Evaluating Generative Text-to-Motion for Occupational Wellbeing Recommender Systems

Gaetano Dibenedetto*, Stefano Labianca, Marco Polignano and Pasquale Lops

University of Bari Aldo Moro - Department of Computer Science, Via Orabona 4, Bari, 70125, Italy

Abstract

Wellbeing Recommender Systems are increasingly being applied to occupational health to prevent Musculoskeletal Disorders through targeted behavioral adjustments. While traditional algorithmic systems can generate precise, mathematically safe postural corrections using skeletal interpolation, these visual overlays often lack natural fluidity. Conversely, recent LLM-driven text-to-motion architectures excel at generating highly fluid, semantically plausible human motion from natural language prompts. This paper investigates the critical trade-off between qualitative semantic naturalness and quantitative biomechanical precision in visual ergonomic feedback. We present a comprehensive, multi-phase evaluation comparing state-of-the-art generative models (MotionGPT-3, FineMoGen, MoMask) against a validated, interpolation-based skeletal baseline (SAFELIFT). Our quantitative analysis reveals that current generative models systematically fail to satisfy strict kinematic safety thresholds (e.g., the Revised NIOSH Lifting Equation) in both zero-shot and domain-adapted settings. However, through a Human-in-the-Loop evaluation methodology, we uncover a "productive tension": users strongly prefer generative models for understanding the natural flow and qualitative nature of a movement, while decisively relying on explicit skeletal interpolation for geometric accuracy and trust. In our valuation, current generative models did not demonstrate sufficient biomechanical robustness to replace algorithmic interpolation for high-stakes physical tasks. We propose design guidelines for future occupational WellRec systems utilizing hybrid visual interfaces that maximize both biomechanical safety and user comprehension.

Keywords

Wellbeing Recommender Systems, Generative AI, Text-to-Motion, Occupational Health, Human-Centered Evaluation, Visual Feedback

1. Introduction

Traditional Recommender Systems have historically focused on suggesting digital content, products, or services. However, the emergence of Wellbeing Recommender Systems (WellRec) represents a paradigm shift toward recommending healthy behaviors and lifestyle adjustments. One of the most critical, yet under-explored, domains for such systems is occupational health, specifically the prevention of Musculoskeletal Disorders (MSDs). MSDs remain a leading cause of occupational injury and long-term disability globally, often resulting from improper posture and unsafe lifting mechanics [1]. While recent advancements in Artificial Intelligence and Computer Vision [2, 3] have enabled automated risk assessment, such as using Human Pose Estimation (HPE) to track joint movements and automatically calculate lifting safety standards like the Revised NIOSH Lifting Equation (RNLE) [4], there remains an open research question regarding the optimal way to visually communicate behavioral corrections back to the user. Our novel contribution is to reframe text-to-motion generation as a visual recommender for occupational wellbeing, and to evaluate it jointly from biomechanical safety and user-perception perspectives against a validated skeletal baseline.

In our previous work, we introduced **SAFELIFT** [5], an automated, safety-aware recommender system capable of detecting risky lifting behaviors from monocular video and generating corrective

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden

*Corresponding author.

gaetano.dibenedetto@uniba.it (G. Dibenedetto); s.labianca10@studenti.uniba.it (S. Labianca); marco.polignano@uniba.it (M. Polignano); pasquale.lops@uniba.it (P. Lops)

https://gaetanodibenedetto.github.io/ (G. Dibenedetto)

0000-0001-6083-3600 (G. Dibenedetto); 0009-0002-9280-4446 (S. Labianca); 0000-0002-3939-0136 (M. Polignano); 0000-0002-6866-9451 (P. Lops)

feedback. SAFELIFT successfully computes the Lifting Index and provides automated postural suggestions using kinematic interpolation. This approach provides geometrically constrained and highly precise biomechanical feedback in our system. As a natural follow-up to SAFELIFT, we seek to explore alternative visual formats for delivering these ergonomic recommendations. Recently, LLM-based text-to-motion architectures utilizing advanced motion encoder/decoder systems (such as MotionGPT [6] or MoMask [7]) have shown remarkable capabilities in generating fluid, natural-looking 3D human motion from textual prompts. We hypothesize that integrating these generative models could offer a different modality of feedback, one that translates "contextual user states" (e.g., "A worker safely lifting a 15kg box from the floor while keeping the spine straight") into highly fluid visual recommendations. Adapting these generative models for wellbeing recommendations, however, introduces a critical tension. We present an investigation into the trade-off between *Semantic Plausibility* and *Biomechanical Accuracy*. Current text-to-motion LLMs are trained to produce animations that are visually pleasing and semantically aligned with the input prompt. Conversely, wellbeing applications like RNLE require strict kinematic precision. An exploratory challenge arises: a generative lifting animation might appear highly natural to the user, yet it may deviate from the strict geometric constraints imposed by traditional interpolation methods like SAFELIFT [8].

