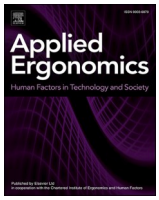
Applied Ergonomics 138 (2027) 104831

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo



Vision-language models for occupational physical exposure assessment: Classification and temporal segmentation of manual material handling tasks

Mohammad Sadra Rajabi [a], Aanuoluwapo Ojelade [b], Sunwook Kim [a], Maury A. Nussbaum [a,*]

[a] Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, 24061, USA
[b] St. Jude Children's Research Hospital, Memphis, TN, 38105, USA

ARTICLE INFO

Keywords:
Physical exposure assessment
Vision-language model (VLM)
Generative pretrained transformer (GPT)
Temporal action segmentation
Human activity recognition (HAR)

ABSTRACT

Effective physical exposure assessment for manual materials handling (MMH) is essential for identifying activities that increase the risk of work-related musculoskeletal disorders and for guiding ergonomic interventions. However, existing methods are labor-intensive and often fail to capture task variability or to effectively estimate task timing characteristics. We evaluated the use of vision-language models (VLMs) to automatically and non-invasively classify eight MMH tasks and specific task conditions (i.e., hand configuration and lifting origin), and to detect task start and end times, using regular RGB video streams. We obtained task classification accuracies of ~82–85%, accuracies for classifying lifting origin of ~94–98%, and mean absolute start and end time errors <0.5 s, superior to prior work in some cases. Classification performance for hand configuration, though, was more variable. These findings demonstrate the potential of VLMs as a practical and scalable tool for physical exposure assessment of MMH tasks.

1. Introduction

Work-related musculoskeletal disorders (WMSDs) risk increases with repeated or long-duration manual material handling (MMH) tasks such as lifting, carrying, and lowering (Hoozemans et al., 2004; Marras and Davis, 1998). MMH tasks are widespread across workplaces, yet differences in task execution among individuals and work settings complicate physical exposure assessment (David, 2005; Deros et al., 2017; Zhao et al., 2022). Effectively and accurately assessing physical exposures—comprising task frequency, duration, and the magnitude of physical load—is essential for recognizing high-risk tasks and employing targeted ergonomic interventions (Punnett and Wegman, 2004). Nevertheless, established approaches are often quite labor-intensive and can fail to adequately describe MMH task variation or provide reliable measurements of exposure duration, frequency, and associated physical loads (Kadikon and Rahman, 2016; Seo and Lee, 2021). Thus, there is a need for more practical tools for real environments (Leggieri et al., 2023; Seo and Lee, 2021).

Physical exposure assessments in the workplace require approaches that can classify and temporally segment MMH tasks effectively and that are compatible with the work environment. Conventional tools include self-report, human observation, or direct measurement (David, 2005; Li and Buckle, 1999; Spielholz et al., 2001). However, the first two approaches are prone to individual biases, potentially leading to an inaccurate assessment of exposure intensity and frequency (David, 2005; Li and Buckle, 1999; Spielholz et al., 2001). Direct measurement approaches—which often involve wearable sensors—offer objective and precise data for evaluating physical exposure (Mudiyanselage et al., 2021; Nath et al., 2018). While some recent wearable sensor systems support real-time data collection and analysis (Lind et al., 2023; Nath et al., 2018), broader adoption might be hindered by equipment costs (Dempsey et al., 2005; Schall et al., 2018), calibration requirements (Guner and Dasdemir, 2023), and the potential for discomfort or altered work behaviors among employees (Antwi-Afari et al., 2018; Golabchi et al., 2016). Therefore, there is an important need for an assessment approach that is both objective and non-intrusive. Automated, non-contact (ambient), vision-based approaches can be a promising alternative for physical exposure assessment (Nath et al., 2018; Ojelade

* Corresponding author.
E-mail address: nussbaum@vt.edu (M.A. Nussbaum).

https://doi.org/10.1016/j.apergo.2026.104831
Received 27 October 2025; Received in revised form 13 February 2026; Accepted 1 June 2026
Available online 5 June 2026

et al., 2025; Santiago-Colón et al., 2021; Wang et al., 2021).

