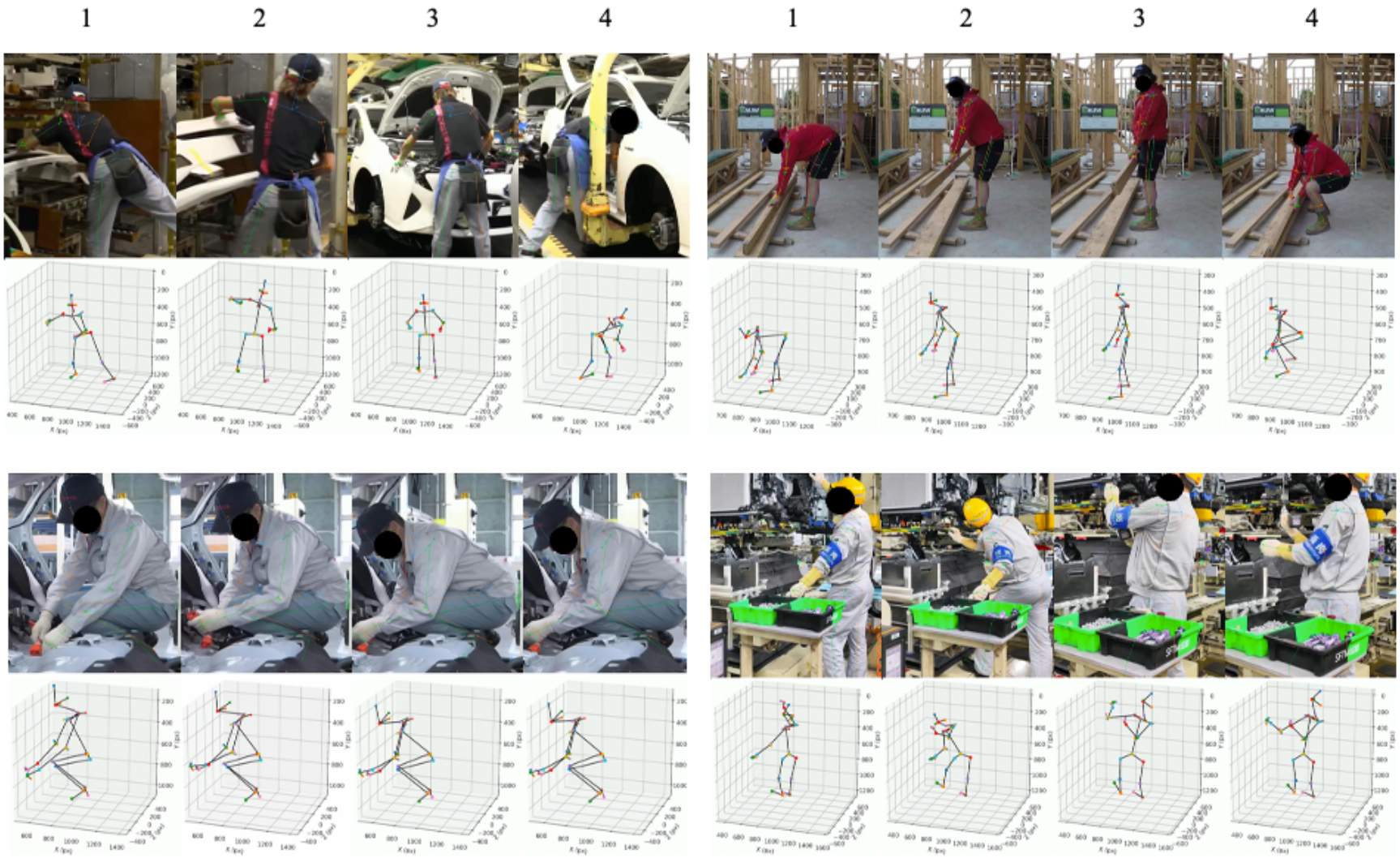


Figure 3.5. Example frames from industry videos with corresponding estimated 3D human poses (part 2).

3.5 Discussion

To interpret joint angle accuracy in the context of WMSD risk assessment, it is important to consider the estimation errors relative to the angle ranges commonly used in observation-based assessment tools. For instance, the neck angle range corresponding to a segment score of 1 in RULA spans 10 degrees, while the leg angle range in REBA spans 30 degrees. These segment scores are further aggregated into a small number of coarse risk levels (e.g., five risk levels in REBA). Given that our model achieves an overall MAE of 5.93º (median absolute error of 3.53º), which is well below the angle range thresholds, such errors would have a minimal impact on the segment score or the final risk level.

While our trained model achieved an overall accurate angle estimation, some variations in angle accuracy can be observed depending on the joint angle. This variation is presented in the MAE results (Table 3.3) and visualized in the error histograms (Figure 3.3), where certain joints display a wider error distribution with a lower peak frequency. Most joints, such as the back (MAE = 3.75°) and knee (MAE = 3.39°), exhibit low angle errors. However, as shown in Figure 3.3, the wrist joint angles displayed a wider error distribution, with an MAE of 9.38°. One potential cause of this accuracy variation is the close proximity of the keypoints used to define the angles (e.g., MCP2, MCP5 on the hand), which amplifies small 3D keypoint inaccuracies from vision-based pose estimation. Despite this accuracy variation, our wrist joint angle estimation accuracy (MAE 9.38°) remains within the angle bin ranges used in common WMSD risk assessment methods (e.g., 30-degree bins for wrist flexion in RULA and REBA), and still outperforms previously reported hand segment angle accuracy by Chu et al. (2020) (MAE = 12.63°) and Fan et al. (2024) (MAE = 15.62°).

Beyond the laboratory evaluation, the qualitative assessment on 68 real-world videos demonstrated the pipeline's ability to produce visually plausible 3D pose estimates under field conditions not represented in the training data. This evaluation serves as an essential step in assessing the pipeline's readiness for real-world deployment, where visual conditions differ substantially from the controlled laboratory environment. The visually plausible results observed across diverse field footage suggest that the adaptations introduced in this chapter may have contributed to the pipeline's generalization capability. Future work can further validate these findings through quantitative evaluation on real-world data, such as comparisons against assessments from certified professional ergonomists.

3.6 Conclusion

This chapter presented an end-to-end computational pipeline for estimating high-DoF joint angles from monocular video for WMSD risk assessment, with specific adaptations to address the challenges of real-world deployment conditions. To bridge the distribution gap between laboratory training data and field deployment, the pipeline leveraged diverse open-source datasets through transfer learning and merge training, and introduced a data-driven camera pitch angle correction method to account for the varying visual conditions and camera viewpoints encountered in field recordings. Quantitative evaluation on laboratory test set videos demonstrated that the pipeline achieved an overall joint angle MAE of 5.93° and median absolute error of 3.53° across 22 high-DoF joint angles. Qualitative evaluation on 68 real-world videos sourced from online repositories and industry partners demonstrated the pipeline's ability to produce visually plausible estimates under field conditions not represented in the training data.

Chapter 4 Intervention Recommendation Applying Scene-Aware Motion Synthesis within Physics-Based Simulations

4.1 Introduction

The previous two chapters focused on vision-based WMSD risk identification, presenting a large-scale human motion dataset and a vision-based computational pipeline for estimating high-DoF joint angles from monocular video needed in WMSD risk assessment. However, identifying risks alone is insufficient; the subsequent step, designing and implementing effective ergonomics interventions to mitigate those risks, is equally important. Unlike risk identification, which has increasingly been automated through vision-based methods, intervention generation has received comparatively limited research attention or industry adoption. As a result, ergonomics intervention design remains a largely manual process that demands substantial resources, including domain expertise, time, and cost, making it particularly challenging for the construction industry given its diverse tasks (Choi 2012), dynamic site conditions (Choi et al. 2016), and high labor turnover (BLS 2025a; b).

Common ergonomics interventions can include helping workers adopt better postures for their specific tasks or redesigning tasks and workstations to reduce biomechanical demand (Choi 2012; Choi et al. 2016). In either case, effectively communicating the intended changes to workers is important. Research on motor imagery, the process of mentally rehearsing a movement, has shown that seeing correct task execution promotes motor learning in physical tasks, particularly when the visual stimuli match the task context (Guillot et al. 2024; Ladda et al. 2021). This

suggests that visualizing these interventions, whether demonstrating safer postures or showing how task and workstation modifications change the resulting motion, can help workers better understand and adopt the intended interventions.

Scene-aware motion synthesis methods, which can synthesize and visualize physically plausible human motions under specified task and site constraints within physics-based simulations (Pan et al. 2025), offer a promising way to provide such visual interventions without needing the substantial resources currently required for manual ergonomics intervention design. The synthesized motions can serve as visual posture demonstrations, showing workers the correct techniques to perform their task under their site constraints. Additionally, it can also enable automated "what-if" analysis of task and workstation modifications, enabling site personnel, such as workers and managers, to design changes toward safer work configurations.

For synthesized motions to be a viable intervention, they must be safe to perform and align with real-world tasks and site constraints where the WMSD risk is identified. However, current scene-aware motion synthesis methods, primarily designed for general applications such as animation, often produce motions with limited consideration of WMSD risks and require substantial effort to input real-world task and site constraints. We propose a scene-aware motion synthesis method for ergonomics interventions that improves motion safety through a custom ergonomics reward. We also developed a computational framework that captures and processes real-world task and site constraints, which serve as key inputs to our scene-aware motion synthesis method, using common consumer devices such as lidar-enabled smartphones. Our proposed method and computational framework enable the synthesis of risk-reduced motion examples that align with real-world task and site constraints that function as ergonomics interventions,

demonstrating the potential to provide site personnel with the means to improve worker postures in high-risk scenarios.

4.2 Literature Review On Current Scene-Aware Motion Synthesis Methods

Recent work on scene-aware motion synthesis can be divided into two categories: data-driven kinematic generation and physics-based character control (Pan et al. 2024). A large number of recent studies have explored data-driven kinematic generation methods, which learn from large-scale motion datasets annotated with natural language descriptions and 3D scene scans (Jiang et al. 2024a; Wang et al. 2024; Zhang and Tang 2022). These methods can generate visually realistic motions conditioned on text prompts and 3D scenes using architectures such as Variational Autoencoders (Zhang and Tang 2022) and diffusion models (Jiang et al. 2024a; Wang et al. 2024). However, a common limitation of data-driven kinematic generation methods is the lack of explicit enforcement of physical constraints, which can result in physically implausible motions such as collisions with the 3D scene (Cen et al. 2024; Wang et al. 2024), unrealistic object manipulation (Peng et al. 2024), and unstable postures (Wang et al. 2024). Moreover, their performance is highly dependent on the task and scene distribution of the training data (Peng et al. 2021), which predominantly consists of daily activities in household or office environments (Araujo et al. 2023; Bhatnagar et al. 2022b). Consequently, they have limited ability to generalize to construction tasks, where domain-specific terminology (e.g., “drilling,” “screeding concrete”) and site-specific conditions are absent from common large-scale motion datasets used in the training process. In our prior work, we experimented with adapting a text-conditioned motion generation method by incorporating a stochastic search process to reduce the WMSD risk; however, the absence of construction tasks in the training dataset resulted in unreliable performance when conditioned on construction task descriptions (Wen et al. 2024).

Physics-based character control methods address some of these limitations by synthesizing motion through physics-based simulations. Instead of learning solely by imitating motions in large-scale datasets, these methods also learn from first principles of physics (Peng et al. 2021). Earlier work focused on designing and optimizing reward functions based on heuristics derived from human motions, using optimization techniques such as trajectory optimization and reinforcement learning (Clegg et al. 2018; Yu et al. 2018). However, formulating effective rewards that consistently produce natural human motions has proven to be highly challenging (Peng et al. 2021). To address this, recent methods have incorporated imitation learning methods, such as Adversarial Motion Priors (AMP), to learn natural human motions from examples in motion datasets (Pan et al. 2025; Peng et al. 2021, 2022). By combining reinforcement learning with motion imitation, these methods require fewer example motions than purely data-driven kinematic generation methods while synthesizing motions that are bounded by the physical task and site constraints.

While physics-based character control offers a path toward realistic motions that conform to task and site constraints, two critical gaps remain for its use in ergonomic interventions. The primary gap is that existing methods lack mechanisms to reduce WMSD risks in the synthesized motions, a fundamental requirement for interventions. Additionally, there is no established framework for capturing and processing real-world task and site constraints into usable inputs for scene-aware motion synthesis methods. Addressing these gaps is essential to enable scene-aware motion synthesis as a potential intervention recommendation tool that helps workers improve their postures.

4.3 Synthesizing Risk-reduced Motions with Custom Ergonomics Rewards

To synthesize risk-reduced motions in scene-aware motion synthesis, we extend TokenHSI (Pan et al. 2025)—a physics-based character control method that has demonstrated strong

performance in synthesizing realistic human motions for MMH tasks using reinforcement learning—by incorporating a custom ergonomics reward $r_t^{ergo.}$ that consists of three components: (1) a guideline reward $r_t^{guildline}$, (2) a reference reward $r_t^{ref.}$, and (3) a knowledge reward $r_t^{knowledge}$.

$$r_t^{ergo.} = r_t^{guildline} + r_t^{ref.} + r_t^{knowledge}$$

The guideline reward forms the foundation of our ergonomics reward by encoding established ergonomics guidelines from standards such as NIOSH, REBA, and RULA into mathematical reward functions. These guidelines address well-defined aspects of safe MMH techniques, such as maintaining back alignment or minimizing load distance from the body, that can be readily expressed as mathematical functions of the humanoid's posture and the task state. However, correct MMH technique encompasses more nuanced, task-specific details that cannot be fully captured through mathematical formulations of general guidelines alone. For this reason,

To address this limitation, we added an additional reference reward that leverages Adversarial Motion Priors (AMP) (Peng et al. 2021) to learn more intricate, task-specific technique details from curated reference motions of correct MMH performance. These reference motions are obtained from videos of correct MMH techniques through a monocular motion capture framework, which extracts 3D human motion from ordinary videos and prepares it for imitation learning in the physics-based simulation, making reference motions accessible to collect without specialized motion capture equipment. By imitating these reference motions, the policy acquires aspects of safe technique that are difficult to specify through explicit reward functions. Yet curating a sufficiently large and diverse set of reference motions for reinforcement learning policy training is resource-intensive, and with a reasonably sized reference motion dataset (e.g., 3-10 motion clips),

the policy often exhibits undesirable behaviors in certain scenarios (e.g., not using handles when lifting containers with handles).

To address these remaining challenges, we developed a framework that systematically captures the tacit knowledge of ergonomics experts regarding correct techniques and integrates it into a knowledge reward function. While such challenges are typically addressed through iterative reward engineering (Araki et al. 2025), which demands significant manual effort and expertise in both reinforcement learning and ergonomics, our framework instead allows ergonomics experts to periodically observe the synthesized motions during training and provide natural language feedback on observed mistakes as well as the desired behaviors. This feedback is then encoded into explicit reward functions using a large language model (LLM), effectively translating tacit expert knowledge into a knowledge reward to encourage correct technique behaviors.

4.3.1 Learning from Published Ergonomics Guidelines

Existing ergonomics guidelines and WMSD risk evaluation methods contain quantifiable rules about human postures that can inform whether a posture is safe. For example, in the NIOSH Revised Lifting Equation, the horizontal distance from the load to the worker's body is used to calculate the Horizontal Multiplier (Waters et al. 2021). In REBA, the back flexion angle is used to calculate the Trunk score (Hignett and McAtamney 2000). We fully leverage this rich information to build the guideline reward.