Because visual recommenders in occupational health must be both trusted and correctly interpreted by workers, computational metrics alone are insufficient to evaluate these systems. We center our research on a Human-in-the-Loop evaluation methodology to understand how users perceive these different feedback formats. We conduct a user study comparing the precise, interpolation-based rendering of SAFELIFT against the fluid, LLM-generated ergonomic feedback. By measuring dimensions such as perceived naturalness, clarity of the recommended movement, and overall user trust, we aim to uncover whether the qualitative naturalness of generative models enhances user comprehension, or if the quantitative precision of interpolation remains superior for safety-critical tasks.

Through this exploratory work, we aim to expand the boundaries of user modeling and personalized recommendation for physical wellbeing. Our main contributions are threefold:

- **Exploration of Visual Formats**: We investigate the application of LLM-based text-to-motion architectures (utilizing motion encoder/decoder systems) as an alternative visual behavioral recommender, expanding upon prior interpolation-based methods.
- **Trade-off Analysis in WellRec**: We provide an empirical analysis of the capabilities and limits of current generative models in wellbeing contexts, mapping the trade-off between qualitative semantic naturalness and quantitative biomechanical precision.
- **Human-Centered Evaluation**: We present a comparative user study demonstrating how different modalities of visual feedback impact user comprehension, naturalness, and trust, offering design guidelines for future ergonomic recommenders.

To facilitate reproducibility, the technical details, additional visualizations, scripts, and prompts used in this study are publicly available in our GitHub[1] repository.

2. Related Work

Our research intersects three distinct domains: health and wellbeing recommender systems, generative text-to-motion architectures, and human-computer interaction for trustworthy AI. In this section, we review the literature in these areas and highlight how our proposed visual feedback format addresses existing gaps.

[1] https://github.com/GaetanoDibenedetto/llm4wellrec

2.1. Visual Feedback in Occupational Wellbeing Recommender Systems

The field of Health Recommender Systems (HRS) and Wellbeing Recommender Systems (WellRec) has traditionally focused on suggesting digital content, dietary plans, or general fitness routines [9, 10, 11, 12, 13]. Recently, there has been a push to apply these systems to occupational health to prevent MSDs. State-of-art automated ergonomic assessment tools often rely on wearable Inertial Measurement Units or multi-camera setups to calculate biomechanical risks like the RNLE [4, 14, 15, 16].

In our previous work, SAFELIFT [5], we demonstrated that accurate RNLE risk assessment and kinematic feedback could be generated using monocular video and HPE without intrusive hardware. However, the visual feedback generated by SAFELIFT relies on algorithmic skeletal interpolation. While mathematically safe, interpolation-based rendering often results in rigid, wireframe-like visual overlays that lack fluidity.

Unlike traditional HRS that output text-based nudges, or systems like SAFELIFT that output rigid skeletal wireframes, we conceptualize *generative text-to-motion models* as adaptive visual recommenders. By treating natural language prompts as contextual user states, we aim to shift the recommender's output from quantitative dashboards and wireframes to fluid, naturalistic human behavioral modeling, introducing a novel format for occupational WellRec.

2.2. Generative Human Motion and the Biomechanical Trade-off

Recent advancements in Generative AI have revolutionized 3D human motion synthesis. Text-to-motion (T2M) architectures typically utilize either Diffusion-based models (e.g., HMDM [17]) or VQ-VAE discrete representations (e.g., T2M-GPT [18], MoMask [7]). These models are predominantly trained on large-scale datasets like HumanML3D [19] to synthesize high-fidelity, diverse motions from natural language descriptions. In standard computer graphics and gaming applications, these models are optimized for semantic alignment (how well the motion matches the text) and visual fidelity (measured via metrics like Fréchet Inception Distance, FID) [19, 7]. Consequently, they often ignore strict physical and environmental constraints, such as gravity, exact joint torques, or the precise geometric angles required for safe load lifting.

While existing literature focuses on the aesthetic and semantic capabilities of T2M models, we benchmark these architectures specifically on *strict biomechanical safety*. We explicitly expose and analyze the trade-off between qualitative semantic naturalness and quantitative biomechanical precision. Our work highlights a critical limitation in current generative models when applied to safety-critical wellbeing tasks, where a visually pleasing animation may still constitute an ergonomic hazard.

2.3. Human-Centered Evaluation in Trustworthy AI

Evaluating generative motion models and recommender systems presents a significant challenge. T2M models are typically evaluated using computational metrics such as R-Precision, Multimodality, and FID [19, 7]. Conversely, standard recommender systems are evaluated using accuracy metrics such as Click-Through Rate or Mean Absolute Error [20].

However, when generating behavioral feedback for human wellbeing, computational metrics are insufficient. A central challenge in intelligent user interfaces is ensuring that feedback is interpretable, trustworthy, and actionable. Explanations, whether textual or visual, strongly influence user trust and understanding [21, 22, 23]. The UMAP community increasingly emphasizes Human-Centered Evaluation, acknowledging that users must interpret the AI's output correctly to enact behavioral change [24].

Recent visual explanation studies demonstrate that users prefer simple, conclusive, and easy-to-interpret formats [25]. For instance, research investigating explanation formats has shown that hybrid *Image+Text* conditions are significantly more helpful and easier to use than text or images alone [26, 27].