Recent advancements in sensor technology and machine learning have supported innovative computer-based approaches for occupational physical exposure assessment, especially related to posture (Anacleto Filho et al., 2024; Menanno et al., 2024). Of relevance here are *vision--language models* (VLMs; Fan, Mei and Li, 2024; Fan et al., 2024; Laurençon et al., 2024; Rajabi et al., 2025; Yong et al., 2024), which refer broadly to models that jointly process visual and textual data. In practice, many VLMs integrate vision encoders, such as *vision transformers* (ViTs), together with large language models like *generative pre-trained transformers* (GPT; Tang et al., 2024; Tzelepi and Mezaris, 2024; Zhang et al., 2024); this integration can offer a non-invasive, cost-effective approach to assessing postural exposures using regular RGB video data (Ferrara, 2024; Yong et al., 2024). Specifically, these models can be trained directly on video data to identify human joint positions and recognize postures or movements that may contribute to WMSDs (Fan et al., 2024). Existing research supports the potential application of VLMs for physical exposure risk assessment. For example, Fan et al. (2024) demonstrated that VLMs can automatically detect postures associated with ergonomic risks—such as lifting heavy materials or maintaining prolonged non-neutral postures—from RGB video in construction settings. These investigators also developed a visual query system that translates images into textual insights for exposure risk assessment (Fan et al., 2024). Moreover, Yong et al. (2024) demonstrated that VLMs with caption augmentation can automatically identify ergonomic risk factors and generate solutions from images in construction.

However, reported applications of VLMs for occupational physical exposure risk assessment have two important limitations. First, it is unclear whether VLMs can effectively classify MMH tasks based on RGB video—particularly for tasks involving complex and/or continuous motion patterns—and whether they can accurately capture task sequences from RGB video streams. Many existing approaches rely on discrete task analysis, which can underestimate cumulative physical exposures (David, 2005; Punnett and Wegman, 2004) and often require extensive manual video processing (Khosrowpour et al., 2014; Spielholz et al., 2001; Yang et al., 2016), making them time-consuming and less practical for real environments. Addressing these limitations is necessary to determine the feasibility of using low-cost, non-contact approaches that rely on regular cameras, reducing equipment costs. Second, in vision-based systems, the choice of encoder—the model component that extracts visual features from video frames—can affect classification performance (Scabini et al., 2024). Different encoders are optimized to capture distinct types of information, such as fine-grained local details or broader spatial context. Therefore, it is important to evaluate how specific encoder architectures affect model accuracy and generalizability. However, prior studies have not systematically compared different vision encoders to determine which are most effective for classifying MMH tasks from RGB video.

In this study, we evaluated the effectiveness of VLM-based pipelines in classifying and temporally segmenting diverse MMH tasks using regular RGB video. Unlike pose-based computer-vision approaches (e.g., Kim et al., 2021; Ojelade et al., 2025), we used ViT-extracted frame-level visual features and a GPT-based temporal model to support both MMH task classification and temporal segmentation from regular RGB video streams. We compared the performance of several VLM pipelines, including different vision encoders, to determine their relative ability to classify complex MMH task sequences. We hypothesized that classification performance would be strong but also vary across VLM pipelines, and that VLMs could effectively segment continuous MMH tasks without the need for manual video processing. While our immediate goal was to assess VLM performance in simulated MMH tasks, our findings could also guide the broader application of automated video-based physical exposure assessment across diverse occupational settings.