Our guideline reward function $r_t^{guideline}$ consists of three components: (a) a back reward r_t^{back} that encourages maintaining a straight back, (b) a box reward r_t^{box} that promotes keeping the load close to the body, and (c) an elbow reward r_t^{elbow} that encourages keeping the elbows within the desired range of motion.

$$r_t^{guideline} = 0.5 * r_t^{back} + 0.25 * r_t^{elbow} + 0.25 * r_t^{box}$$

The back reward function yields the maximum reward (i.e., 1.0) when the trunk remains straight, with flexion within a range of ±20°. This threshold is determined based on the trunk score angle threshold defined in the REBA (Hignett and McAtamney 2000). A modification to the reward function is applied when the knees are deeply bent in a power-lift posture (Loyola University Chicago 2024). In such cases, moderate back flexion is tolerated to allow the load to be maintained closer to the body.

$$r_t^{back} = \begin{cases} 1.0 & \left|\theta_t^{back}\right| < \frac{1}{9}\pi \\ \exp\left(-5 * \left|\theta_t^{back} - \frac{1}{9}\pi\right| * \min\left(1, \frac{2\pi - 2\theta_t^{knee}}{\pi}\right)\right) & o.w. \end{cases}$$

Where θ_t^{back} is the back flexion angle at time t, and θ_t^{knee} is the minimum knee flexion angle.

The box reward yields its maximum value when the horizontal distance between the load and the center of the feet is minimized, following the Horizontal Multiplier in the NIOSH Lifting Equation (Waters et al. 2021).

$$r_t^{box} = \begin{cases} 0.0 & \left\|x_t^{obj} - x_t^{start}\right\| < 0.1 \\ exp\left(-5 * min\left(0, \left\|x_t^{obj_{xy}} - x_t^{feet_{xy}}\right\| - \frac{w_{obj}}{2}\right)\right) & o.w. \\ 1.0 & \left\|x_t^{obj} - x_t^{goal}\right\| < 0.1 \end{cases}$$

Where $x_t^{obj_{xy}}$ is the object center's position projected onto the horizontal plane at time t, $x_t^{feet_{xy}}$ is the feet center's position projected onto the horizontal plane, w_{obj} is the width of the object, x_t^{obj} is the 3D location of the object, x_t^{start} is the 3D object start location for the MMH task, $x_t^{goal}$is the object goal location for the MMH task.

The elbow reward yields its maximum value when the elbow flexion angle is maintained within 60°–100°, consistent with the lower-arm angle thresholds defined in REBA (Hignett and McAtamney 2000).

$$r_t^{elbow} = \begin{cases} 0.0 & \left\| x_t^{obj} - x_t^{start} \right\| < 0.1 \\ \sum_{s \in \{l,r\}} 0.5 * \exp\left(-5 * \max\left(0, \left|\theta_t^{elbow,s} - \frac{4}{9}\pi\right| - \frac{1}{9}\pi\right)\right) & o.w. \\ 1.0 & \left\| x_t^{obj} - x_t^{goal} \right\| < 0.1 \end{cases}$$

Where $\theta_t^{elbow,s}$ is the elbow flexion angle on the body side s at time t.

4.3.2 Learning from Reference Motions via Imitation Learning

To capture aspects of correct MMH techniques that cannot be easily encoded in reward functions, such as avoiding jerking by using smooth motions or using the legs rather than the back to lift (NIOSH 2007), we incorporate imitation learning to further reduce the WMSD risk of the synthesized motions. This requires reference motions demonstrating the correct techniques for each task category.

We developed a pipeline for capturing reference motions suitable for imitation learning (Figure 4.1). The pipeline uses monocular motion capture to extract 3D human motion from reference videos demonstrating the correct MMH techniques, which is then retargeted to a humanoid model in Isaac Gym (Makoviychuk et al. 2021), a suitable format for imitation learning. For each task category, we captured approximately 2 minutes of reference video (18-21 clips) using a smartphone, demonstrating correct MMH techniques under varying task parameters (e.g., different start heights, different object weights)

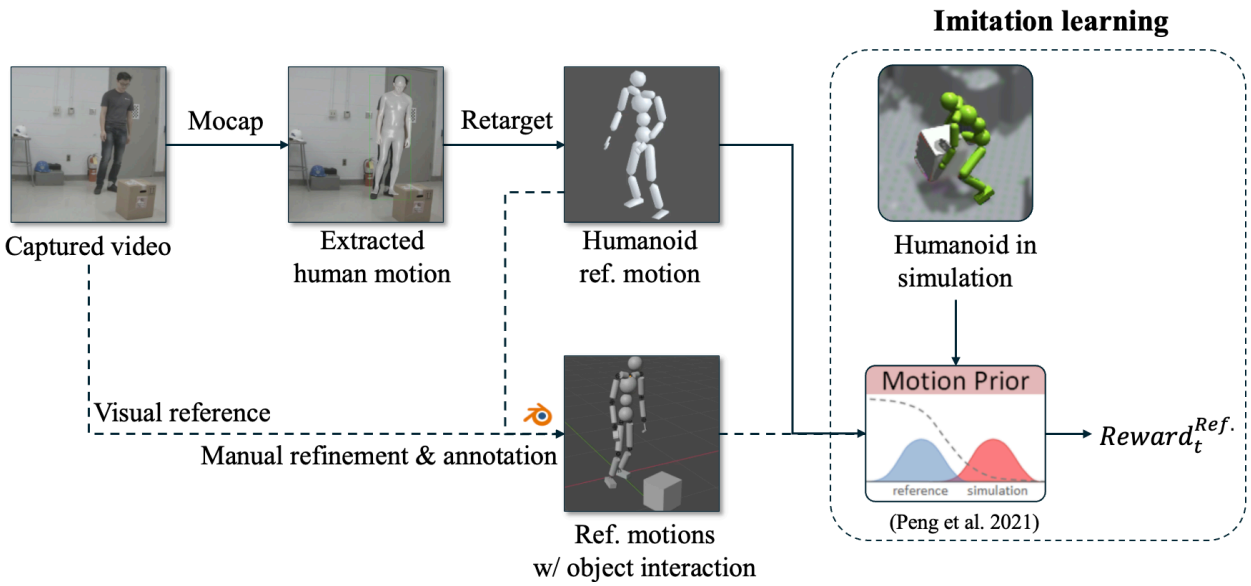


Figure 4.1. Reference motion capture pipeline for imitation learning.

To further assist the reinforcement learning policy training process, for a subset of these motions (2-3 clips per task category), we annotate the human-object interactions in Blender (Blender Foundation 2026), manually specifying the physical movement of interaction objects in space by referencing the captured RGB videos. We also manually refine the captured motions to resolve spatial conflicts such as hand-object penetrations and limb collisions. These annotated human-object interactions contain paired humanoid and object states, allowing the simulation environment to be initialized at realistic intermediate states during training, such as mid-lift, providing the policy with diverse and physically consistent starting conditions to learn from.

Both types of reference motions, the automatically captured motions and the manually refined motions with human-object interactions, are incorporated into the training process using Adversarial Motion Priors (AMP) (Peng et al. 2021). AMP trains a discriminator network, implemented as a multi-layer perceptron (MLP), to distinguish between state transitions from the reference motions and those synthesized by the policy. The discriminator outputs the reference reward $r_t^{ref.}$ that encourages the policy to produce motions whose state-transition distribution

matches that of the reference motions. Through this adversarial training process, the policy learns to reproduce the detailed aspects of correct MMH techniques, such as smooth motion trajectories and proper leg-driven lifting patterns, without explicitly encoding them in the mathematical function.

4.3.3 Learning from Expert Interactions for Task-Specific Tacit Knowledge

Despite the guideline reward and reference reward, reinforcement learning policies occasionally exhibit unexpected behaviors that, while satisfying the designed rewards, do not reflect correct real-world MMH techniques. For example, when trained to lift containers with handles, the policy may learn to grip the handles on the adjacent side of the box and hug the container, rather than using the handles on opposite sides as proper technique requires. Correcting such behaviors traditionally requires either a substantially larger imitation dataset demonstrating the correct behaviors (Jiang et al. 2025), which is time-consuming and resource-intensive to collect, or extensive manual reward engineering, which demands expertise in both reinforcement learning and ergonomics (Araki et al. 2025).

To address this, we introduce a framework that captures and stores the task-specific tacit knowledge of ergonomics experts as a knowledge reward function using an LLM (Figure 4.2). We first train the policy with the guideline reward and reference reward. The trained policy's behaviors are then visualized and reviewed by an ergonomics expert knowledgeable in correct MMH techniques for the specific task category. When the expert identifies incorrect or undesirable behaviors, they provide natural language feedback describing the observed issue and the desired behavior. This feedback is then processed by an LLM, which translates the expert's natural language descriptions into mathematical reward functions that can be incorporated into the next round of policy training. For this purpose, we selected Claude Opus 4.6 (Anthropic 2026) for its

leading performance in code generation. This process effectively encodes the expert's tacit knowledge of correct techniques into explicit reward functions. For example, in the handle container case, the expert's feedback that the humanoid should grip the handles on the opposite side of the box is translated into a reward function that encourages the hands to reach toward those handles.

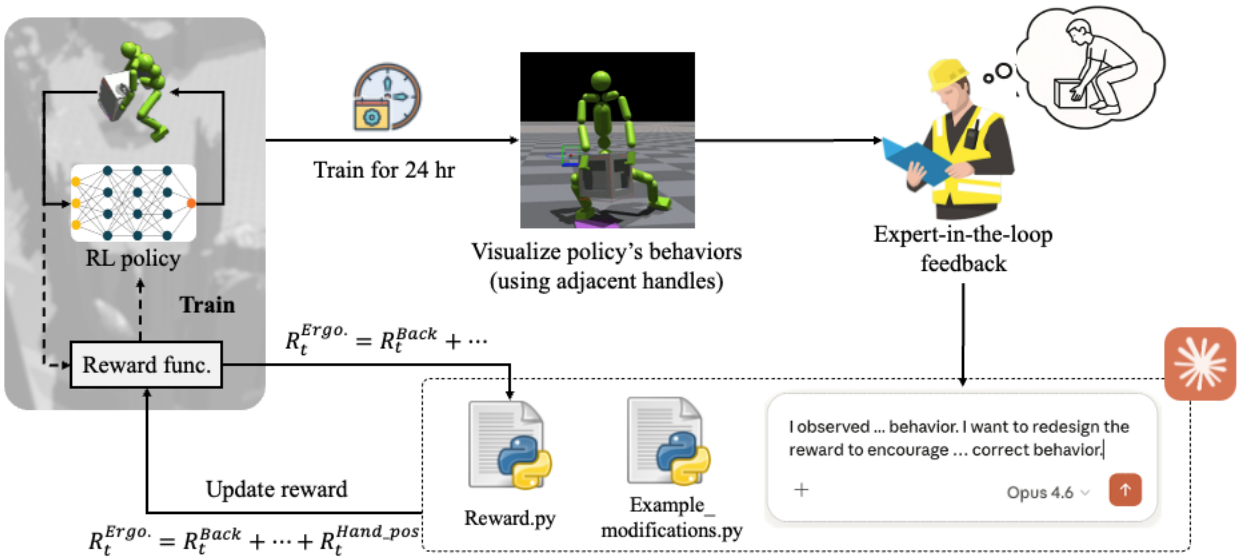


Figure 4.2. Tacit knowledge capture pipeline through experts' natural language feedback.

4.4 Computational Framework: Generating Task and Site-specific Ergonomics Intervention Applying Motion Synthesis

For scene-aware motion synthesis to be a practical tool for intervention recommendations in real-world scenarios, site personnel, such as workers and managers, need to be able to specify site and task constraints, essential inputs to scene-aware motion synthesis, using readily available devices. Once WMSD risks are detected (e.g., through vision-based assessment methods (Wen et al. 2026)), our computational framework enables site personnel to describe task constraints through natural language and to capture site constraints with smartphone-based LiDAR (Figure 4.3).

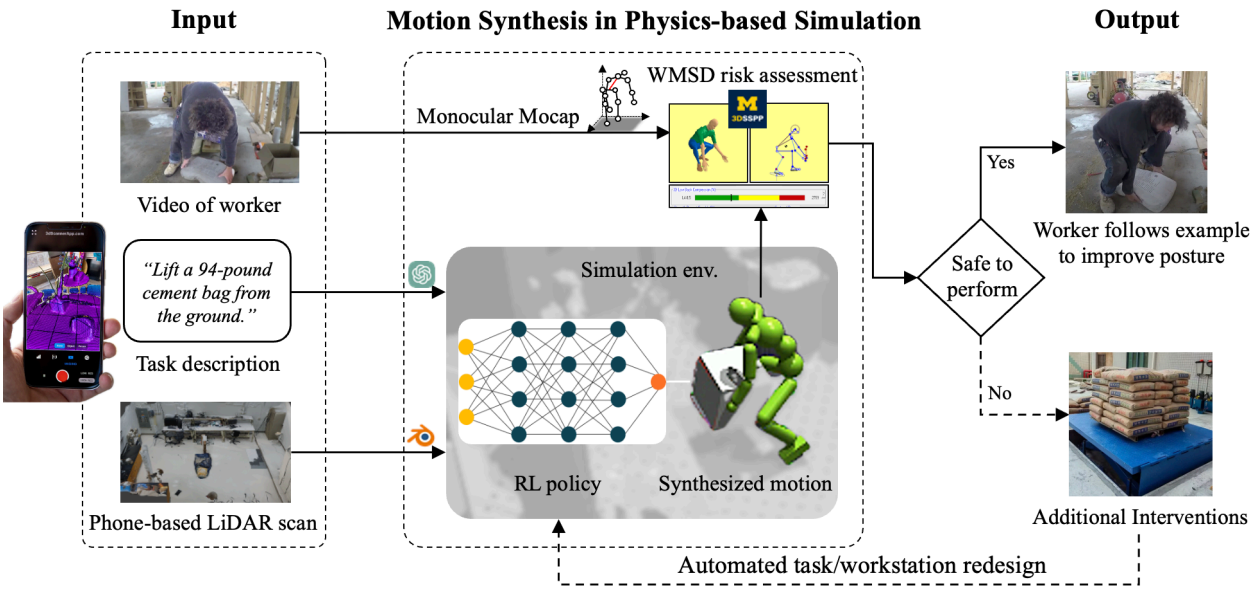


Figure 4.3. Computational framework for motion synthesis–based ergonomics interventions.

The computational framework extracts task and site constraints and feeds them into the scene-aware motion synthesis method. Task constraints define the MMH task scenario to be simulated and are extracted from the site personnel's natural language task descriptions. The description is parsed using a prompt-engineered GPT-5 model, selected for its leading natural language processing capabilities (OpenAI 2025), into a structured JSON file containing the extracted task parameters. These parameters include the object type, which determines the corresponding asset loaded into the Isaac Gym simulation environment for the interaction (e.g., container, lumber, non-rigid bag); the object dimensions and weight, which define the geometry and mass properties of the simulated asset; and the lift height, which specifies the initial position of the object at the start of the MMH task. Together, these parameters define the reinforcement learning task specification within the physics-based simulation environment.