Building on this literature, our study investigates not only the impact of static and hybrid visualizations but the role of *generative motion-based feedback*. Moving beyond standard algorithmic benchmarking, we conduct a Human-in-the-Loop evaluation comparing generative T2M feedback against the traditional

skeletal rendering of SAFELIFT. By actively measuring end-user perceptions of naturalness, clarity of instruction, and system trust, our study provides empirical design guidelines for how generative visual recommenders should be presented to maximize comprehension and safety in occupational environments.

3. Methodology and Evaluation Framework

To evaluate whether LLM-driven generative models can effectively replace the skeletal interpolation used in SAFELIFT, we designed a comprehensive evaluation pipeline. Our goal is to determine if these architectures can accurately interpret ergonomic instructions and translate them into biomechanically safe animations.

The pipeline proceeds in three phases: (1) a zero-shot evaluation of pre-trained T2M models, (2) a domain-adaptation phase involving fine-tuning on a custom occupational dataset, and (3) a human-centered user study protocol.

3.1. Phase 1: Zero-Shot Evaluation Protocol

In the first phase, we evaluated three representative T2M architectures without any domain-specific fine-tuning: **MotionGPT-3** [28] (continuous latent representation), **FineMoGen** [29] (diffusion-based representation), and **MoMask** [7] (discrete token representation).

3.1.1. Prompting Strategies for Ergonomic Control

To test the models' ability to generate safe postures, we designed experimental tasks based on common occupational hazards (e.g., reaching overhead). For each model, we applied three progressive prompting strategies:

1. **Qualitative Baseline (Qualitative):** Natural language descriptions without quantitative constraints (e.g., *"a person reaches up..."*).
2. **Direct Quantitative Specifications (Metric):** Prompts enriched with explicit metric values (e.g., *"...reaches up 30 cm above head height"*).
3. **Relative Anatomical References (Anatomical):** Prompts substituting abstract numbers with bodily proportions (e.g., *"...approximately one forearm's length above head"*).

3.1.2. Biomechanical Verification

To measure geometric accuracy, the quantitative evaluation was conducted specifically on the overhead reaching task. For each model, we tested two distinct prompting strategies to define the target extension: *Metric* specifications in centimeters (10, 30, and 60 cm) and *Anatomical* references (hand width $\approx$ 10 cm, forearm length $\approx$ 30 cm, and half arm length $\approx$ 60 cm). We extracted the 3D kinematic parameters directly from the output sequences by tracking the vertical coordinates of the wrist joints relative to the head. We then computed the Mean Absolute Error (MAE) between the generated movement and the quantitative target specified in the prompt.

3.2. Phase 2: Domain Adaptation via Fine-Tuning Protocol

To address potential zero-shot limitations, we designed a fine-tuning pipeline to adapt a generative architecture (MoMask [7]) to the specific domain of ergonomic lifting.

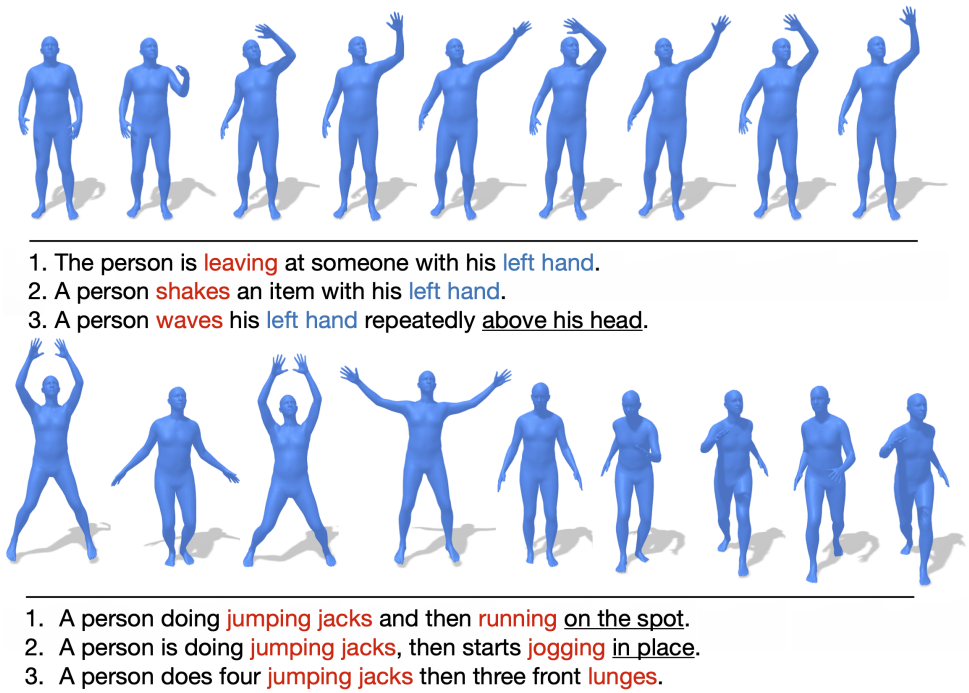


Figure 1: Motion-text pairs from the HumanML3D dataset [19].