2. Methods

2.1. Overview of the dataset

Data used herein were obtained in a prior study by Ojelade et al. (2025), in which 36 healthy young adults (22 males and 14 females; Appendix A.1) performed eight simulated MMH tasks under various conditions in a controlled laboratory setting. The tasks included (see Fig. 1, and Appendix A.2): symmetric lifting from the floor or knee to hip height (Task 1); asymmetric lifting (Task 2); load carriage (Task 3); pushing a box across a table (Task 4); pulling a box toward the body (Task 5); cart pushing (Task 6); lifting a box from cart to overhead height (Task 7); and lowering a box from overhead to either floor or knee height (Task 8). In addition, a ninth category (Task 9) captured non-MMH activities such as unloaded walking, standing, and approaching or leaving the task setup. Such non-task intervals were included for better segmentation during the model training described below. Collectively, these tasks represent core components of mixed MMH workflows commonly observed in occupational settings, in which workers routinely transition between MMH tasks, consistent with some prior MMH task classification studies (e.g., Kim and Nussbaum, 2014; Porta et al., 2021).

Participants completed 36 task conditions involving three *Hand Configurations* (broad = 52 cm handle spacing, narrow = 33 cm, and one-handed; see Figure A.1); three *Box Mass* levels (6, 9, and 12 kg for two-hand lifts; 5, 7, and 9 kg for one-hand lifts); two *Lift Origins* (floor and knee height), and two *Starting Positions* (see Appendix A.3). Whole-body motion was recorded using regular RGB video data captured by three synchronized cameras, which recorded at 30 Hz (Appendix A.4; Figure A.2).

2.2. Overview of the MMH task classification and segmentation pipeline

MMH task classification and temporal segmentation can be viewed as a structured, multi-stage pipeline (Fig. 2). This pipeline architecture was inspired by the TimeChat framework of Ren et al. (2024), which integrates time-sensitive visual encoding and sequence modeling for video understanding. Our pipeline is divided into two steps. Step 1 is the classification process, in which regular RGB video frames are labeled, transformed into tensors, and then processed using VLM-pretrained ViT image encoders to extract features. These features are then classified using a GPT-based model to classify *Task Type*, *Hand Configuration*, and *Lift Origin* for Task 1, sequentially. Step 2 is the temporal segmentation process, which uses the trained general MMH task classifier to assign frame-wise labels across entire video sequences, followed by a majority voting filter to refine task temporal segmentation. Each of these steps is discussed in detail subsequently.

2.3. Data labeling and processing

The RGB video recordings were manually labeled on a frame-wise basis in terms of *Task Type*, *Hand Configuration*, and *Lift Origin* by two independent research assistants; these labels served as ground truth for our classification approach. Task boundaries were defined based on participants' interaction with the box or cart, and non-task frames were labeled as Task 9. The labeled videos were then processed to extract individual frames and converted into a standardized format that was compatible with the ViT pipeline. See Appendix A.5 for details on labeling criteria and processing steps.

2.4. Feature extraction

Three ViTs from different VLMs (see Appendix A.6) were used to extract visual features from each video frame: 1) CLIP (Contrastive Language–Image Pre-training; Radford et al., 2021); 2) MAE (Masked AutoEncoder; Tong et al., 2022); and 3) DINO (self-DIstillation with NO