Site constraints define the physical workspace within which the motion is synthesized. Site personnel capture the 3D geometry of the workspace using smartphone-based LiDAR scanning. The captured 3D scans are converted into 2D height maps using the Blender Python API (Blender

Foundation 2026), which represent terrain elevation, obstacles, and spatial boundaries of the workspace. This height map is imported as the ground and collision geometry in the Isaac Gym simulation environment, enabling the simulated humanoid to respect real-world spatial constraints such as confined spaces, uneven ground, and nearby obstacles that limit feasible postures.

With the extracted task and site constraints, the framework synthesizes motions demonstrating risk-reduced techniques for performing the specified task in the captured environment. The synthesized motions are then evaluated through an automated analysis pipeline that supports both WMSD risk assessment and biomechanical analysis. For WMSD risk assessment, a custom Python script computes the NIOSH Lifting Index (Waters et al. 2021) from the synthesized motions. For biomechanical analysis, the synthesized motions are exported as batch input files and fed into 3D SSPP (VelocityEHS 2024b) through a custom wrapper that interfaces with the software programmatically, bypassing the manual GUI workflow and enabling automated batch analysis. The analysis results are automatically exported, and the potential WMSD risk reduction is quantified by comparing the synthesized motions against the captured task motions. The evaluation results also inform the intervention recommendation. If the synthesized posture improvements are sufficient to reduce WMSD risk to an acceptable level, workers can use the visualized motion demonstrations as guides to improve their technique when performing the specified task. However, in cases where posture improvements alone are insufficient (i.e., the task is too physically demanding regardless of technique), the framework automatically proposes task or workstation modifications (e.g., reducing object weight or adjusting lift height) and performs iterative "what-if" analyses within the simulated environment, re-synthesizing motions under the modified conditions until an acceptable level of WMSD risk is

achieved. This process identifies the extent of task or workstation redesign needed to bring WMSD risk within acceptable limits.

4.5 Evaluation of the Proposed Intervention Recommendation Framework Under Various Task and Site Constraints

We evaluated the proposed intervention recommendation framework under two types of scenarios: unconstrained task scenarios and site-constrained task scenarios. For the unconstrained task scenarios, we designed 12 scenarios across four MMH task categories common in construction: lifting containers, lifting containers with handles, lifting irregular objects such as lumber, and lifting non-rigid objects such as cement bags, with varying task parameters such as object weight, size, and difficulty level across scenarios. For the site-constrained task scenarios, we designed scenarios where the MMH tasks are performed in the presence of site constraints, such as a raised step simulating a pallet placed below the object. We first verified that the computational framework correctly captures and processes the task and site constraints across all 16 scenarios and outputs suitable task- and site-specific interventions. We then evaluated whether the proposed motion synthesis method with ergonomics reward can synthesize motions that reduce WMSD risk relative to workers' original motions.

4.5.1 Verification of the Computational Framework

To verify that the computational framework correctly captures and processes real-world task and site constraints, we tested the end-to-end pipeline in a controlled laboratory environment. We captured a 3D LiDAR scan of the laboratory using an iPhone 14 Pro with a free 3D scanning application (Polycam 2022). The scan was processed into a 2D height map using the Blender Python API and imported as the ground and collision geometry in the Isaac Gym simulation environment, and the resulting simulation terrain was visually verified to match the spatial layout

of the scanned laboratory. We then specified task parameters in natural language across all 16 evaluation scenarios, and the prompt-engineered GPT-5 model parsed each description into structured JSON files. The extracted parameters were verified to correctly reflect the intended object type, dimensions, weight, and lift height for each scenario. The framework successfully synthesized motions that carried out the specified MMH tasks within the scanned environment using the extracted constraints.

4.5.2 Test Subjects and Data Collection

We aim to assess whether the proposed motion synthesis method can produce motions with reduced WMSD risk compared to workers' captured motions, rather than to characterize variability in human lifting behavior. For this purpose, we recruited three male university graduate students within the same BMI classification and instructed them to lift the specified objects following their natural technique without particular consideration for ergonomics. Each subject's captured motion serves as the baseline against which the synthesized motion is compared, and recruiting multiple subjects allows us to verify that the risk reduction pattern is consistent across subjects. In total, this yields 48 comparisons (36 unconstrained and 12 site-constrained). Table 4.1 summarizes their demographic information, including height, weight, and BMI. Their motions were recorded using a smartphone camera, and 3D human motion was extracted from the recorded videos using monocular motion capture for subsequent WMSD risk evaluation.

Table 4.1 Height, weight, and BMI of test subjects.

Participant #	Gender	Height (m)	Weight (Kg)	BMI
#1	Man	1.77	88	28.1
#2	Man	1.71	84	28.7
#3	Man	1.80	80	25.3

4.5.3 Evaluation metric

We evaluated the WMSD risk of both the captured task motions and the recommended motions synthesized under the same task and site constraints using two established evaluation tools: the NIOSH Lifting Equation (Waters et al. 2021) and 3D SSPP (VelocityEHS 2024b).

For the NIOSH Lifting Equation, we identified key lift start and lift end frames from each motion and calculated the multiplier factors directly from the 3D human motion data, whether from the simulated humanoid or the captured human motion. These multipliers are used to compute the Recommended Weight Limit (RWL) and the Lifting Index (LI). The RWL represents the weight that nearly all healthy workers could lift without increased risk under the specified body posture and MMH task conditions, and the LI provides a relative estimate of the physical stress for the MMH task (Waters et al. 2021).

For 3D SSPP, we analyzed the posture at the lift start position, representing the most biomechanically demanding point of the lift. The 3D human motion data were formatted as a batch file and imported into the 3D SSPP software. All postures analyzed in 3D SSPP were manually refined within the software following its user instructions to match the visuals from the video recordings or the visualized physics-based simulation. From 3D SSPP, we extracted the L4/L5 back compression force as well as the joint-wise strength percent capable scores, which indicate the percentage of the population capable of safely exerting the required strength at each joint for the analyzed posture. Together, they serve as valuable metrics for evaluating the WMSD risk of MMH task postures and for determining whether the task can be performed safely.

4.5.4 Evaluation of Synthesized Motions on Unconstrained MMH Task Scenarios

We selected four MMH task categories that are common in construction, and each requires distinct lifting techniques to perform safely: lifting containers, lifting containers with handles, lifting irregular objects such as lumber, and lifting non-rigid objects such as concrete bags. Figure 4.4 illustrates the task parameters for each scenario. The container and container with handle categories each include four scenarios varying in object size (13 in. and 18 in.) and weight (10 lbs. and 20 lbs.), while the irregular object category includes two scenarios varying in length (3 ft and 5 ft), and the non-rigid bag category includes a 25 lbs. bag of cement and a 45 lbs. bag of stones, resulting in 12 unconstrained task scenarios in total.

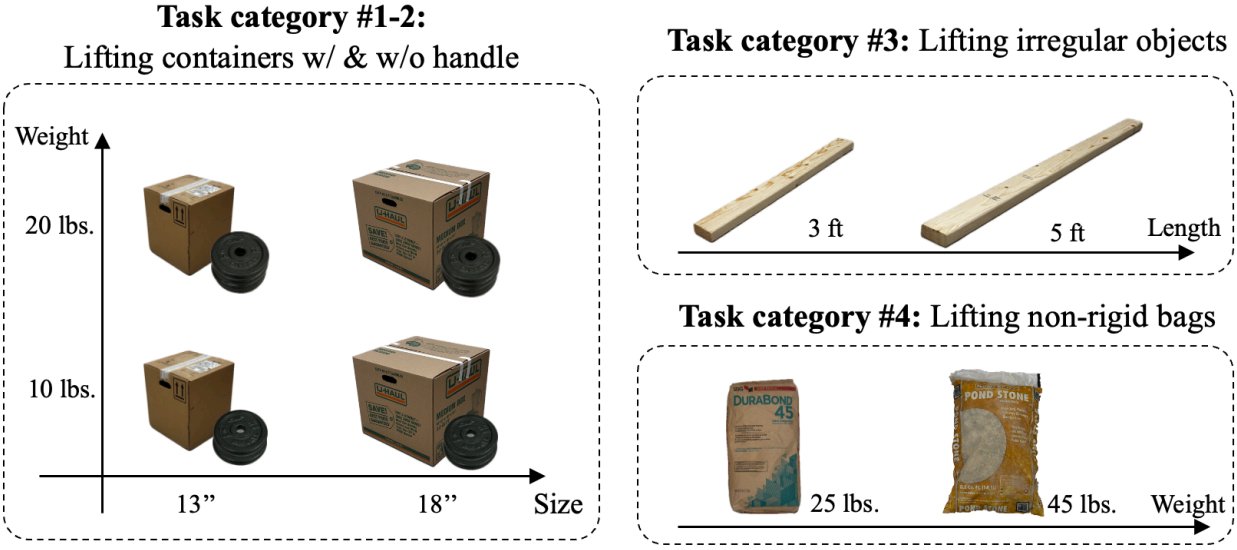


Figure 4.4. Task parameters for the 12 unconstrained MMH task scenarios across four task categories.

4.5.4.1 Visualization of the Synthesized and Captured Motions

We first visualized the captured task motions alongside the synthesized motions for each of the four task categories (Figure 4.5). Compared to the captured task motions, the synthesized motions consistently lift with the knees and bring the object as close to the body as possible, reflecting the ergonomics reward’s emphasis on proper lifting technique.

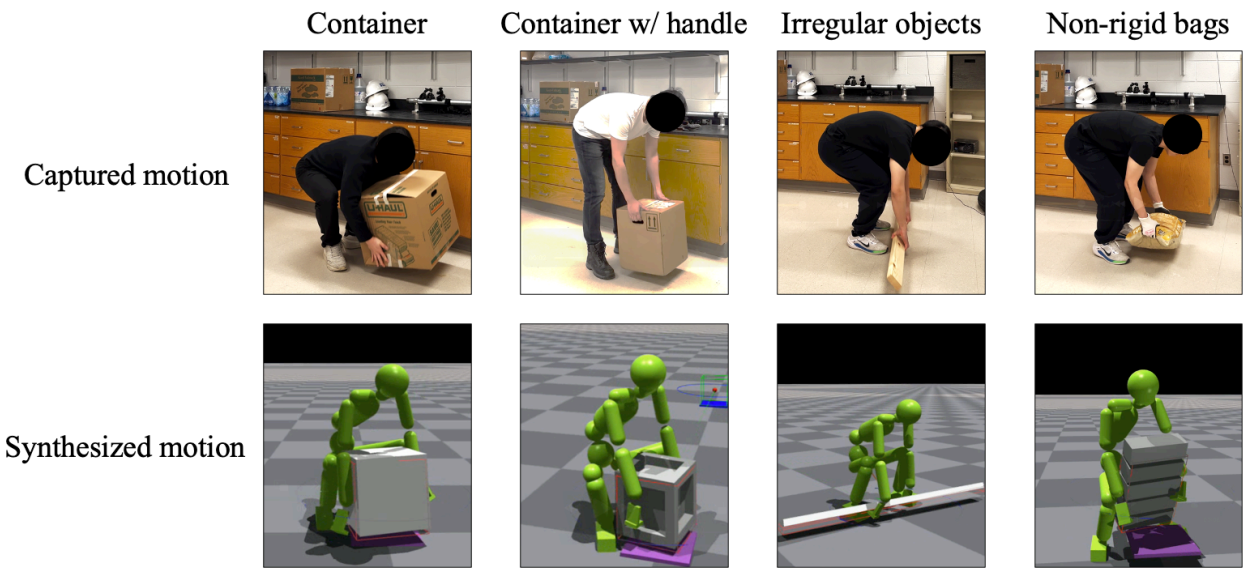


Figure 4.5. Visual comparison of captured task motions and synthesized motions across the four MMH task categories.

4.5.4.2 NIOSH Lifting Index and Recommended Weight Limit

We then compared the NIOSH LI and RWL between the captured task motions and the synthesized motions. Table 4.2 reports the LI and RWL for each of the 12 unconstrained task scenarios, where the captured motion values represent the average across the three test subjects. Across all scenarios, the synthesized motions achieved a lower LI than the captured task motions, with an average reduction of 40%. The RWL increased by an average of 4.0 kg, indicating that the synthesized postures allow workers to safely handle heavier loads under the same task conditions.

Table 4.2. NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in unconstrained task scenarios.

MMH tasks	Scenarios	NIOSH Lifting Index ↓			NIOSH RWL (kg) ↑		
		Captured motions	Synthesized motions	Difference (%)	Captured motions	Synthesized motions	Difference
Container	10 lbs. / S	0.68	0.42	-38%	6.7	10.8	4.1
	20 lbs. / S	1.45	0.88	-39%	6.3	10.3	4.0
	10 lbs. / L	0.77	0.47	-39%	5.9	9.6	3.7
	20 lbs. / L	1.46	0.82	-44%	6.2	11.1	4.9
Container w/o handle	10 lbs. / S	0.67	0.46	-30%	7.0	9.8	2.8
	20 lbs. / S	1.25	0.77	-38%	7.4	11.8	4.4
	10 lbs. / L	0.72	0.56	-22%	6.3	8.1	1.8
	20 lbs. / L	1.4	0.99	-29%	6.5	9.1	2.7
Irregular object	3 ft	0.2	0.12	-41%	6.8	11.4	4.6
	5 ft	0.3	0.2	-33%	7.5	11.3	3.7
Non-rigid bag	25 lbs.	1.6	0.86	-46%	7.1	13.2	6.1
	45 lbs.	3.21	1.69	-47%	6.4	12.1	5.7
Average		**1.14**	**0.69**	**-40%**	**6.7**	**10.7**	**4.0**

To test whether the synthesized motions achieve statistically significant WMSD risk reduction compared to the captured task motions, we use a linear mixed model (LMM) with motion source (captured vs. synthesized) as a fixed effect and scenario as a random intercept. The random intercept accounts for the substantial differences in baseline WMSD risk across scenarios due to variations in task category, object weight, and object shape (e.g., lifting a 3 ft piece of lumber weighing 3 lbs. vs. a 45 lbs. bag of stones). Without this structure, pooling observations across scenarios can conflate between-scenario variability with the effect of interest, potentially masking or inflating the true difference between captured and synthesized motions.

We visualized the distribution of individual observations for each motion source as box plots in Figure 4.6, displaying the interquartile range (IQR), median, mean, individual data points,

and LMM test significance. The synthesized motions achieved significantly lower LI and significantly higher RWL compared to the captured motions ($p < 0.001$).