3.2.1. Occupational Lifting Dataset Creation

Because standard T2M datasets like HumanML3D [19] are predominantly composed of daily activities or sports captured via omnidirectional Motion Capture (MoCap) systems (Fig. 1), we constructed a custom dataset of occupational lifting scenarios. We enriched the HumanML3D dataset with our custom SAFELIFT dataset [5, 30].

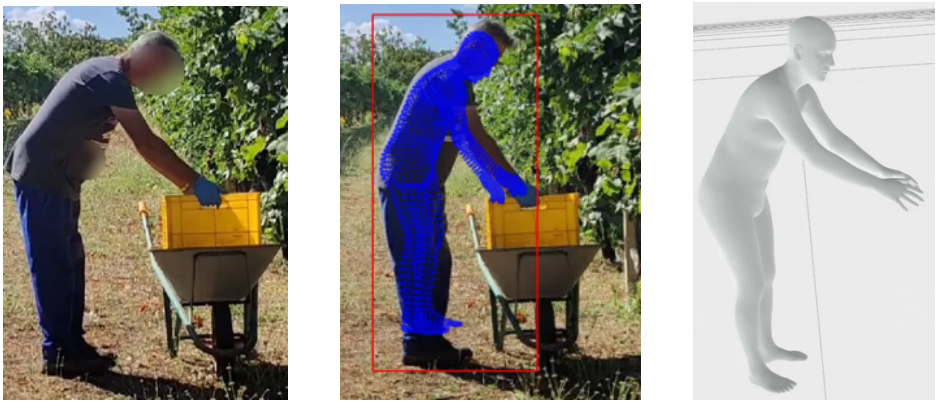
To increase spatial variability and double the available training samples, we applied horizontal flipping to the extracted kinematic sequences. In total, this augmented dataset introduced 578 new occupational lifting poses and 2,312 specific textual descriptions. The lifting videos featured 17 subjects recorded across four distinct real-world environments: a laboratory, an office, a classroom, and a vineyard. Unlike HumanML3D, our kinematic data was extracted "in-the-wild" using HPE from monocular video, restricted to two fixed camera angles (side and rear-side).

To extract the 3D poses accurately and ensure representational compatibility with MoMask, we utilized SMPLer-X [31] to generate volumetric body meshes in the SMPL-X format [32], capturing subtle postural nuances including hand grips (Fig. 2).

3.2.2. Model Fine-Tuning Strategies

To determine the optimal method for integrating strict ergonomic constraints, we conducted two distinct fine-tuning experiments:

- **MoMask Retraining** (Mixed-Domain): Retraining the model on the combined dataset (HumanML3D + SAFELIFT). This process required approximately 15 days for 150 epochs on an NVIDIA RTX 3090 GPU (24GB VRAM).
- **MoMask Finetuning** (Task-Specific): Fine-tuning the model exclusively on the SAFELIFT dataset, starting directly from the original pre-trained checkpoint. This focused adaptation took roughly 6 hours for 50 additional epochs on an NVIDIA TITAN X GPU (12GB VRAM).



(a) Original frame (b) Mesh overlay (c) 3D side view

Figure 2: 3D pose extraction pipeline using SMPLer-X.

Table 1
The four-way comparative questionnaire used in Study 1 to evaluate user perception of the visual feedback formats.

Evaluation Dimension	Question Prompt (Translated)
Naturalness	*"Which animation appears most natural and fluid in representing the movement?"*
Perceived Accuracy	*"Which animation seems most accurate in showing how to correctly execute the movement?"*
Instructiveness	*"Which animation would you be most inclined to follow as a guide to execute this movement at work?"*

3.3. Phase 3: Human-Centered Evaluation Protocol

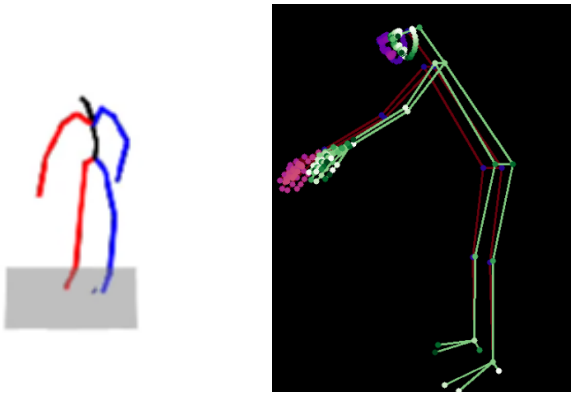
To move beyond computational metrics, we designed two user studies to capture perceptions of the visual feedback formats.

General Procedure: To ensure a fair comparison and prevent visual bias, all T2M animations were rendered using a standard skeletal representation rather than their supported volumetric meshes (Fig. 1). This prevented the baseline skeletal visualization (SAFELIFT) from being unfairly disadvantaged by differences in rendering fidelity (Fig. 3). Across both studies, we utilized a direct comparative (forced-choice) approach rather than Likert-scale evaluations to minimize cognitive load and focus on the relative differences between visual formats. To prevent bias and order effects, all visualizations were presented anonymously in a randomized order. The underlying generative systems were kept strictly blind to the users, and participants were allowed to replay the animated feedback as many times as needed before finalizing their evaluations.