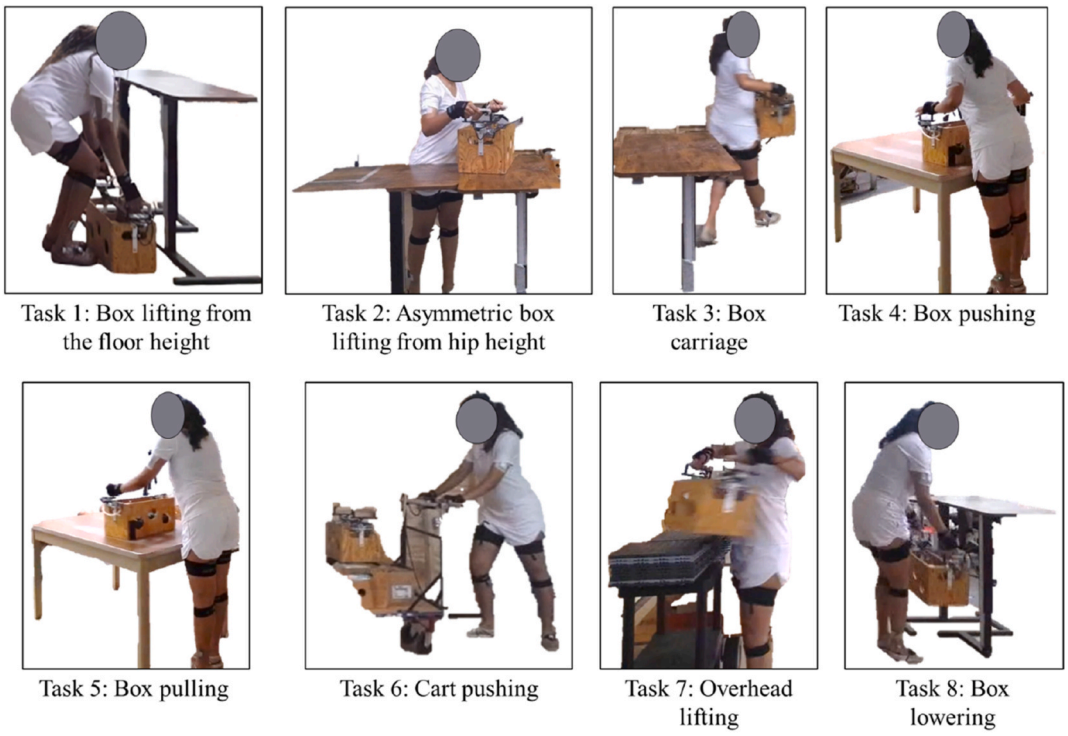


Fig. 1. Illustrations of the eight simulated MMH tasks in the broad hand configuration.

labels, self-supervised learning framework; Caron et al., 2021). These three ViTs were trained using distinct learning objectives and were tested to evaluate their different feature extraction strategies. The vision transformer in CLIP processes frames into high-dimensional features aimed at capturing semantic similarities with textual descriptions and is trained using contrastive learning (Radford et al., 2021). Contrastive learning is a training technique in which a model learns to distinguish between similar instances (positive pairs) and dissimilar instances (negative pairs) by comparing examples (Belcaid et al., 2022; Sabiri et al., 2024). Such a training approach makes the model particularly adept at capturing semantic-level information from frames (Belcaid et al., 2022). The MAE vision transformer is trained through a masked image modeling approach, in which random patches of an image are masked, and the model learns to reconstruct those masked patches (Tong et al., 2022). This training approach enables the transformer to understand spatial structures, all within a self-supervised learning framework (Tong et al., 2022). The DINO vision transformer is trained using self-distillation, a training technique whereby a model learns to generate consistent features across different augmented versions of the same frame—such as those created through random cropping, color jittering, or blurring—within an unsupervised learning framework (Caron et al., 2021). This approach makes the model particularly effective at capturing structural patterns and object-level information within scenes (Caron et al., 2021). These three ViTs were loaded with pre-trained weights with their backbone weights frozen to enable feature extraction from video frames without requiring computationally expensive fine-tuning on our specific dataset.

2.5. Classification approach

We used frame features and corresponding labels as input to a GPT-based model for frame classification. The classification process involved crafting a model architecture, data handling, model training, and evaluation, each of which is discussed in detail subsequently.