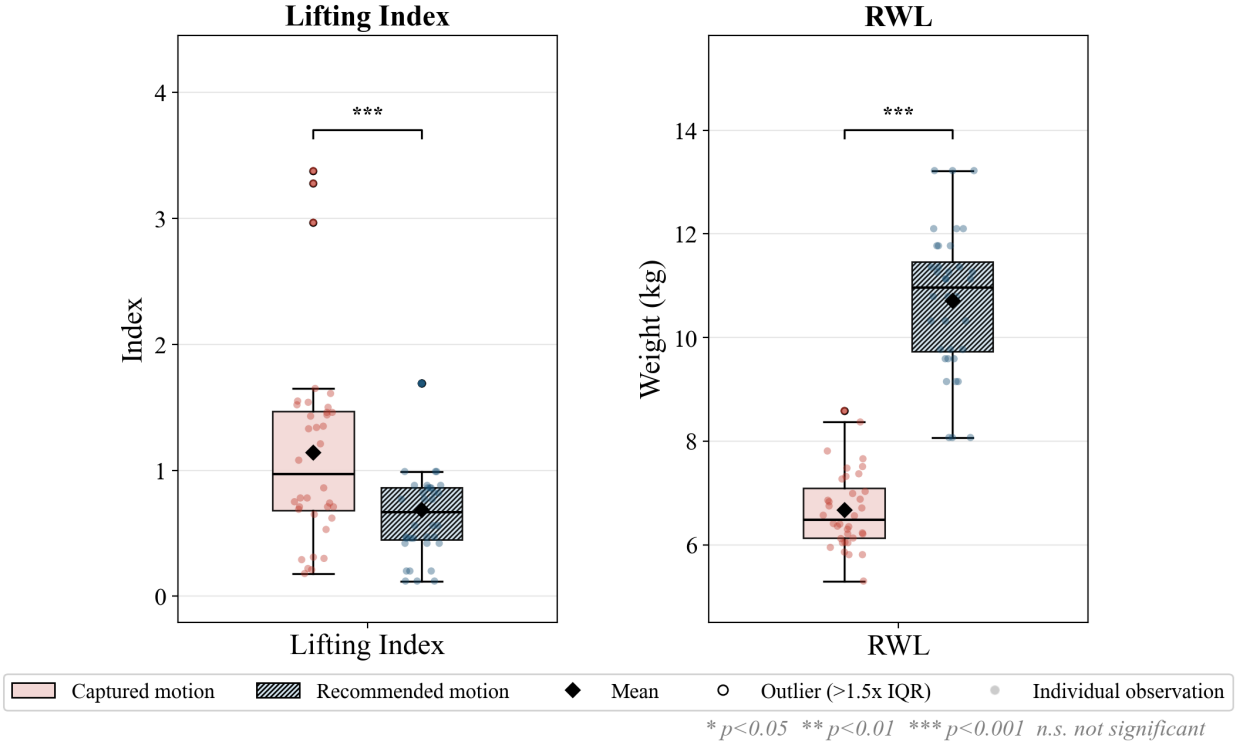


Figure 4.6. Box plots comparing the NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in unconstrained task scenarios.

4.5.4.3 3D SSPP Back Compression Force and Joint Strength Precent Capable

We also compared the 3D SSPP L4/L5 compression force and joint strength percent capable scores between the captured and synthesized motions. We visualized the distribution of individual observations for each motion source as box plots in Figure 4.7, and reported the scenario-level comparison in Table 4.3. All differences are statistically significant ($p < 0.001$).

The synthesized motions reduced the L4/L5 compression force by an average of 11% and improved the hip strength percent capable from 66.0% to 93.5%, indicating substantially reduced back compression force and improved hip joint strength requirements. However, the knee strength percent capable decreased from 96.0% to 74.7%. This tradeoff is expected, as the synthesized

motions emphasize lifting with the legs to reduce back loading, which in turn places greater biomechanical demand on the knee joints.

Table 4.3. 3D SSPP L4/L5 compression force and joint strength percent capable scores of synthesized and captured motions in unconstrained task scenarios.

MMH tasks	Scenarios	L4/L5 Compression Force (N) ↓			Hip Strength Percent Capable (%) ↑			Knee Strength Percent Capable (%) ↑		
		Captured motions	Synthesized motions	Difference (%)	Captured motions	Synthesized motions	Difference (pp)	Captured motions	Synthesized motions	Difference (pp)
Container	10 lbs. / S	3329	3135	-6%	77.9	95.8	18%	99.1	66.5	-33%
	20 lbs. / S	3757	3465	-8%	69.4	94.2	25%	92.6	80.5	-12%
	10 lbs. / L	3422	3274	-4%	73.4	95.2	22%	96.0	88.6	-7%
	20 lbs. / L	3741	3460	-8%	63.7	94.8	31%	83.1	87.7	5%
Container w/o handle	10 lbs. / S	3321	2587	-22%	65.5	92.7	27%	99.8	80.1	-20%
	20 lbs. / S	3584	3161	-12%	57.9	89.5	32%	99.7	79.4	-20%
	10 lbs. / L	3392	3252	-4%	61.4	89.4	28%	99.9	86.1	-14%
	20 lbs. / L	3651	3264	-11%	50.3	87.6	37%	99.7	88.5	-11%
Irregular object	3 ft	2985	2600	-13%	76.5	96.4	20%	96.7	89.1	-8%
	5 ft	3062	2593	-15%	77.1	96.6	19%	93.6	91.9	-2%
Non-rigid bag	25 lbs.	3795	3430	-10%	72.2	95.6	23%	97.9	34.9	-63%
	45 lbs.	4689	3898	-17%	46.6	94.2	48%	93.6	23.2	-70%
Average		**3561**	**3176**	**-11%**	**66.0**	**93.5**	28%	**96.0**	**74.7**	**-21%**

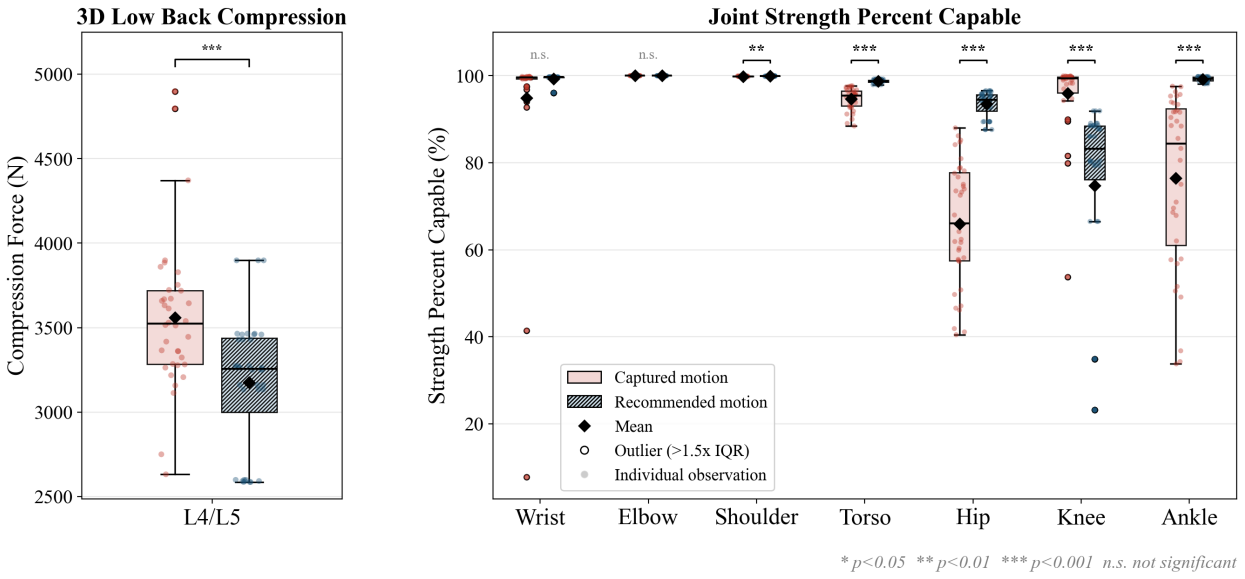


Figure 4.7. Box plots comparing the 3D SSPP L4/L5 compression force and joint strength percent capable scores of captured and synthesized motions in unconstrained task scenarios.

4.5.5 Evaluation of Synthesized Motions on Site-constrained MMH Task Scenarios

In practice, site constraints such as pallets, deep shelves, or obstacles can prevent workers from adopting the optimal posture for their task, even when the correct technique is known. For example, when objects are placed on a raised pallet, the pallet prevents the worker from standing close to the object, increasing the horizontal distance between the body and the load during the lift. In such cases, the recommended motion must be adapted to the site conditions. For instance, stepping onto the pallet brings the body closer to the object, reducing the horizontal distance and the resulting biomechanical demand.

To evaluate the framework's ability to generate interventions that account for such site constraints, we designed a scenario where the MMH task is performed in the presence of a raised step simulating a pallet placed below the object. The site geometry was scanned and processed using the computational framework introduced in Section 4.3. For the site-constrained evaluation, we selected one scenario per task category, resulting in four site-constrained scenarios in total. We

evaluated these scenarios using the same metrics and comparison procedure as the unconstrained task scenarios.

4.5.5.1 Visualization of the Synthesized and Captured Motions

We visualize the captured task motions alongside the synthesized motions for the site-constrained scenarios (Figure 4.8). In the captured motions, the test subjects typically reach over the raised step to lift the object from a standing position on the ground, especially for lighter objects such as lumber, resulting in an increased horizontal distance between the body and the load. In contrast, the synthesized motions step onto the raised step before lifting, bringing the body closer to the object and reducing the horizontal reach required during the lift.

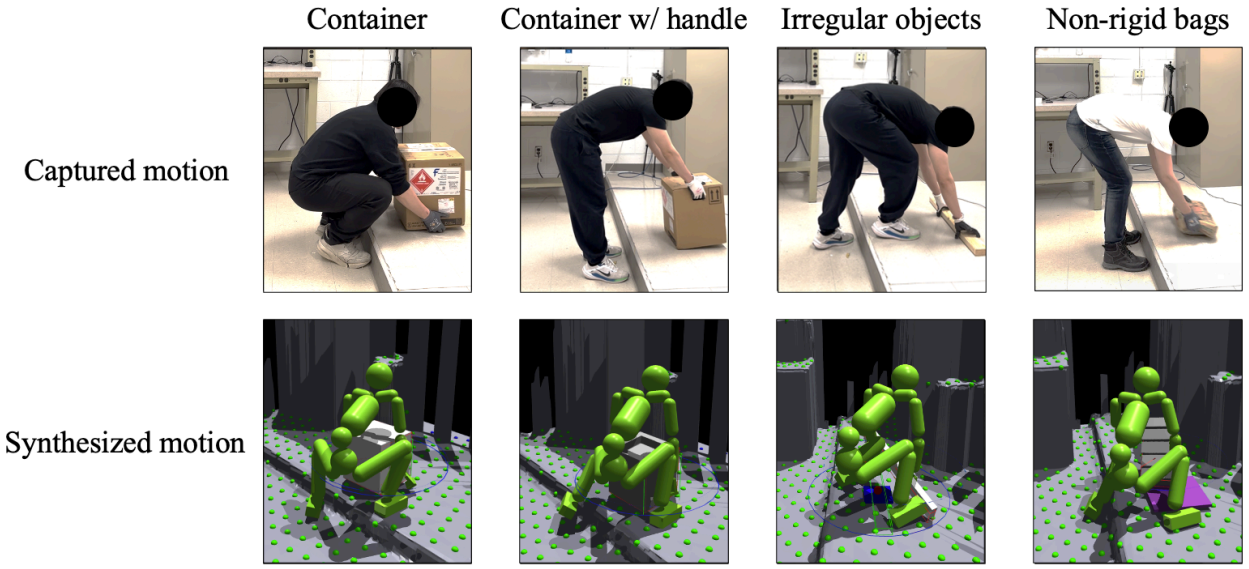


Figure 4.8. Visual comparison of captured task motions and synthesized motions across the four MMH task categories under site constraints.

4.5.5.2 NIOSH Lifting Index and Recommended Weight Limit

We compared the NIOSH LI and RWL between the captured and synthesized motions for the four site-constrained scenarios (Figure 4.9, Table 4.4). The synthesized motions achieved a lower LI in all four scenarios, with an average reduction of 38%. The RWL increased by an average of 5.0 kg. The LMM test confirms that both differences are statistically significant ($p < 0.001$).

The irregular object scenario showed the largest LI reduction (62%), as test subjects tended to be less conscious of their posture when lifting lighter objects such as lumber, resulting in a greater gap between their natural technique and the synthesized motion.

Table 4.4. NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in site-constrained task scenarios.

MMH tasks	**Scenarios**	**NIOSH Lifting Index ↓**			**NIOSH RWL (kg) ↑**		
		Captured motions	Synthesized motions	Difference (%)	Captured motions	Synthesized motions	Difference
Container	10 lbs. / S	0.75	0.46	-38%	6.1	9.8	3.7
Container w/o handle	10 lbs. / S	0.74	0.41	-44%	6.2	11.0	4.8
Irregular object	3 ft	0.25	0.09	-62%	5.6	14.4	8.8
Non-rigid bag	25 lbs.	1.95	1.32	-33%	5.9	8.6	2.8
Average		**0.92**	**0.57**	**-38%**	**5.9**	**11.0**	**5.0**

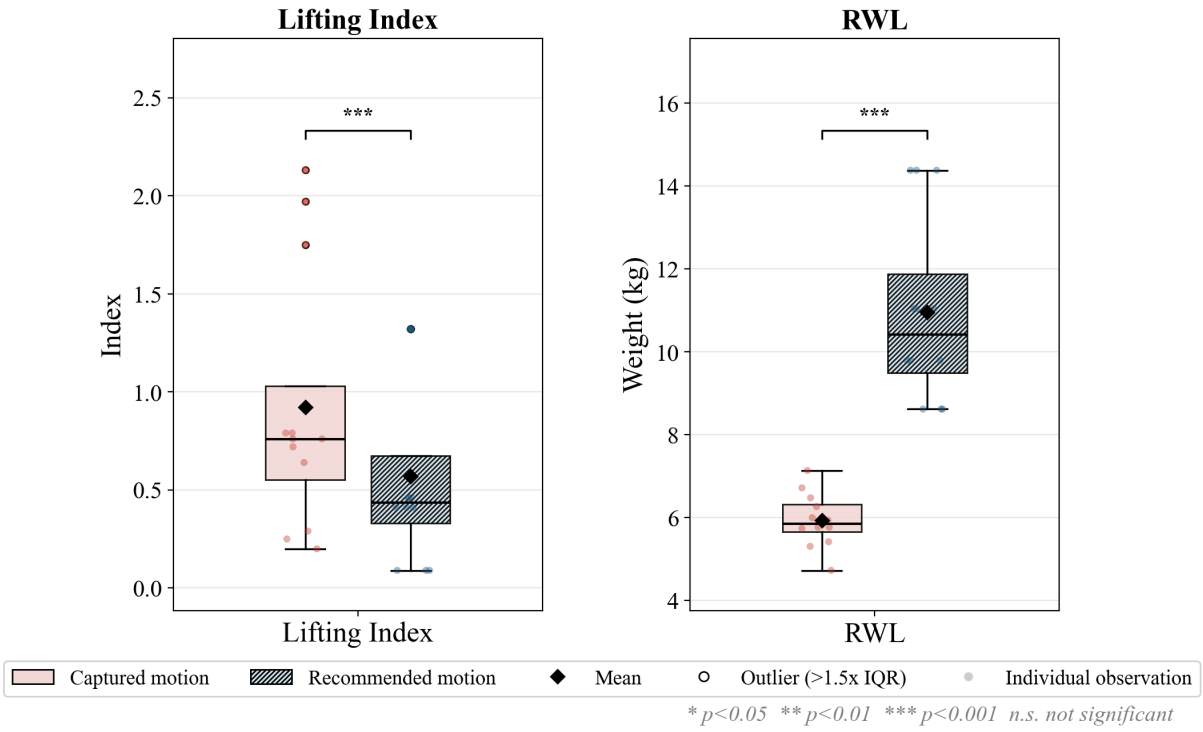


Figure 4.9. Box plots comparing the NIOSH Lifting Index and Recommended Weight Limit of captured and synthesized motions in constrained task scenarios.