3.3.1. Study 1: Zero-Shot Generative vs. Algorithmic Baseline

To translate our preliminary qualitative evaluation into empirical validation, we designed a study to test whether generative visualizations effectively communicate ergonomic information to non-experts, compared with the validated SAFELIFT approach [5].

We recruited 33 participants (aged 18–65) from the general public who had no prior experience with the task. Each participant viewed four visualizations of the same ergonomically correct movement: animations generated by zero-shot MotionGPT-3, FineMoGen, MoMask, and the baseline skeletal rendering produced by SAFELIFT. After observing all four visualizations, participants were asked to choose a single preferred animation across four distinct dimensions. The exact questionnaire prompts are detailed in Table 1.



(a) Generative T2M skeletal visualization

(b) SAFELIFT skeletal visualization

Figure 3: Sample of visualizations compared in the user studies.

Table 2
The two-way comparative questionnaire used in Study 2 to evaluate user perception of the fine-tuned generative model versus the skeletal baseline.

Evaluation Dimension	Question Prompt (Translated)
Comprehension	*"Which format helped you understand the movement better?"*
Effectiveness	*"Which format do you think is most effective for improving posture during a lift?"*
Naturalness	*"Which format shows more natural movements?"*
Deployment Preference	*"Which format would you prefer to see in a real-world context?"*
Unclarity	*"Which format do you think is the least clear?"*
Interpretation Time	*"Which format do you think would require more time to interpret?"*

3.3.2. Study 2: Fine-Tuned Generative vs. Algorithmic Baseline

To verify if domain-specific fine-tuning offered qualitative benefits despite the degraded computational metrics, we designed a second user study. This allowed us to assess whether the adapted generative model (*MoMask*) was more informative and interpretable in real-world scenarios than the SAFELIFT approach [5, 33].

We recruited 30 participants (aged 18–27) who had no prior experience with the task. Each participant accessed a web-based questionnaire and viewed two distinct visualizations of the same ergonomically correct lifting movement: an animation generated by the fine-tuned MoMask model and the baseline skeletal rendering produced by SAFELIFT. Participants evaluated the formats across six distinct dimensions focusing on clarity, effectiveness, and real-world deployment preference. The exact questionnaire prompts are detailed in Table 2.

4. Experimental Results and Discussion

This section presents the findings from our evaluation pipeline, analyzing both computational metrics and human-centered feedback.

4.1. Phase 1 Results: Zero-Shot Quantitative Evaluation

The zero-shot quantitative evaluation was conducted on the overhead reaching task using *Metric* and *Anatomical* prompting strategies. Table 3 summarizes the results.

Table 3
Quantitative evaluation of overhead reaching tasks using *Metric* and *Anatomical* prompts.

	Metric Prompts				Anatomical Prompts			
Model	**Target**	**Measured**	**Abs. Err.**	**MAE**	**Target**	**Measured**	**Abs. Err.**	**MAE**
MotionGPT-3	10 cm	41.7 cm	31.7 cm	24.3 cm	Hand (10 cm)	36.6 cm	26.6 cm	21.1 cm
	30 cm	22.3 cm	7.7 cm		Forearm (30 cm)	35.7 cm	5.7 cm	
	60 cm	26.4 cm	33.6 cm		Arm (60 cm)	28.9 cm	31.1 cm	
FineMoGen	10 cm	49.1 cm	39.1 cm	27.7 cm	Hand (10 cm)	58.3 cm	48.3 cm	21.7 cm
	30 cm	64.1 cm	34.1 cm		Forearm (30 cm)	44.7 cm	14.7 cm	
	60 cm	69.9 cm	9.9 cm		Arm (60 cm)	62.0 cm	2.0 cm	
MoMask	10 cm	40.2 cm	30.2 cm	51.2 cm	Hand (10 cm)	29.8 cm	19.8 cm	20.6 cm
	30 cm	85.3 cm	55.3 cm		Forearm (30 cm)	29.8 cm	0.2 cm	
	60 cm	128.0 cm	68.0 cm		Arm (60 cm)	18.3 cm	41.7 cm	

Table 4
Comparison of motion realism (FID) and semantic alignment (R-Precision Top-1) across adaptation strategies.

Test Set	**Model**	**FID ↓**	**R-Precision ↑**
HumanML3D	MoMask	$\mathbf{0.045^{\pm.002}}$	$\mathbf{0.521^{\pm.002}}$
	MoMask Retraining	$9.707^{\pm.067}$	$0.300^{\pm.002}$
	MoMask Finetuning	$116.964^{\pm.047}$	$0.109^{\pm.004}$
HumanML3D + SAFELIFT	MoMask	$\mathbf{0.125^{\pm.005}}$	$\mathbf{0.558^{\pm.006}}$
	MoMask Retraining	$9.981^{\pm.073}$	$0.289^{\pm.004}$
	MoMask Finetuning	$115.537^{\pm.082}$	$0.035^{\pm.001}$
SAFELIFT	MoMask	$\mathbf{7.650^{\pm.246}}$	$0.106^{\pm.007}$
	MoMask Retraining	$8.948^{\pm.214}$	$\mathbf{0.120^{\pm.006}}$
	MoMask Finetuning	$7.693^{\pm.247}$	$0.109^{\pm.004}$