2.5.1. Model architecture

We used a modified GPT-2 architecture to classify frame labels while modeling temporal dependencies in sequences of consecutive video frames (Radford et al., 2019). GPT-2 is a deep transformer model with self-attention that is well suited for modeling long-range dependencies in sequences (Radford et al., 2019), and similar applications have been adopted in recent video–language modeling pipelines as temporal reasoning backbones over video frame features (e.g., Ren et al., 2024; Tang et al., 2024; Zhang et al., 2023). In our classification approach, the model takes extracted features from video frames and learns both the content of each frame and the relationship between frames over a sequence of video frames. Here, a "sequence" refers to an ordered set of consecutive frames taken from the video streams. Considering sequences—rather than treating frames independently—is important because MMH tasks involve continuous, coordinated body motions, and accurate classification of a frame often depends on surrounding frames that capture body kinematics over time. The output layer then assigns each frame to a label. Further technical details of the architecture are provided in Appendix A.7.

2.5.2. Data handling, model training, and evaluation

Each sequence was set to a fixed length of consecutive frames (i.e., 100 frames). Preprocessing steps, including frame padding, attention masking, and batching, were applied (see Appendix A.8). Our dataset was imbalanced, with certain MMH tasks (e.g., box carrying) contributing substantially more sequences than others (e.g., box pushing and pulling). This imbalance can bias the model toward more frequent tasks and reduce accuracy for underrepresented tasks. To mitigate this, two techniques were used (Qin et al., 2018; Rezaei-Dastjerdehei et al., 2020): 1) weighted random sampling during data loading to increase the likelihood of balanced class representation in each batch; and 2) a weighted cross-entropy loss function to amplify the contribution of errors from underrepresented classes during training.

Model training was performed for up to 100 epochs, with early stopping applied to prevent overfitting. Specifically, training was terminated if the validation loss—the prediction error measured on