4.5.5.3 3D SSPP Back Compression Force and Joint Strength Precent Capable

We compared the 3D SSPP results between captured and synthesized motions for the site-constrained scenarios (Figure 4.10, Table 4.5). The synthesized motions reduced the L4/L5 compression force by an average of 12% and improved the hip strength percent capable from 83.7% to 93.6%. The LMM test confirms that these differences are statistically significant ($p < 0.001$ for L4/L5 compression, $p < 0.05$ for hip strength percent capable). The knee strength percent capable decreased from 97.4% to 91.7%, though this difference is not statistically significant. Similar to the unconstrained scenarios, the slight decrease in knee strength percent capable reflects the greater reliance on the knees during lifting in the synthesized motions.

Table 4.5. 3D SSPP L4/L5 compression force and joint strength percent capable scores of synthesized and captured motions in constrained task scenarios.

MMH tasks	Scenarios	L4/L5 Compression Force (N) ↓			Hip Strength Percent Capable (%) ↑			Knee Strength Percent Capable (%) ↑		
		Captured motions	Synthesized motions	Difference (%)	Captured motions	Synthesized motions	Difference (pp)	Captured motions	Synthesized motions	Difference (pp)
Container	10 lbs. / S	3611	3269	-9%	84.0	91.8	8%	95.0	96.2	1%
Container w/o handle	10 lbs. / S	3210	2838	-12%	80.4	92.4	12%	98.4	95.9	-2%
Irregular object	3 ft	3245	2737	-16%	88.0	96.8	9%	98.1	77.1	-21%
Non-rigid bag	25 lbs.	4400	3941	-10%	82.4	93.4	11%	98.2	97.6	-1%
Average		**3616**	**3196**	**-12%**	**83.7**	**93.6**	**10%**	**97.4**	**91.7**	**-6%**

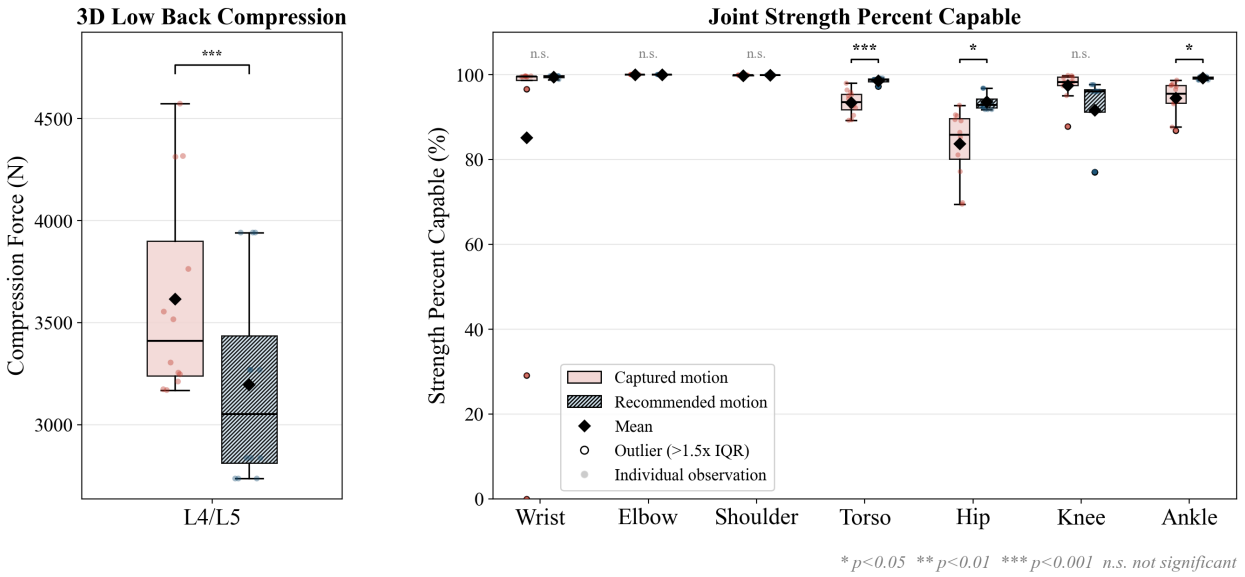


Figure 4.10. Box plots comparing the 3D SSPP L4/L5 compression force and joint strength percent capable scores of captured and synthesized motions in constrained task scenarios.

4.6 Discussion

The proposed motion synthesis method with ergonomics reward consistently produced motions with reduced NIOSH LI and L4/L5 back compression force compared to the captured task motions across all tested scenarios. The NIOSH LI improvement is consistent with the design of the guideline reward, which directly incorporates principles from the NIOSH Lifting Equation, such as minimizing the horizontal distance between the load and the body. The L4/L5 compression force reduction, while not directly optimized in our reward function, is a natural consequence of the improved lifting postures. 3D SSPP is a standalone Windows software that cannot be integrated into the reinforcement learning training loop for reward calculations, but the proper back and lifting techniques encouraged by our ergonomics reward are closely aligned with the biomechanical principles that reduce back compression, resulting in consistent improvements in this metric as well.

On the other hand, the 3D SSPP strength percent capable scores revealed a consistent tradeoff in joint strength requirements between the hip and ankle joints on one side and the knee

joint on the other. In both unconstrained and site-constrained scenarios, the synthesized motions consistently improved hip and ankle strength percent capable while decreasing knee strength percent capable (Figure 4.7, Figure 4.10). This tradeoff is inherent to lifting motions predominantly using the legs, as transferring the load from the back to the lower extremities reduces back compression but increases the biomechanical demand on the knees. However, by adopting the recommended postures, the lowest joint strength percent capable across all joints increased from 66.0% to 74.7% in unconstrained scenarios, and from 83.7% to 91.7% in site-constrained scenarios, indicating that a larger proportion of the population can safely perform the task by following the synthesized motions. However, individuals with prior knee injuries or conditions may not be suitable candidates to adopt the recommended lifting technique and may benefit from alternative interventions such as task or workstation modifications.

Beyond the quantitative improvements in WMSD risk metrics, the synthesized motions demonstrated proper MMH techniques, including lifting with greater knee flexion while maintaining a straighter back, aligning with NIOSH lifting guidelines that recommend lifting with the legs rather than the upper body or back (NIOSH 2007). The synthesized motions also straddled objects between the legs when feasible and positioned near the corner of the object during lifts to keep the body closer to the load, a practical technique recommended in ergonomics guidelines (Loyola University Chicago 2024) and workplace health and safety training programs (Elite Force Safety 2025). Notably, this straddling technique was not explicitly encoded in the reward function but emerged from the combination of encouraging the load to stay close to the body and imitating safe motion examples, demonstrating that the ergonomics reward can capture proper MMH techniques beyond what is explicitly specified.

While this chapter evaluated the framework on four MMH task categories, the proposed approach is designed to be adaptable to new task categories beyond those presented here. Adapting to a new task category involves three steps: collecting reference videos of correct techniques using a smartphone, running the monocular motion capture pipeline to extract reference motions for imitation learning, and iteratively applying the tacit knowledge capture framework to refine undesirable behaviors until the policy learns the correct techniques specific to the new task. Once the policy exhibits the desired behaviors, no further expert interaction is needed for that task category. The accessibility of each step, requiring only smartphone video and natural language feedback from ergonomics experts rather than reinforcement learning or simulation expertise, makes it possible to extend the framework to new MMH tasks as they are encountered on construction sites.

While the proposed intervention recommendation framework demonstrates the potential for an accessible and resource-efficient way to generate interventions, this study has several limitations that should be acknowledged. First, the evaluation was conducted with three male graduate students and a limited set of site constraint scenarios (one site constraint type across four task categories), which limits the statistical power and the diversity of the evaluation. Future work should include a larger and more diverse participant pool and a broader range of site constraint scenarios to strengthen the generalizability of the findings. Second, further validation through real-world deployment is needed to confirm that the recommended task- and site-specific interventions, including both posture demonstrations and task or workstation redesign suggestions, are suitable for real-world scenarios and lead to actual WMSD risk reduction in practice.

4.7 Conclusion

This chapter presented a simulation-driven computational framework for generating task- and site-specific ergonomics interventions through scene-aware motion synthesis. A key challenge for using motion synthesis as an intervention tool is synthesizing motions that demonstrate correct MMH techniques and are safe to perform. To address this, we developed a custom ergonomics reward that leverages established ergonomics guidelines and imitation learning to guide the reinforcement learning policy toward risk-reduced motions. To further refine the policy for task-specific MMH techniques, we captured tacit knowledge from ergonomics experts through natural language feedback, which is translated into reward functions using a large language model.

We evaluated the proposed framework under 12 unconstrained task scenarios and 4 site-constrained scenarios. The synthesized motions consistently reduced WMSD risk compared to captured task motions, achieving an average 39% reduction in NIOSH Lifting Index and 11% reduction in L4/L5 back compression force across all scenarios. These results demonstrate the framework's potential as an intervention recommendation tool. The risk-reduced synthesized motions can serve as posture demonstrations for workers, providing visual guidance on correct MMH techniques for their specific task and site conditions. When posture improvements alone are insufficient to reduce WMSD risk to an acceptable level, the synthesized human motions enable the framework to further perform automated "what-if" analysis to quantitatively identify necessary task or workstation modifications. By requiring only smartphone-captured inputs and natural language task descriptions, the framework provides an accessible and resource-efficient means to generate ergonomics interventions.

Chapter 5 Quantitative Evaluation Of Human Pose Representations for WMSD Risk Assessments

5.1 Introduction

In Chapters 2 and 3, we proposed a vision-based WMSD risk assessment pipeline using keypoint-based pose representations, where a custom set of 66 3D anatomical landmarks and joint centers is used to represent the human posture and calculate the high-DoF joint angles. While this keypoint-based representation supports accurate joint angle calculations in the definitions required by WMSD risk assessments, it also imposes a structural limitation: the fidelity of downstream joint angle estimation is bounded by the sparsity and coverage of the keypoint set defined during data collection and model training.

Recently, mesh-based human body models have gained increasing adoption in the computer vision community as an alternative to sparse keypoint representations (Ma et al. 2023; Yin et al. 2025). Methods such as SMPL (Loper et al. 2015) parameterize the human body as a deformable surface mesh consisting of thousands of vertices that densely cover the body surface (e.g., 6,890 vertices in SMPL), driven by a compact set of pose and shape parameters. Because the mesh encodes the full body surface, any anatomical landmark on the body surface can be mapped to a corresponding vertex or interpolated surface point, making it available for high-DoF joint angle calculation without requiring that landmark to have been explicitly defined before data collection and model training. This makes mesh-based representations inherently more versatile for adapting to different downstream tasks, such as computing joint angles under alternative

anatomical definitions required by different WMSD risk assessment tools, or retargeting the captured motion to humanoid robots or digital human models for simulation-based analysis.

Despite the potential benefits, the impact of mesh-based representations on joint angle estimation accuracy compared to keypoint-based representations has not been systematically investigated. This question is critical for WMSD risk assessment, where joint angle accuracy directly determines the validity of risk scores. This chapter addresses this question through a controlled comparison, training both a keypoint-based and a mesh-based model on the same dataset with the same backbone architecture, and evaluating their joint angle estimation accuracy against motion capture ground truth on the joint angles used in WMSD risk assessments.

5.2 Quantitative Comparison between Keypoint-Based and Mesh-Based Pose Estimation Models

We designed an experiment to test whether the choice of pose representation (keypoint-based vs. mesh-based) significantly affects joint angle estimation accuracy for WMSD risk assessment. The null hypothesis is that there is no significant difference in joint angle estimation error between the two representations when trained on the same dataset with the same backbone architecture (Figure 5.1). To conduct this test, we need to (1) establish the correspondence between mesh vertices and the anatomical keypoints used for joint angle calculation, (2) convert the motion capture dataset into mesh format to create paired training data, and (3) train both models with the same architecture and configuration. With these steps completed, we calculated 22 joint angles from both models' 3D outputs using the same angle definitions, evaluated the estimation error of each model against motion capture ground truth, and compared the errors between the two models using paired t-tests with Bonferroni correction.

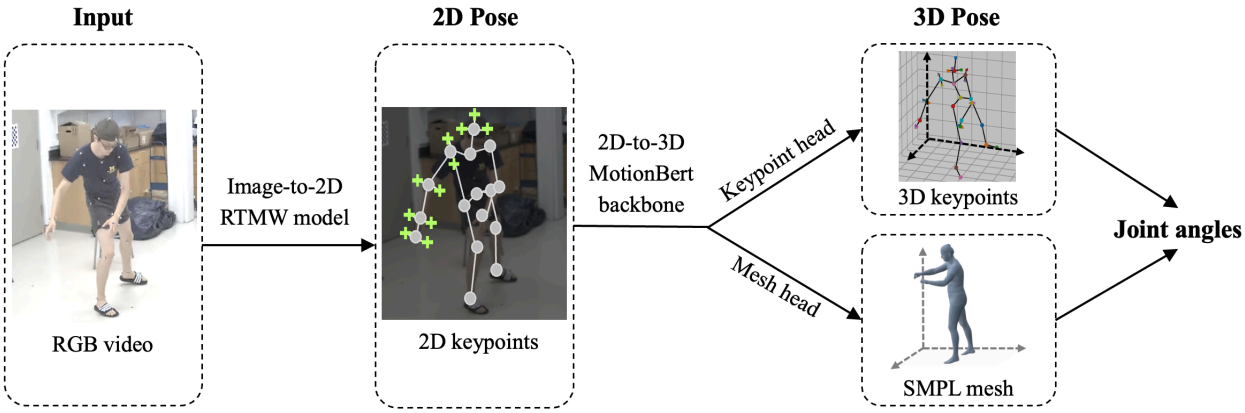


Figure 5.1. Model architecture for keypoint-based and mesh-based 3D pose estimation models.

5.2.1 Mesh to Body Landmark Mapping

To enable high-DoF joint angle calculation from the mesh representation of human pose, we first established the correspondence between the SMPL mesh model and the 66 keypoints defined in the VEHS-6.7M dataset (Chapter 2). Of the 66 keypoints, 47 are body surface landmarks (e.g., right acromion, left anterior superior iliac spine, C7) that lie on the skin surface and can be directly mapped to specific vertices on the mesh surface. We identified the corresponding vertex index for each of these 47 landmarks by visually locating the matching vertex on the SMPL template mesh in Blender. Table 5.1 lists the full correspondence mapping between the 47 surface landmarks and the SMPL vertex indices.

Table 5.1. Mapping between 47 body surface keypoints and SMPL vertex indices.