Regarding the *Metric* prompts, none of the three models passed the accuracy test. While current occupational health literature does not define a strict kinematic error tolerance for visual feedback, we established a conservative threshold of ±5 cm as a reasonable heuristic for biomechanical safety; no model met this criteria. MotionGPT-3 produced substantially constant reaching values regardless of the specified target, completely ignoring the numerical values. FineMoGen showed weak ordinal sensitivity, measurements grew monotonically as the target increased, but systematically overestimated the movements. MoMask exhibited the most problematic behavior, with extreme errors and structural anomalies, including severe instability in the skeleton's scale normalization.

With *Anatomical* prompts, isolated successes emerged (FineMoGen passed the 60 cm target, and MoMask passed the 30 cm target). However, these appear to be coincidental rather than evidence of genuine geometric control, as the models failed on adjacent references, resulting in massive errors.

Ultimately, without explicit quantitative supervision, models did not demonstrate to control precise geometric parameters regardless of their architectural sophistication. Under our heuristic, none of the evaluated models achieved acceptable accuracy for formal NIOSH applications.

4.2. Phase 2 Results: Domain Adaptation via Fine-Tuning

We evaluated the adapted MoMask models using standard generative motion metrics: Fréchet Inception Distance (FID) for motion realism, and R-Precision for text-motion semantic alignment [7].

As shown in Table 4, domain adaptation yielded severe degradation in both motion realism (higher FID) and semantic alignment (lower R-Precision) compared to the baseline model.

This degradation highlights a critical real-world deployment challenge for WellRec systems. We attribute this performance drop to two primary factors:

Table 5
Absolute preferences from *Study 1* for the zero-shot generative vs. algorithmic baseline comparison ($N = 33$).

Question	MotionGPT-3	FineMoGen	MoMask	SAFELIFT
Most **natural**	**23**	5	4	1
Most **accurate**	10	5	3	**15**
Most **instructive**	**15**	4	5	9

Table 6
Absolute preferences from *Study 2* comparing the fine-tuned generative model against the algorithmic baseline ($N = 30$).

Evaluation Dimension	SAFELIFT	Fine-Tuned MoMask
Comprehension	21	9
Effectiveness	21	9
Naturalness	22	8
Deployment Preference	20	10
Unclarity	8	22
Interpretation Time	9	21

- The Kinematic Domain Gap: HumanML3D relies on pristine, omnidirectional MoCap data, whereas our lifting data relies on monocular HPE. The partial spatial occlusion inherent in fixed-angle video creates noisy skeletal inferences that standard T2M models struggle to assimilate.
- Caption Complexity: Standard HumanML3D textual descriptions are highly descriptive (averaging 12 words), capturing sequential micro-actions. Conversely, our automated occupational descriptions were highly repetitive and shorter (averaging 9.25 words), reducing the text encoder's ability to map distinct semantic features to precise spatial embeddings. Complete details regarding the textual descriptions generation process are provided in our GitHub[1] repository.

4.3. Human-Centered Evaluation Findings

4.3.1. Study 1: Zero-Shot Generative vs. Algorithmic Baseline Comparison

Table 5 summarizes the absolute preferences from our first user study ($N = 33$).

A clear separation between *naturalness* and *perceived accuracy* emerged. On the naturalness dimension, MotionGPT-3 dominated (69.7% of preferences). Conversely, on the perceived accuracy dimension, the pattern reversed: SAFELIFT's skeletal rendering was preferred by 45.5% of participants, surpassing all generative models. Qualitative feedback indicated that while participants found generative animations more fluid, they relied on the explicit geometry of the skeletal rendering to judge correctness. Interestingly, for *instructiveness*, MotionGPT-3 returned to the top (45.5%), suggesting that for ordinary users, naturalness heavily influences the perceived utility of a learning tool.

4.3.2. Study 2: Fine-Tuned Generative vs. Algorithmic Baseline Comparison

Table 6 summarizes the preferences from our follow-up study ($N = 30$). Comparing the fine-tuned *MoMask Retraining* model against SAFELIFT mirrored the quantitative failures: users preferred the traditional skeletal interpolation in 57% of cases. Crucially, participants reported that the fine-tuned MoMask animations were less clear and required more cognitive effort to interpret. The degradation in generation quality caused the generative model to lose its primary advantage, naturalness, making the rigid but highly accurate skeletal wireframe the unambiguously preferred format.

4.4. Discussion and Design Implications for WellRec

The results highlight a "productive tension" between semantic naturalness and biomechanical accuracy. Generative animations effectively communicate the "how" of the movement (fluidity, coordination), while traditional skeletal overlays communicate the "how much" (final position, exact angles).

For future Wellbeing Recommender Systems, our findings suggest that an optimal visual interface should combine these modalities: using fine-tuned text-to-motion animations to communicate the overall qualitative nature of a healthy posture, overlaid or paired with standard skeletal rendering to guarantee the strict geometric precision required by occupational safety standards like the RNLE.