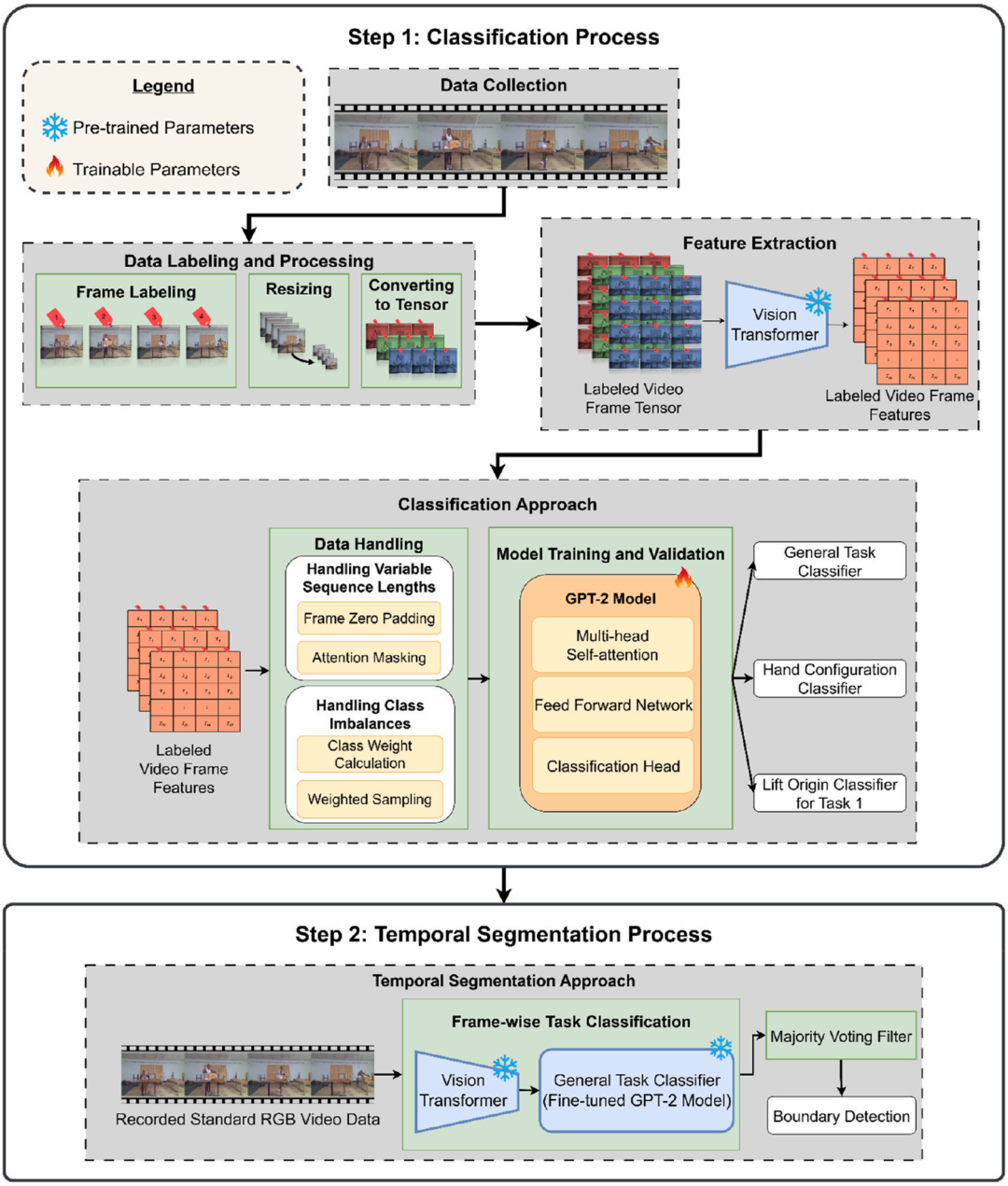


Fig. 2. Overview of the MMH task classification and temporal segmentation pipeline. Step 1: the classification process, where labeled video frames are processed through vision transformers, and a GPT-2-based model classifies *Task Type*, *Hand Configurations,* and *Lift Origins* for Task 1. Step 2: the temporal segmentation process, which refines the prior frame-wise classification using a majority voting filter.

held-out validation data— failed to improve over seven consecutive epochs (e.g., Anam et al., 2024). A leave-one-subject-out approach was used for model validation (e.g., Gholamiangonabadi et al., 2020); training was replicated 36 folds—each fold using 35 participants for training and the remaining one for validation. While this method introduces high variance due to individual differences (Jordao et al., 2018), it more closely replicates real environments, in which models are trained on known data and then used on an unseen participant (Jordao et al., 2018). Validation loss was measured at every epoch, and if a new least validation loss is achieved, the model's state was saved at the end of each fold. These saved models were later used for temporal segmentation (see Section 2.6). Further technical details of the model training are provided in Appendix A.9.

Model classification performance was evaluated using four common metrics (see Figure A.3)—overall (micro) Accuracy, macro-averaged Precision, Recall, and F1-score (Sokolova and Lapalme, 2009). Accuracy measures the ratio of correct predictions to total predictions. Macro-averaging is calculated by computing each metric separately for each class and then averaging the results across all classes. Precision is the ratio of true positives to all predicted positives, and Recall is the ratio of true positives to all actual positives. The F1-score is the harmonic mean of Precision and Recall. We do not report macro accuracy because per-class accuracy is mathematically identical to per-class Recall; thus, a macro-averaged accuracy would duplicate macro-Recall. To further clarify MMH task classification performance, confusion matrices were determined. These matrices were then normalized, so that the values in each row summed to 1 (i.e., making them comparable across tasks regardless of class size), and subsequently the normalized matrices from all 36 folds were combined by computing the mean across folds to create an 'overall' matrix. For certain evaluations, the overall matrix was

filtered by excluding non-MMH tasks (Task 9) so that performance could be assessed more directly for the eight MMH tasks of interest.