No.	Region	Keypoint	Vertex ID	No.	Region	Keypoint	Vertex ID
1	Head	HDTP	411	26	Foot	RMTP1	6739
2		REAR	3941	27		RMTP5	6745
3		LEAR	517	28		LMTP1	3337
4		MDFH	335	29		LMTP5	3344
5	Torso	C7	829	30		RHEEL	6786
6		C7_d	1305	31		LHEEL	3387
7		SS	3171	32	Shoulder	RAP	4721
8		T8	3024	33		RAP_b	5283
9		XP	3077	34		RAP_f	5354
10	Pelvis	RPSIS	5246	35		LAP	1238
11		RASIS	6573	36		LAP_b	2888
12		LPSIS	1781	37		LAP_f	1317
13		LASIS	3156	38	Elbow	RLE	5090
14	Thigh	RIC	5257	39		RME	5202
15		RGT	4927	40		LLE	1621
16		LIC	1794	41		LME	1732
17		LGT	1454	42	Wrist	RRS	5573
18	Knee	RMFC	4842	43		RUS	5568
19		RLFC	4540	44		LRS	2112
20		LMFC	1369	45		LUS	2108
21		LLFC	1054	46	Hand	RMCP2	5595
22	Ankle	RMM	6832	47		RMCP5	5636
23		RLM	6728	48		LMCP2	2135
24		LMM	3432	49		LMCP5	2290
25		LLM	3327				

The remaining 19 keypoints are joint centers (e.g., hip joint center, knee joint center) that lie inside the body and do not correspond to any surface vertex. These joint centers were computed from the 47 surface landmarks using the same definitions as in Chapter 2. Through this two-step process, the full set of 66 keypoints can be derived from a predicted body mesh: first by extracting the 47 surface landmarks via their vertex indices, then by computing the 19 joint centers. This ensures that the same joint angle definitions are applied to both the keypoint-based and mesh-based model outputs.

5.2.2 Mesh Dataset Preparation

To conduct a controlled comparison between keypoint-based and mesh-based pose representations, both models need to be trained on the same underlying motion data. The VEHS-6.7M dataset (Chapter 2) provides optical motion capture recordings with 66 keypoints defined on the body surface and at joint centers, but does not include SMPL mesh parameters. We therefore converted the keypoint-based motion capture data into SMPL mesh representations to create a paired dataset suitable for training both models.

We performed this conversion using MoSh++, an optimization-based method that fits SMPL body model parameters to optical marker-based motion capture data (Ghorbani and Black 2021; Loper et al. 2014; Mahmood et al. 2019). Given a set of labeled markers and their known correspondences to locations on the body mesh vertex indexes (defined in Section 5.2.1), MoSh++ optimizes the SMPL pose parameters, shape parameters, and per-frame global translation to minimize the distance between the model surface and the observed marker positions. The method works with arbitrary marker sets and jointly recovers body shape across the sequence, producing SMPL parameters that are consistent in shape across all frames for a given subject. For each motion capture sequence, we produced per-frame SMPL pose parameters (72 dimensions), shape parameters (10 dimensions), and global translation (3 dimensions). The resulting SMPL parameters, paired with the original synchronized video frames, constitute the mesh-based training data. This conversion yields a paired dataset where each video frame is associated with both a keypoint-based ground truth and a mesh-based ground truth, ensuring that any observed difference in joint angle accuracy can be attributed to the pose representation rather than differences in training data.

5.2.3 Model Architecture and Training

To isolate the effect of pose representation on joint angle accuracy, we designed the comparison so that the keypoint-based and mesh-based models share the same architecture and training configuration, differing only in their output head (Figure 5.1**)**. Both models use the same two-stage architecture established in Chapter 3. The first stage uses the same RTMW model (Jiang et al. 2024b) we trained in Chapter 3 to estimate 2D pose from the input image. The second stage uses MotionBERT (Zhu et al. 2023) as the backbone to lift 2D pose estimates to 3D. The two models diverge only at the output head after the MotionBert backbone: the keypoint-based model predicts 3D keypoint coordinates, while the mesh-based model predicts SMPL pose and shape parameters. Both models were trained on the paired dataset described in Section 5.2.2, using identical training configurations including learning rate, number of training epochs, and optimizer. This controlled setup ensures that any observed difference in joint angle accuracy is attributable to the choice of pose representation.

5.3 Evaluation of Pose Representations on Joint Angle Accuracy

We evaluated both trained models on the test set of the VEHS-6.7M dataset by computing 22 joint angles (14 unique angle types across left and right sides) from each model's 3D output using the angle definitions established in Chapter 2, and comparing them against the corresponding joint angles computed from motion capture ground truth. We report the mean absolute error (MAE) for each joint angle, averaged across all frames and subjects. To test whether the two representations produce significantly different errors, we computed the MAE per video sequence for each model and performed two-tailed paired t-tests for each of the 14 joint angle types. We applied Bonferroni correction to account for the 14 simultaneous comparisons, yielding an adjusted significance level of $\alpha = 0.05/14 = 3.6e\text{-}3$. We reported the MAE of both models, the

absolute and percentage differences, and the p-values in Table 5.2, and visualized the error distributions in Figure 5.2.

Table 5.2. Joint angle MAE comparison between keypoint-based and mesh-based models.

Angle Name		Keypoint Model	Mesh Model	Difference	Difference (%)	P Value
Neck	Flexion/extension	6.4°	13.1°	6.7°	51%	< 3.6e-3
	Lateral bend	3.5°	6.0°	2.5°	42%	< 3.6e-3
	Rotation	5.2°	15.9°	10.7°	67%	< 3.6e-3
Shoulder	Flexion/extension	4.7°	11.6°	6.9°	60%	< 3.6e-3
	Abduction/adduction	5.5°	10.4°	4.9°	47%	< 3.6e-3
	Rotation	10.2°	17.0°	6.8°	40%	< 3.6e-3
Elbow	Flexion	4.6°	8.6°	3.9°	46%	< 3.6e-3
Wrist	Flexion/extension	8.3°	11.4°	3.1°	27%	< 3.6e-3
	Radial/ulnar deviation	6.6°	7.4°	0.8°	11%	< 3.6e-3
	Pronation/supination	17.4°	23.3°	5.9°	25%	< 3.6e-3
Back	Flexion/extension	2.3°	4.0°	1.8°	44%	< 3.6e-3
	Lateral flexion	5.4°	7.2°	1.8°	24%	8.5E-02
	Rotation	3.8°	5.0°	1.3°	25%	< 3.6e-3
Knee	Flexion	3.5°	6.5°	3.0°	46%	< 3.6e-3
	Overall	**6.3°**	**10.5°**	**4.3°**	**41%**	< 3.6e-3

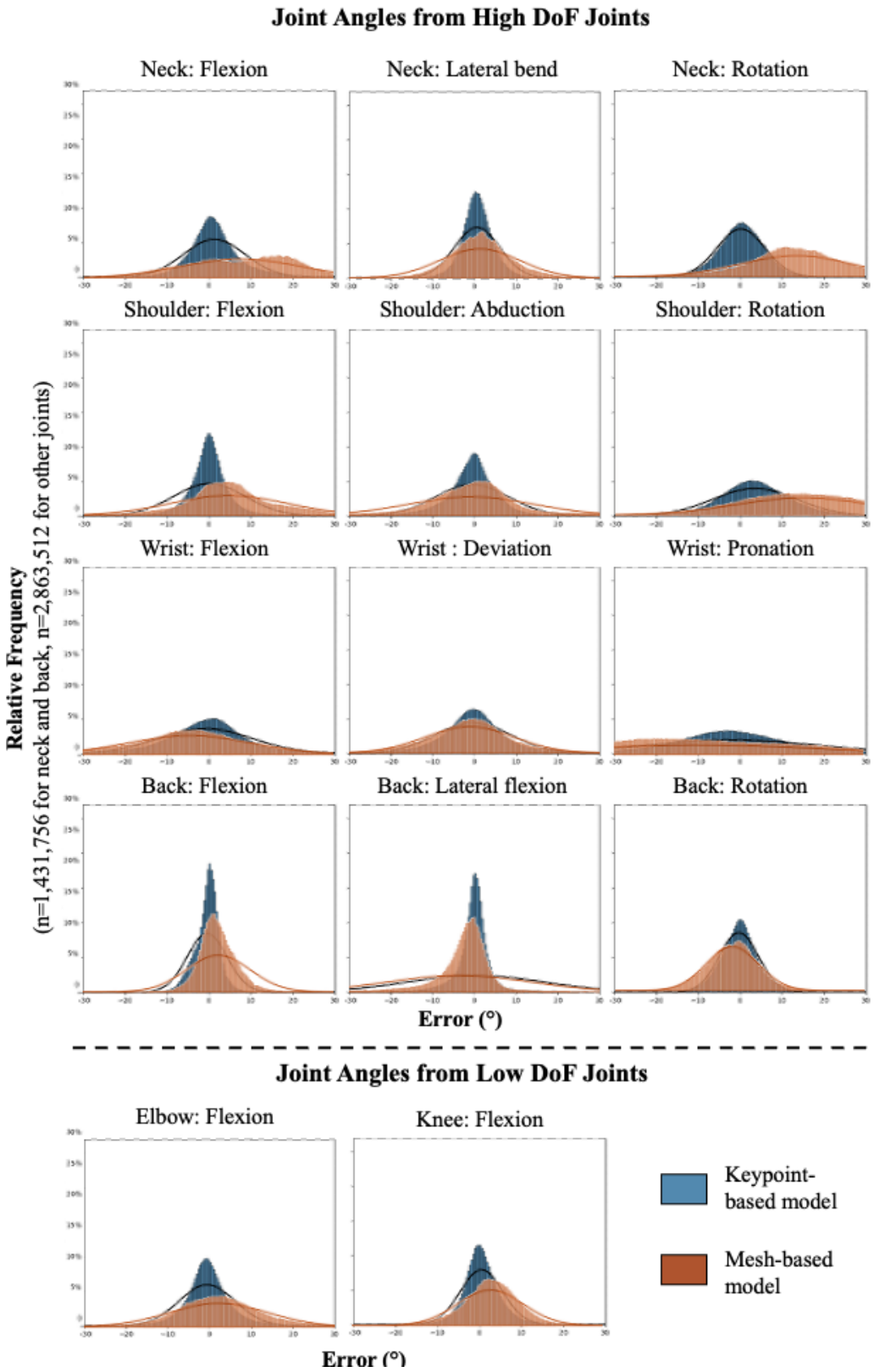


Figure 5.2. Histograms of angle estimation errors with RGB video input, for both the keypoint-based and mesh-based models.

Based on these results, we rejected the null hypothesis for 13 of the 14 joint angles. The keypoint-based model produced significantly more accurate joint angles than the mesh-based model across nearly all degrees of freedom relevant to WMSD risk assessment, when both models shared the same training data, backbone architecture, and training configuration.

The keypoint-based model achieved lower MAE than the mesh-based model on all 14 joint angles. The overall MAE was 6.3° for the keypoint-based model compared to 10.5° for the mesh-based model, a difference of 4.3° (41%). This pattern is also visible in the error distributions in Figure 5.2, where the mesh-based model consistently shows lower peaks and wider spreads, as well as larger bias from zero. This difference was statistically significant ($p < 3.6e\text{-}3$) for 13 of the 14 joint angles. The only exception was back lateral flexion ($p = 8.5e\text{-}2$), though the keypoint-based model still showed numerically lower error (5.4° vs. 7.2°).

The largest absolute errors for the mesh-based model appeared in joint angles involving rotation along the limb axis: wrist pronation/supination (23.3°), shoulder rotation (17.0°), and neck rotation (15.9°). These axial rotations are inherently difficult to estimate from visual input because the limb's visual appearance changes minimally as it rotates around its longitudinal axis.

Both models achieved their lowest errors on back flexion/extension (2.3° and 4.0°) and knee flexion (3.5° and 6.5°), which are primary sagittal-plane angles with large and well-defined ranges of motion during manual material handling tasks. The smallest differences appeared in wrist radial/ulnar deviation (0.8°, 11%), back rotation (1.3°, 25%), and back lateral flexion (1.8°, 24%). For these degrees of freedom, the two representations performed more comparably, though the keypoint-based model maintained a consistent advantage.

5.4 Discussion

The keypoint-based model significantly outperformed the mesh-based model across nearly all joint angles, despite sharing the same training data, backbone architecture, and training configuration. A likely explanation is that the keypoint-based model receives direct supervision on the exact anatomical landmarks used for joint angle calculation, whereas the mesh-based model is trained to predict the full body surface geometry, and the anatomical landmarks are extracted post-hoc from mesh vertices. Predicting the positions of 6,890 mesh vertices to reconstruct the entire body surface is a substantially more complex regression task than predicting 66 targeted keypoint coordinates, and this additional complexity does not contribute to joint angle accuracy at the specific degrees of freedom needed for WMSD risk assessment. The accuracy gap is particularly pronounced for rotational angles such as shoulder rotation (6.8° difference), neck rotation (10.7° difference), and wrist pronation/supination (5.9° difference). These axial rotations produce minimal change in the body surface shape, making them inherently difficult to recover from a surface mesh representation. In contrast, the keypoint-based model can capture these rotations through the relative displacement of targeted surface markers placed specifically to be sensitive to such movements. This result highlights a tradeoff between the two representations: mesh-based models offer greater flexibility for adapting to different downstream tasks without retraining, but this generality comes at the cost of lower accuracy on the specific joint angles required for WMSD risk assessment.

A further source of error in the mesh-based approach arises from the correspondence between the SMPL mesh surface and the human body. The SMPL mesh is a learned statistical approximation of the body surface, and anatomical surface landmarks placed on a real human body may not move identically to their corresponding vertices on the mesh during motion. Soft tissue

deformation, skin sliding over underlying bone, and individual anatomical variation all contribute to discrepancies between marker trajectories on a real body and vertex trajectories on the SMPL surface. These discrepancies introduce an additional layer of error when extracting landmark positions from the predicted mesh for joint angle calculation, an error source that the keypoint-based model avoids entirely by predicting landmark positions directly.

Our controlled experiment isolated the effect of pose representation by fixing all other variables. However, in practice, mesh-based pose estimation models are often trained on different large-scale open-source datasets with different model architectures. To assess whether the observed accuracy gap generalizes beyond our specific training setup, we also evaluated SMPLest-X (Yin et al. 2025), a state-of-the-art one-stage image-to-mesh model selected for its leading performance on mesh estimation benchmarks. We applied SMPLest-X, pretrained on its original open-source datasets, to our test set and computed joint angles using the same vertex-to-landmark mapping and angle definitions. SMPLest-X produced an overall MAE of 12.2°, slightly worse than our mesh model trained on the VEHS-6.7M dataset (10.5°) and substantially worse than the keypoint-based model (6.3°). This result suggests that the accuracy gap between mesh-based and keypoint-based representations for joint angle estimation is not an artifact of our particular training setup but reflects a more general characteristic of the mesh representation itself.

5.5 Conclusion

This chapter presented a quantitative comparison between keypoint-based and mesh-based pose representations for joint angle estimation accuracy in WMSD risk assessment. We trained both models on the same dataset with the same backbone architecture and training configuration, differing only in the output representation, and evaluated joint angle errors against motion capture ground truth across 22 joint angles (14 unique angle types across left and right sides) used in

WMSD risk assessments. The keypoint-based model achieved significantly lower errors than the mesh-based model on 13 of the 14 angle types, with an overall MAE of 6.3° compared to 10.5° (41% reduction). This finding was further supported by the evaluation of SMPLest-X, a leading open-source mesh recovery model, which produced a similar overall MAE of 12.2°. These results indicate that keypoint-based representations provide improved joint angle accuracy for WMSD risk assessment, though at the cost of reduced flexibility to adapt to new joint angle definitions or downstream tasks. For practitioners and researchers who prioritize accurate joint angle estimation, keypoint-based models trained on targeted anatomical landmarks remain the stronger choice. For applications that require frequent adaptation to different assessment protocols or retargeting to other domains, mesh-based representations offer a practical alternative despite their lower angular accuracy.