5. Conclusion and Future Work

In this exploratory work, we investigated the viability of LLM-based text-to-motion architectures as visual behavioral recommenders for occupational wellbeing [13]. By conceptualizing generative models as a novel modality for delivering ergonomic feedback, we aimed to address the limitations of rigid, interpolation-based skeletal renderings like those used in our previous work [5].

Our evaluation exposed a critical trade-off between qualitative semantic naturalness and quantitative biomechanical precision [1]. *Phase 1* demonstrated that while pre-trained generative models excel at producing fluid and semantically plausible animations, they systematically fail to adhere to the strict geometric constraints required by safety standards like the RNLE. Furthermore, *Phase 2* revealed that fine-tuning these discrete latent architectures on "in-the-wild" monocular HPE data introduces significant real-world deployment challenges; the kinematic domain gap and descriptive ambiguities fundamentally degraded both motion realism and semantic alignment.

Despite these computational limitations, our Human-in-the-Loop evaluation methodology shows that users exhibited a strong preference for the naturalness of generative animations when conceptualizing the "flow" of a movement, but decisively relied on explicit skeletal interpolation to accurately judge postural correctness and establish trust.

Our results suggest that, in the current setting, unconstrained generative motion models are not yet robust enough to replace algorithmic interpolation for safety-critical occupational feedback. These findings motivate future hybrid interfaces aimed at balancing biomechanical safety with user comprehensibility, grounded on designing hybrid visual interfaces. By layering the strict, trustworthy geometric precision of traditional interpolation over the fluid, easily interpretable kinematics of generative models, we could develop adaptive recommendation strategies that are both biomechanically safe and highly comprehensible for the end-user.

Acknowledgments

The research is partially funded by PNRR - Mission 4 ("Education and research") – Component 2 ("From research to business"), Investment 3.3 ("Introduction of innovative doctorates that respond to the innovation needs of companies and promote the hiring of researchers by companies") D.M.n. 117/2023 - CUP: H91I23000170007. We extend our sincere gratitude to Naps Lab S.r.l.s. for their support and collaboration in the realisation of this research. Moreover, we would like to thank Andrea Romano for his valuable support throughout this project.

Declaration on Generative AI

During the preparation of this work, the author(s) used MotionGPT-3, FineMoGen, and MoMask in order to: Generate 3D human motion animations evaluated in the experimental phases and user studies. Limited Generative AI assistance was used exclusively for minor language editing tasks, such as grammar checking, spelling correction, and occasional sentence rephrasing to improve clarity. All intellectual and analytical content presented in the paper remains solely the work of the authors.

References

[1] M. Kumaresan, S. B. Darivemula, S. Bala, S. Kadas, Musculoskeletal disorders among long-standing workers working for more than 6-hours a day in an automobile factory in south india, Journal of Emergencies, Trauma, and Shock 18 (2025) 119–125. doi:10.4103/jets.jets_161_24.

[2] S. Jung, L. Qing, B. Su, X. Xu, Toward a fully automated niosh lifting equation using computer vision, Human Factors 0 (0) 00187208261418858. doi:10.1177/00187208261418858, pMID: 41605555.

[3] G. Dibenedetto, S. Sotiropoulos, M. Polignano, G. Cavallo, P. Lops, Comparing human pose estimation through deep learning approaches: An overview, Computer Vision and Image Understanding (2025) 104297. doi:https://doi.org/10.1016/j.cviu.2025.104297.

[4] T. R. WATERS, V. PUTZ-ANDERSON, A. GARG, L. J. F. and, Revised niosh equation for the design and evaluation of manual lifting tasks, Ergonomics 36 (1993) 749–776. doi:10.1080/00140139308967940, pMID: 8339717.

[5] G. Dibenedetto, P. Lops, P. Lovreglio, M. Polignano, R. Ravallese, H. Torkamaan, SAFELIFT: safety-aware feedback for ergonomic lifting & injury-free tasks, in: Proceedings of the 31st International Conference on Intelligent User Interfaces, IUI 2026, Paphos, Cyprus, March 23-26, 2026, ACM, 2026, pp. 988–1003. doi:10.1145/3742413.3789143.

[6] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, T. Chen, Motiongpt: Human motion as a foreign language, in: Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 20067–20079. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/3fbf0c1ea0716c03dea93bb6be78dd6f-Paper-Conference.pdf.

[7] C. Guo, Y. Mu, M. G. Javed, S. Wang, L. Cheng, Momask: Generative masked modeling of 3d human motions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1900–1910.

[8] A. R. Sahili, N. Neji, H. Tabia, Text-driven motion generation: Overview, challenges and directions, 2025. arXiv:2505.09379.

[9] H. Schäfer, Personalized support for healthy nutrition decisions, in: Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 455–458. doi:10.1145/2959100.2959105.

[10] A. El Majjodi, A. D. Starke, C. Trattner, Integrating digital food nudges and recommender systems: Current status and future directions, IEEE Access 13 (2025) 123002–123017. doi:10.1109/ACCESS.2025.3588663.