2.6. Temporal segmentation approach

Temporal segmentation was performed by applying framewise classification, with the best general MMH task classifier (i.e., the model corresponding to the minimal validation loss) identified through cross-validation. Each video frame was first converted into feature representations and then passed sequentially through the trained GPT-based classifier to predict frame labels. To reduce noise and occasional misclassifications, a majority voting filter (kernel size = 15 frames) was applied (e.g., Trubakov and Trubakova, 2020) and task boundaries were identified based on label transitions. Temporal segmentation performance was evaluated using absolute errors in task start and end times, and Intersection over Union (IoU) for task intervals (see Figure A.4).

2.7. Statistical analyses

To assess the effects of *Model* and *Task Condition* on classification and temporal segmentation performance, a series of repeated-measures analyses of variance (RANOVAs) was conducted. First, separate one-way RANOVAs were performed for each classifier (general, hand configuration, and lift origin) to evaluate the effect of *Model* on classification accuracy (accuracy was computed across all tasks combined). Second, separate two-way RANOVAs were used for each classifier to assess the effects of *Model* and *Task Condition* (eight levels for general task, three for hand configuration, and two for lift origin) on Precision, Recall, and F1-score. For the temporal segmentation metrics (i.e., absolute error for task starts and end times and IoU), two-way RANOVAs were also used to assess the effects of *Model* and *MMH Task*. For each RANOVA model, biological sex (*Sex*) was included as a blocking effect. Where relevant, significant interaction effects were explored using simple-effects testing, and *post hoc* paired comparisons were completed using the Tukey's HSD procedure. All statistical analyses were performed with JMP Pro 18 (SAS, Cary, NC) using the restricted maximum likelihood (REML) method. Logit transformations were applied to Accuracy, Precision, Recall, and F1-scores of the general MMH task classifier, and logarithmic transformations were applied to start and end time absolute errors, in all cases to obtain a normal distribution of model residuals. Statistical significance was determined when $p < 0.05$, and summary data are reported as least-square means based on the statistical model fits.

3. Results

All ANOVA results are summarized in Tables A.1–A.5, and Figures A.5–A.7 provide confusion matrices for each classifier. Significant main or interaction effects of *Model* and *Task Condition* were found across all classification and temporal segmentation performance metrics. More detailed results are provided below. In the rest of the paper, the terms CLIP, MAE, and DINO refer to the full VLMs evaluated in this study, each consisting of a frozen ViT integrated with a GPT-2-based temporal classification model.

3.1. Classifying MMH tasks

Accuracy: There was a significant main effect of *Model* (Table A.1). The DINO model achieved significantly better mean accuracy (~85.0%) vs. MAE (~83.3%) or CLIP (~82.5%), though the differences between models were only ~2–3%. The DINO model achieved strong classification performance, with most predictions on the diagonal of the confusion matrix (Fig. 3), and with misclassifications occurring between some tasks, such as asymmetric lifting (Task 2) and overhead lifting (Task 7).

Precision and F1-score: There was significant main effect of *Model* on both Precision and F1-score (Table A.2). Both metrics were relatively large across models (i.e., ~70–80%), with the DINO model consistently achieving significantly better Precision (~74%) and F1-score (~80%) vs. MAE (~71% and ~77%, respectively) or CLIP (~70% and ~75%, respectively). However, the differences between models were small (~3–5%). *MMH Task* × *Sex* interaction effects were significant for both metrics (Table A.2, Fig. 4 and A.8). Mean Precision and F1-score were generally smaller among females vs. males for all eight tasks (although only some of these sex-based differences were significant). In addition, *MMH Task* had a significant simple effect on Precision and F1-score among both males and females ($p < 0.0001$).

Recall: There were significant main effects of both *MMH Task* and *Model* on Recall (Table A.2). Recall was relatively large across models (i.e., ~85–89%), with the DINO model yielding significantly larger Recall (~89%) vs. MAE (~86%) or CLIP (~85%). As was the case for Precision and F1-score, the differences were quite small (~3–4%). Recall was also relatively large across *MMH tasks* (~83–90%; Fig. 5), with Task 8 achieving the largest value (~90%) and Task 4 the smallest (~83%), though neither was significantly different from all other tasks.