Chapter 6 Conclusions and Recommendations

6.1 Summary of research

This research pursued two overarching goals: (1) to develop an ergonomics-oriented human motion dataset and a computational pipeline for accurate high-DoF joint angle estimation from monocular video for WMSD risk assessment; and (2) to develop a simulation-driven computational framework for generating task- and site-specific ergonomics intervention recommendations through risk-reduced motion synthesis. With these goals in mind, this research addressed four specific objectives as follows: (1) to construct and evaluate an ergonomics-oriented 3D human motion dataset that captures richer keypoint representations for high-DoF joint angle estimation; (2) to develop and evaluate a computational pipeline that derives accurate high-DoF joint angles from monocular video under real-world deployment conditions; (3) to develop and evaluate a motion synthesis method that reduces WMSD risks in synthesized human motions; and (4) to develop and evaluate a computational framework that integrates simulation-driven motion synthesis into a task- and site-specific ergonomics intervention recommendation framework. A summary of these studies' results and implications is as follows.

1. **Ergonomics-oriented 3D human motion dataset (Chapter 2):** This study defined a set of 66 3D body keypoints and derived joint angle calculation methods that provide sufficient 3D spatial information for calculating 22 angles across 6 body joints, including 4 high-DoF joints (neck, shoulder, back, and wrist) that are challenging to calculate using the sparse joint center representations typically adopted by prior vision-based WMSD risk assessment methods.

Using these designs, a large-scale 3D human motion dataset, VEHS-6.7M, containing 6.7 million frames across nine MMH and assembly task categories, was collected via marker-based motion capture. Evaluation of a baseline 3D human pose estimation model trained on the dataset with ground truth 2D pose input demonstrated accurate joint angle estimation, achieving an angle MAE of 2.4°. The proposed keypoint set and angle calculation methods serve as a foundation for training vision-based pose estimation models tailored for WMSD risk assessment.

2. **Vision-based WMSD risk assessment computational pipeline (Chapter 3):** This study presented an end-to-end computational pipeline for estimating high-DoF joint angles from monocular video for WMSD risk assessment, with specific adaptations to address real-world deployment conditions. The pipeline adopted a two-stage architecture (image-to-2D and 2D-to-3D). To improve the pipeline's ability to generalize to challenging real-world visual conditions, transfer learning with backbone freezing and merge training with diverse open-source datasets were incorporated. A camera pitch angle correction method was also introduced to account for varying camera viewpoints encountered in real-world recordings. Quantitative evaluation on laboratory test set videos demonstrated a 2D PCK@0.2 of 96.4%, 3D MPJPE of 16.7 mm, and an overall joint angle MAE of 5.9° across 22 high-DoF joint angles. Qualitative evaluation on 68 real-world MMH videos sourced from online repositories and industry partners showed that the pipeline produced visually plausible pose results under field conditions not represented in the training data, supporting its applicability for WMSD risk assessment in construction.
3. **Risk-reduced motion synthesis with ergonomics reward (Chapter 4):** This study developed a custom ergonomics reward that leverages established ergonomics guidelines, imitation

learning from curated safe motion examples, and tacit knowledge captured from ergonomics experts through natural language feedback. Across 12 unconstrained and 4 site-constrained MMH task scenarios, the synthesized motions achieved an average 39% reduction in NIOSH Lifting Index and 11% reduction in L4/L5 back compression force compared to captured task motions, with statistically significant differences observed for both metrics. These results demonstrate the potential of the custom ergonomics rewards for producing risk-reduced motion alternatives for workers in MMH tasks.

4. **Computational framework for task- and site-specific intervention recommendation (Chapter 4):** Chapter 4 also introduced a computational framework that enables site personnel to specify task constraints through natural language descriptions and capture site constraints using smartphone-based LiDAR, translating real-world conditions into the physics-based simulation environment for motion synthesis. The framework supports two types of ergonomics interventions: the synthesized risk-reduced motions serve as posture demonstrations for workers, and when posture improvements alone are insufficient, the framework performs automated "what-if" analysis by modifying task or workstation parameters until an acceptable level of WMSD risk is achieved. This framework provides site personnel with accessible means to generate task- and site-specific ergonomics interventions without requiring specialized expertise.
5. **Quantitative evaluation of human pose representations for WMSD risk assessment (Chapter 5):** Mesh-based human pose estimation has emerged as a promising alternative for vision-based WMSD risk assessment, as mesh-based models capture versatile motion data that enables flexible adaptation to new downstream tasks without collecting additional datasets with new keypoint annotations. To evaluate whether mesh-based representations can achieve

comparable accuracy in joint angle estimation, this study conducted a quantitative comparison between keypoint-based and mesh-based approaches. Across all joint angles, the keypoint-based model achieved an overall MAE of 6.3°, compared to 10.6° for the mesh-based model trained under identical conditions, a 4.3° (41%) difference. The difference was statistically significant for 13 of the 14 joint angles evaluated. While mesh-based representations offer greater flexibility, the results indicate that keypoint-based representations currently provide more accurate joint angle estimation for WMSD risk assessment applications.

6.2 Contributions

This dissertation makes several contributions to the body of knowledge in vision-based WMSD risk identification and simulation-driven intervention recommendation. To address the insufficient 3D keypoint in existing human pose datasets for high-DoF joint angle estimation, we designed and collected a large-scale motion capture dataset annotated with 66 body surface landmarks and joint center keypoints, specifically tailored to support angle calculations in WMSD risk assessments. Building on this dataset, we developed a two-stage monocular motion capture pipeline with specific adaptations to account for challenging real-world conditions, including occlusion, varying camera viewpoints, and cluttered backgrounds. Beyond risk identification, we proposed a simulation-driven framework that synthesizes risk-reduced motions using a custom ergonomics reward function and translates real-world task and site conditions into the simulation environment using accessible devices. To our knowledge, this is the first framework that enables the generation of visualized, task- and site-specific interventions, such as posture adjustments and workstation redesign, for WMSD risk mitigation.

From a practical standpoint, the vision-based WMSD risk assessment pipeline and the intervention recommendation framework only require inputs from accessible devices such as a

smartphone. This enables site personnel, including workers and managers, to access ergonomics guidance without the typically resource-intensive process of manual risk identification and intervention design. Ultimately, equipping site personnel with these tools has the potential to shift WMSD risk management from a reactive, resource-intensive process to a proactive practice integrated into daily construction operations.

6.3 Future research

This dissertation focused on vision-based WMSD risk identification and simulation-driven intervention recommendation for construction workers. Several promising research directions emerge from this work.

Can human motion prediction enable proactive safety risk mitigation? This dissertation identified WMSD risks from previously captured motions. A natural extension is to predict human motions before they occur, enabling proactive interventions such as alerting workers before they adopt a hazardous posture or warning workers approaching a danger zone of potential struck-by incidents. Such predictive capabilities would shift WMSD risk management from retrospective assessment to real-time prevention.

Can simulation-driven motion demonstrations be applied to broader construction skill training? This dissertation generated visual demonstrations of correct MMH techniques as ergonomics interventions. The same approach of synthesizing task- and site-specific motion demonstrations may be applicable to broader skill training in construction, where workers need to learn correct techniques for tasks such as scaffolding assembly, formwork installation, or equipment operation under varying site conditions.

Can the tacit knowledge capture framework generalize to other expert-dependent tasks in construction? The framework developed in this dissertation captures and stores expert

tacit knowledge through structured interactions between domain experts and a reinforcement learning system. This approach may be applicable to other construction tasks where correct techniques are traditionally transferred through hands-on mentorship, such as welding, rebar tying, or crane signaling.

Bibliography

Andriluka, M., L. Pishchulin, P. Gehler, and B. Schiele. 2014. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anthropic. 2026. *System Card: Claude Opus 4.6*.

Araki, T., Y. Mukuta, T. Osa, and T. Harada. 2025. "Few-shot Imitation Learning by Variable-Length Trajectory Retrieval from a Large and Diverse Dataset." In: *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, 829–836.

Araujo, J. P., J. Li, K. Vetrivel, R. Agarwal, D. Gopinath, J. Wu, A. Clegg, and C. K. Liu. 2023. "CIRCLE: Capture In Rich Contextual Environments." arXiv.

Badler, N. I., M. S. Palmer, and R. Bindiganavale. 1999. "Animation control for real-time virtual humans." *Commun. ACM*, 42 (8): 64–73. https://doi.org/10.1145/310930.310975.

Baer, J. L., and R. G. Cohen. 2023. "Does neck flexion improve performance? Effects on reaction time depend on whether responses are expected." *SN Applied Sciences*, 5 (4): 106. https://doi.org/10.1007/s42452-023-05335-6.

Bertaux, A., M. Gueugnon, F. Moissenet, B. Orliac, P. Martz, J.-F. Maillefert, P. Ornetti, and D. Laroche. 2022. "Gait analysis dataset of healthy volunteers and patients before and 6 months after total hip arthroplasty." *Scientific Data*, 9: 399. https://doi.org/10.1038/s41597-022-01483-3.

Bhatnagar, B. L., X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. 2022a. "BEHAVE: Dataset and Method for Tracking Human Object Interactions." arXiv.

Bhatnagar, B. L., X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. 2022b. "BEHAVE: Dataset and Method for Tracking Human Object Interactions." arXiv.

Blender Foundation. 2026. "Blender - The Free and Open Source 3D Creation Software."

BLS. 2022. "Workplace Injuries and Job Requirements for Construction Laborers : Spotlight on Statistics : U.S. Bureau of Labor Statistics." *Bureau of Labor Statistics*. Accessed September 17, 2025. https://www.bls.gov/spotlight/2022/workplace-injuries-and-job-requirements-for-construction-laborers/home.htm.

BLS. 2025a. "Table 18. Annual average hires rates by industry and region, not seasonally adjusted - 2025 M01 Results." *Bureau of Labor Statistics*. Accessed September 5, 2025. https://www.bls.gov/news.release/jolts.t18.htm.

BLS. 2025b. "Table 20. Annual average total separations rates by industry and region, not seasonally adjusted - 2025 M01 Results." *Bureau of Labor Statistics*. Accessed September 5, 2025. https://www.bls.gov/news.release/jolts.t20.htm.

Bogo, F., A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. 2016. "Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image." *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling, 561–578. Cham: Springer International Publishing.

Bonakdar, A., N. Riahi, M. Shakourisalim, L. Miller, M. Tavakoli, H. Rouhani, and A. Golabchi.

2025. “Validation of markerless vision-based motion capture for ergonomics risk assessment.” *International Journal of Industrial Ergonomics*, 107: 103734. https://doi.org/10.1016/j.ergon.2025.103734.

Bureau of Labor Statistics. 2023. “Employer-Reported Workplace Injuries and Illnesses - 2021-2022.”

Cen, Z., H. Pi, S. Peng, Z. Shen, M. Yang, S. Zhu, H. Bao, and X. Zhou. 2024. “Generating Human Motion in 3D Scenes from Text Descriptions.” In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1855–1866. Seattle, WA, USA: IEEE.

Chen, W., D. Gu, and J. Ke. 2024. “Real-time ergonomic risk assessment in construction using a co-learning-powered 3D human pose estimation model.” *Computer-Aided Civil and Infrastructure Engineering*, 39 (9): 1337–1353. https://doi.org/10.1111/mice.13139.

Choi, S. 2012. “A study of trade-specific occupational ergonomics considerations in the U.S. construction industry.” *Work (Reading, Mass.)*, 42: 215–22. https://doi.org/10.3233/WOR-2012-1344.

Choi, S. D., L. Yuan, and J. G. Borchardt. 2016. “Musculoskeletal disorders in construction: practical solutions from the literature.” *Professional Safety*, 61(01), 26–32.

Chu, W., S. Han, X. Luo, and Z. Zhu. 2020. “Monocular Vision–Based Framework for Biomechanical Analysis or Ergonomic Posture Assessment in Modular Construction.” *Journal of Computing in Civil Engineering*, 34 (4): 04020018. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000897.

Ciriello, V. M., S. H. Snook, L. Hashemi, and J. Cotnam. 1999. “Distributions of manual materials handling task parameters.” *International Journal of Industrial Ergonomics*, 24 (4): 379–388. https://doi.org/10.1016/S0169-8141(99)00005-0.

Clegg, A., W. Yu, J. Tan, C. K. Liu, and G. Turk. 2018. “Learning to dress: synthesizing human dressing motion via deep reinforcement learning.” *ACM Transactions on Graphics*, 37 (6): 1–10. https://doi.org/10.1145/3272127.3275048.

CPWR. 2025. “CPWR Fatal and Nonfatal Injuries.” Accessed September 12, 2025. https://www.cpwr.com/research/data-center/the-construction-chart-book/interactive-7th/injuries-illnesses-health/fatal-and-nonfatal-injuries/.

Diego-Mas, J. A. 2023. “Postural Assessment Using The OWAS Method.” *Ergonautas, Universidad Politécnica de Valencia*. Accessed December 2, 2024. https://www.ergonautas.upv.es/ergoniza/app_en/land/index.html?method=owas.

Duan, H., K. Lin, S. Jin, W. Liu, C. Qian, and W. Ouyang. 2019. “TRB: A Novel Triplet Representation for Understanding 2D Human Body.” arXiv.

Elite Force Safety. 2025. “Manual Handling: Tailored Health & Safety Training.” Accessed September 13, 2025. https://www.eliteforcesafety.com/manualhandling.

ErgoPlus. 2024. “ErgoPlus Industrial – Next Generation Ergonomics and Injury Prevention Software.” Accessed November 22, 2024. https://ergo-plus.com/software/.

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

Ghorbani, N., and M. J. Black. 2021. “SOMA: Solving Optical Marker-Based MoCap Automatically.” arXiv.