[11] G. Dibenedetto, E. Musacchio, M. Polignano, P. Lops, Fine-tuning large multimodal models for fitness action quality assessment, in: Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2025, New York City, NY, USA, June 16-19, 2025, ACM, 2025, pp. 39–44. doi:10.1145/3708319.3733684.

[12] I. Coppens, T. De Pessemier, L. Martens, Balancing habit repetition and new activity exploration: A longitudinal micro-randomized trial in physical activity recommendations, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1147–1151. doi:10.1145/3640457.3691715.

[13] M. Khwaja, M. Ferrer, J. O. Iglesias, A. A. Faisal, A. Matic, Aligning daily activities with personality: towards a recommender system for improving wellbeing, RecSys '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 368–372. doi:10.1145/3298689.3347020.

[14] S. E. Mudiyanselage, P. H. D. Nguyen, M. S. Rajabi, R. Akhavian, Automated workers' ergonomic risk assessment in manual material handling using semg wearable sensors and machine learning, Electronics 10 (2021). doi:10.3390/electronics10202558.

[15] G. Zhou, V. Aggarwal, M. Yin, D. Yu, A computer vision approach for estimating lifting load contributors to injury risk, IEEE Transactions on Human-Machine Systems 52 (2022) 207–219. doi:10.1109/THMS.2022.3148339.

[16] N. Sabetta, M. Bernabei, S. Colabianchi, D. Colangelo, F. Costantino, et al., Ergonomic training tool: a pose detection-based digitalization of iso/tr 12295 and iso 11228-1, in: Proceedings of the Summer School Francesco Turco, 2024.

[17] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, A. H. Bermano, Human motion diffusion model, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=SJ1kSyO2jwu.

[18] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, Y. Shan, Generating human motion from textual descriptions with discrete representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14730–14740.

[19] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, L. Cheng, Generating diverse and natural 3d human motions from text, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5152–5161.

[20] F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, 2022. doi:10.1007/978-1-0716-2197-4.

[21] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, L. Getoor, Personalized explanations for hybrid recommender systems, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019, ACM, 2019, pp. 379–390. doi:10.1145/3301275.3302306.

[22] C. Tsai, P. Brusilovsky, Designing explanation interfaces for transparency and beyond, in: Joint Proceedings of the ACM IUI 2019, Los Angeles, USA, March 20, 2019, volume 2327 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2327/IUI19WS-IUIATEC-4.pdf.

[23] N. Tintarev, J. Masthoff, Evaluating the effectiveness of explanations for recommender systems - methodological issues and empirical studies on the impact of personalization, User Model. User Adapt. Interact. 22 (2012) 399–439. doi:10.1007/S11257-011-9117-5.

[24] B. P. Knijnenburg, M. C. Willemsen, Evaluating recommender systems with user experiments, in: Recommender Systems Handbook, Springer, 2015, pp. 309–352. doi:10.1007/978-1-4899-7637-6_9.

[25] M. A. Chatti, M. Guesmi, A. Muslim, Visualization for recommendation explainability: A survey and new perspectives 14 (2024). doi:10.1145/3672276.

[26] E. Lukianova, J. Jeong, J. Jeong, A picture is worth a thousand words? investigating the impact of image aids in AR on memory recall for everyday tasks, in: Proceedings IUI 2025, Cagliari, Italy, March 24-27, 2025, ACM, 2025, pp. 106–126. doi:10.1145/3708359.3712087.

[27] M. Szymanski, M. Millecamp, K. Verbert, Visual, textual or hybrid: the effect of user expertise on different explanations, in: IUI '21, College Station, TX, USA, April 13-17, 2021, ACM, 2021, pp. 109–119. doi:10.1145/3397481.3450662.

[28] B. Zhu, B. Jiang, S. Wang, S. Tang, T. Chen, L. Luo, Y. Zheng, X. Chen, Motiongpt3: Human motion as a second modality, 2025. arXiv:2506.24086.

[29] M. Zhang, H. Li, Z. Cai, J. Ren, L. Yang, Z. Liu, Finemogen: Fine-grained spatio-temporal motion generation and editing, in: Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 13981–13992. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/2d52879ef2ba487445ca2e143b104c3b-Paper-Conference.pdf.

[30] G. Dibenedetto, P. Lops, M. Polignano, H. Torkamaan, Lift it up right: A recommender system for safer lifting postures, in: Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys 2025, Prague, Czech Republic, September 22-26, 2025, ACM, 2025, pp. 1222–1227. doi:10.1145/3705328.3759314.

[31] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang, C. C. Loy, L. Yang, Z. Liu, SMPLer-X: Scaling up expressive human pose and shape estimation, in: Advances in Neural Information Processing Systems, 2023.

[32] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, M. J. Black, Expressive body capture: 3D hands, face, and body from a single image, in: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10975–10985.

[33] G. Dibenedetto, M. Polignano, P. Lops, G. Semeraro, Human pose estimation for explainable corrective feedbacks in office spaces, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2024, Cagliari, Italy, July 1-4, 2024, ACM, 2024. doi:10.1145/3631700.3665184.