Guillot, A., F. D. Rienzo, C. Frank, U. Debarnot, and T. E. MacIntyre. 2024. “From simulation to motor execution: a review of the impact of dynamic motor imagery on performance.” *International Review of Sport and Exercise Psychology*, 17 (1): 420–439.

https://doi.org/10.1080/1750984X.2021.2007539.
Hassan, M., V. Choutas, D. Tzionas, and M. J. Black. 2019. “Resolving 3D Human Pose Ambiguities with 3D Scene Constraints.” arXiv.
Hignett, S., and L. McAtamney. 2000. “Rapid Entire Body Assessment (REBA).” *Applied Ergonomics*, 31 (2): 201–205. https://doi.org/10.1016/S0003-6870(99)00039-3.
Huang, C., W. Kim, Y. Zhang, and S. Xiong. 2020. “Development and Validation of a Wearable Inertial Sensors-Based Automated System for Assessing Work-Related Musculoskeletal Disorders in the Workspace.” *International Journal of Environmental Research and Public Health*, 17 (17): 6050. https://doi.org/10.3390/ijerph17176050.
Huang, C.-H. P., H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black. 2022. “Capturing and Inferring Dense Full-Body Human-Scene Contact.” arXiv.
Ionescu, C., D. Papava, V. Olaru, and C. Sminchisescu. 2014. “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (7): 1325–1339. https://doi.org/10.1109/TPAMI.2013.248.
Jeong, S., and J. Kook. 2023. “CREBAS: Computer-Based REBA Evaluation System for Wood Manufacturers Using MediaPipe.” *Applied Sciences*, 13 (2): 938. https://doi.org/10.3390/app13020938.
Jiang, N., Z. He, Z. Wang, H. Li, Y. Chen, S. Huang, and Y. Zhu. 2024a. “Autonomous Character-Scene Interaction Synthesis from Text Instruction.” arXiv.
Jiang, T., X. Xie, and Y. Li. 2024b. “RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation.” arXiv.
Jiang, Z., Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. J. Fan, and Y. Zhu. 2025. “DexMimicGen: Automated Data Generation for Bimanual Dexterous Manipulation via Imitation Learning.” In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 16923–16930.
Jin, S., L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo. 2020. “Whole-Body Human Pose Estimation in the Wild.” arXiv.
Joo, H., H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. 2015. “Panoptic Studio: A Massively Multiview System for Social Motion Capture.” In: *2015 IEEE International Conference on Computer Vision (ICCV)*, 3334–3342. Santiago, Chile: IEEE.
Karhu, O., P. Kansi, and I. Kuorinka. 1977. “Correcting working postures in industry: A practical method for analysis.” *Applied Ergonomics*, 8 (4): 199–201. https://doi.org/10.1016/0003-6870(77)90164-8.
Kaufmann, M., J. Song, C. Guo, K. Shen, T. Jiang, C. Tang, J. J. Zárate, and O. Hilliges. 2023. “EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild.” In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14586–14597. Paris, France: IEEE.
Kee, D., M. K. Chung, and J.-H. Kim. 2011. “Legal system and its effect for prevention of work-related musculoskeletal disorders in Korea.” *International Journal of Industrial Ergonomics*, 41 (3): 224–232. https://doi.org/10.1016/j.ergon.2011.02.001.
Kudo, Y., K. Ogaki, Y. Matsui, and Y. Odagiri. 2018. “Unsupervised Adversarial Learning of 3D Human Pose from 2D Joint Locations.” arXiv.
Kwon, Y.-J., D.-H. Kim, B.-C. Son, K.-H. Choi, S. Kwak, and T. Kim. 2022. “A Work-Related

Musculoskeletal Disorders (WMSDs) Risk-Assessment System Using a Single-View Pose Estimation Model." *International Journal of Environmental Research and Public Health*, 19 (16): 9803. https://doi.org/10.3390/ijerph19169803.

Ladda, A. M., F. Lebon, and M. Lotze. 2021. "Using motor imagery practice for improving motor performance – A review." *Brain and Cognition*, 150: 105705. https://doi.org/10.1016/j.bandc.2021.105705.

Lassner, C., J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. 2017. "Unite the People: Closing the Loop Between 3D and 2D Human Representations." arXiv.

Li, X., S. Han, M. Gül, and M. Al-Hussein. 2019. "Automated post-3D visualization ergonomic analysis system for rapid workplace design in modular construction." *Automation in Construction*, 98: 160–174. https://doi.org/10.1016/j.autcon.2018.11.012.

Li, Y., S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, and S.-T. Xia. 2022. "SimCC: a Simple Coordinate Classification Perspective for Human Pose Estimation." arXiv.

Li, Z., R. Zhang, C.-H. Lee, and Y.-C. Lee. 2020. "An Evaluation of Posture Recognition Based on Intelligent Rapid Entire Body Assessment System for Determining Musculoskeletal Disorders." *Sensors*, 20 (16): 4414. https://doi.org/10.3390/s20164414.

Liberty Mutual. 2021. *Liberty Mutual Workplace Safety Index Annual report from Risk Control Services*.

Liberty Mutual. 2022. *Liberty Mutual Workplace Safety Index Annual report from Risk Control Services*.

Loper, M., N. Mahmood, and M. J. Black. 2014. "MoSh: motion and shape capture from sparse markers." *ACM Transactions on Graphics*, 33 (6): 1–13. https://doi.org/10.1145/2661229.2661273.

Loper, M., N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. 2015. "SMPL: a skinned multi-person linear model." *ACM Transactions on Graphics*, 34 (6): 1–16. https://doi.org/10.1145/2816795.2818013.

Loyola University Chicago, E. S. department. 2024. "LUC Guidelines for Safe Lifting Techniques." https://www.luc.edu/media/lucedu/environmentalservices/pdfs/LUC%20Guidelines%20for%20Safe%20Lifting%20Techniques.pdf.

Lugaresi, C., J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. 2019. "MediaPipe: A Framework for Building Perception Pipelines." arXiv.

Ma, X., J. Su, C. Wang, W. Zhu, and Y. Wang. 2023. "3D Human Mesh Estimation From Virtual Markers."

Mahmood, N., N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black. 2019. "AMASS: Archive of Motion Capture As Surface Shapes." In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5441–5450. Seoul, Korea (South): IEEE.

Makoviychuk, V., L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State. 2021. "Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning." arXiv.

McAtamney, L., and E. N. Corlett. 1993. "RULA: a survey method for the investigation of work-related upper limb disorders." *Applied Ergonomics*, 24 (2): 91–99. https://doi.org/10.1016/0003-6870(93)90080-S.

Mehta, D., H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. 2017. "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision." arXiv.

NIOSH. 2007. *Ergonomic guidelines for manual material handling.* U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health.

OpenAI. 2025. "Introducing GPT-5." Accessed September 13, 2025. https://openai.com/index/introducing-gpt-5/.

Pan, L., J. Wang, B. Huang, J. Zhang, H. Wang, X. Tang, and Y. Wang. 2024. "Synthesizing Physically Plausible Human Motions in 3D Scenes." In: *2024 International Conference on 3D Vision (3DV)*, 1498–1507.

Pan, L., Z. Yang, Z. Dou, W. Wang, B. Huang, B. Dai, T. Komura, and J. Wang. 2025. "TokenHSI: Unified Synthesis of Physical Human-Scene Interactions through Task Tokenization." arXiv.

Paterson, K. L., R. S. Hinman, B. R. Metcalf, K. L. Bennell, and T. V. Wrigley. 2017. "Plug-in-Gait calculation of the knee adduction moment in people with knee osteoarthritis during shod walking: comparison of two different foot marker models." *Journal of Foot and Ankle Research*, 10: 8. https://doi.org/10.1186/s13047-017-0187-4.

Paudel, P., Y.-J. Kwon, D.-H. Kim, and K.-H. Choi. 2022. "Industrial Ergonomics Risk Analysis Based on 3D-Human Pose Estimation." *Electronics*, 11 (20): 3403. https://doi.org/10.3390/electronics11203403.

Pavlakos, G., V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. 2019. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image." arXiv.

Peng, X. B., Y. Guo, L. Halper, S. Levine, and S. Fidler. 2022. "ASE: Large-Scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters." *ACM Transactions on Graphics*, 41 (4): 1–17. https://doi.org/10.1145/3528223.3530110.

Peng, X. B., Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. 2021. "AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control." *ACM Transactions on Graphics*, 40 (4): 1–20. https://doi.org/10.1145/3450626.3459670.

Peng, X., Y. Xie, Z. Wu, V. Jampani, D. Sun, and H. Jiang. 2024. "HOI-Diff: Text-Driven Synthesis of 3D Human-Object Interactions using Diffusion Models." arXiv.

Plantard, P., H. P. H. Shum, A.-S. Le Pierres, and F. Multon. 2017. "Validation of an ergonomic assessment method using Kinect data in real workplace conditions." *Applied Ergonomics*, 65: 562–569. https://doi.org/10.1016/j.apergo.2016.10.015.

Polycam. 2022. "Polycam - LiDAR & 3D Scanner for iPhone & Android." Accessed October 7, 2022. https://poly.cam/.

Ryu, J., M. M. Diraneyya, C. T. Haas, and E. Abdel-Rahman. 2021. "Analysis of the Limits of Automated Rule-Based Ergonomic Assessment in Bricklaying." *Journal of Construction Engineering and Management*, 147 (2): 04020163. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001978.

Salehi, M., A. Taheri, S. Choi, and J. H. Kim. 2026. "Evaluation of a markerless motion capture to measure 3D joint kinematics during occupational lifting tasks using mobile devices." *Applied Ergonomics*, 134: 104743. https://doi.org/10.1016/j.apergo.2026.104743.

Schall Jr., M. C., N. B. Fethke, and V. Roemig. 2018. "Digital Human Modeling in the Occupational Safety and Health Process: An Application in Manufacturing." *IISE Transactions on Occupational Ergonomics and Human Factors*, 6 (2): 64–75. Taylor & Francis. https://doi.org/10.1080/24725838.2018.1491430.

Schwartz, A. H., T. J. Albin, and S. G. Gerberich. 2019. "Intra-rater and inter-rater reliability of the rapid entire body assessment (REBA) tool." *International Journal of Industrial*

Ergonomics, 71: 111–116. https://doi.org/10.1016/j.ergon.2019.02.010.
Seo, J., and S. Lee. 2021. "Automated postural ergonomic risk assessment using vision-based posture classification." *Automation in Construction*, 128: 103725. https://doi.org/10.1016/j.autcon.2021.103725.
Sers, R., S. Forrester, E. Moss, S. Ward, J. Ma, and M. Zecca. 2020. "Validity of the Perception Neuron inertial motion capture system for upper body motion analysis." *Measurement*, 149: 107024. https://doi.org/10.1016/j.measurement.2019.107024.
Sigal, L., A. O. Balan, and M. J. Black. 2010. "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion." *International Journal of Computer Vision*, 87 (1–2): 4–27. https://doi.org/10.1007/s11263-009-0273-6.
Sundareshan, B. 2024. *Evaluation of 3D Human Pose Estimation*.
Trumble, M., A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. 2017. "Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors." In: *Procedings of the British Machine Vision Conference 2017*, 14. London, UK: British Machine Vision Association.
Van Crombrugge, I., S. Sels, B. Ribbens, G. Steenackers, R. Penne, and S. Vanlanduit. 2022. "Accuracy Assessment of Joint Angles Estimated from 2D and 3D Camera Measurements." *Sensors*, 22 (5): 1729. https://doi.org/10.3390/s22051729.
Varcin, F., and M. G. Boocock. 2026. "The accuracy, validity and reliability of Theia3D markerless motion capture for studying the biomechanics of human movement: A systematic review." *Artificial Intelligence in Medicine*, 173: 103332. https://doi.org/10.1016/j.artmed.2025.103332.
VelocityEHS. 2024a. "All-in-One Industrial Ergonomics Solution." *VelocityEHS*. Accessed November 22, 2024. https://www.ehs.com/solutions/ergonomics/industrial/.
VelocityEHS. 2024b. "3D SSPP - 3D Static Strength Prediction Program." *VelocityEHS*. Accessed January 12, 2024. https://www.ehs.com/solutions/ergonomics/3d-sspp/.
Vicon. 2021. "Plug-in Gait Reference Guide."
Vicon. 2022. "Recommended FLIR Blackfly S video hardware for Nexus - Nexus 2.13 Documentation - Vicon Documentation." Accessed September 21, 2022. https://docs.vicon.com/display/Nexus213/Recommended+FLIR+Blackfly+S+video+hardware+for+Nexus.
Vicon. 2023. "Vero | Compact Super Wide Camera by Vicon." *Vicon*. Accessed September 8, 2023. https://www.vicon.com/hardware/cameras/vero/.
Vijayakumar, R., and J. Choi. 2022. "Emerging Trends of Ergonomic Risk Assessment in Construction Safety Management: A Scientometric Visualization Analysis." *International Journal of Environmental Research and Public Health*, 19 (23): 16120. https://doi.org/10.3390/ijerph192316120.
Von Marcard, T., R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. 2018. "Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera." *Computer Vision – ECCV 2018*, edited by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, 614–631. Cham: Springer International Publishing.
Wang, Z., Y. Chen, B. Jia, P. Li, J. Zhang, J. Zhang, T. Liu, Y. Zhu, W. Liang, and S. Huang. 2024. "Move as you Say, Interact as you can: Language-Guided Human Motion Generation with Scene Affordance." In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 433–444. Seattle, WA, USA: IEEE.
Waters, T. R., V. Putz-Anderson, and A. Garg. 2021. *Applications manual for the revised NIOSH*

lifting equation. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health.

Wen, L., D. Kim, V. Talreja, M. Liu, J. Penfield, R. Barker, and S. Lee. 2026. "3D human pose keypoints and corresponding joint angle calculation for vision-based WMSD risk assessments." *Ergonomics*, 1–18. https://doi.org/10.1080/00140139.2026.2639614.

Wen, L., G. Yong, F. Baek, and S. Lee. 2024. "Text-to-Motion Generation for Construction Worker's Posture Coaching Applying Stochastic Ergonomic Risk Minimization." In: *ASCE International Conference on Computing in Civil Engineering*. Pittsburgh, PA.

Wu, G., F. C. T. van der Helm, H. E. J. (DirkJan) Veeger, M. Makhsous, P. Van Roy, C. Anglin, J. Nagels, A. R. Karduna, K. McQuade, X. Wang, F. W. Werner, and B. Buchholz. 2005. "ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand." *Journal of Biomechanics*, 38 (5): 981–992. https://doi.org/10.1016/j.jbiomech.2004.05.042.

Wu, G., S. Siegler, P. Allard, C. Kirtley, A. Leardini, D. Rosenbaum, M. Whittle, D. D. D'Lima, L. Cristofolini, H. Witte, O. Schmid, and I. Stokes. 2002. "ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine." *Journal of Biomechanics*, 35 (4): 543–548. https://doi.org/10.1016/S0021-9290(01)00222-6.

Wu, W., C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. 2018. "Look at Boundary: A Boundary-Aware Face Alignment Algorithm." arXiv.

Yang, S. T., M. H. Park, and B. Y. Jeong. 2020. "Types of manual materials handling (MMH) and occupational incidents and musculoskeletal disorders (MSDs) in motor vehicle parts manufacturing (MVPM) industry." *International Journal of Industrial Ergonomics*, 77: 102954. https://doi.org/10.1016/j.ergon.2020.102954.

Yin, W., Z. Cai, R. Wang, A. Zeng, C. Wei, Q. Sun, H. Mei, Y. Wang, H. E. Pang, M. Zhang, L. Zhang, C. C. Loy, A. Yamashita, L. Yang, and Z. Liu. 2025. "SMPLest-X: Ultimate Scaling for Expressive Human Pose and Shape Estimation." arXiv.

Yong, G., M. Liu, and S. Lee. 2024. "Explainable Image Captioning to Identify Ergonomic Problems and Solutions for Construction Workers." *Journal of Computing in Civil Engineering*, 38 (4): 04024022. https://doi.org/10.1061/JCCEE5.CPENG-5744.

Yong, G., M. Liu, and S. Lee. 2026. "Pose-Aware Image Captioning for Ergonomic Problem and Solution Identification." *Journal of Construction Engineering and Management*, 152 (5): 04026043. American Society of Civil Engineers. https://doi.org/10.1061/JCEMD4.COENG-17775.

Yu, W., G. Turk, and C. K. Liu. 2018. "Learning symmetric and low-energy locomotion." *ACM Transactions on Graphics*, 37 (4): 1–12. https://doi.org/10.1145/3197517.3201397.

Zhang, S., Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, and S. Tang. 2022. "EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices." arXiv.

Zhang, Y., and S. Tang. 2022. "The Wanderings of Odysseus in 3D Scenes." In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20449–20459. New Orleans, LA, USA: IEEE.

Zhu, W., X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. 2023. "MotionBERT: A Unified Perspective on Learning Human Motion Representations." arXiv.