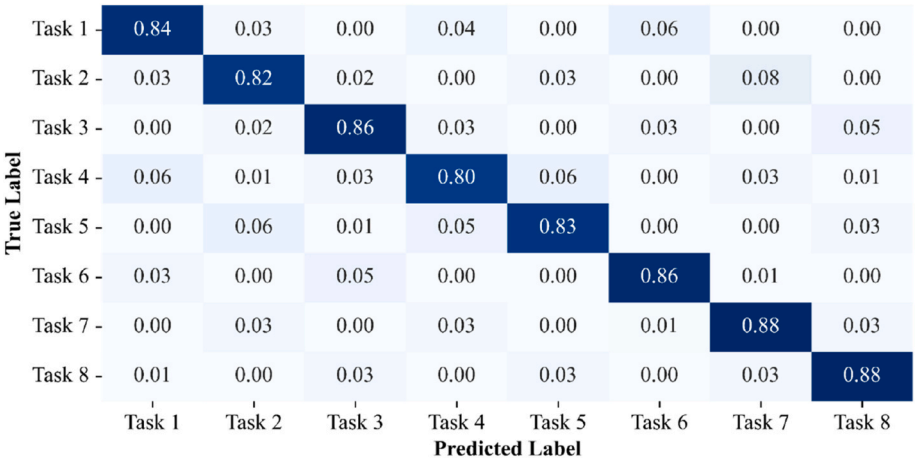


Fig. 3. Overall filtered and normalized confusion matrix across 36 folds using the DINO model, showing classification results for the eight MMH tasks. Note that here and below, each cell represents the mean normalized number of predictions across folds, with darker shades indicating larger (better) values. Entries along the main diagonal and off-diagonal represent correct and incorrect classifications, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

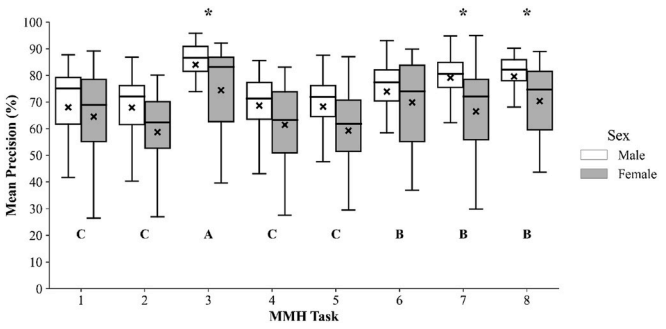


Fig. 4. Interaction effect of *MMH Task* × *Sex* on Precision using the general MMH task classifier. Based on *post hoc* paired comparisons, *MMH tasks* that do not share a common letter are significantly different. Asterisks denote significant simple effects of *Sex* within a specific task.

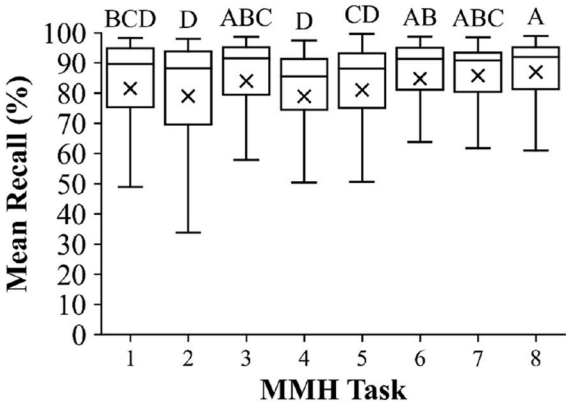


Fig. 5. Main effect of *MMH task* on Recall using the general MMH task classifier.

3.2. Classifying hand configuration

Accuracy: There was a significant main effect of *Model* on accuracy (Table A.1). Similar to results from the general MMH task classifier, DINO achieved larger accuracy (~73%) vs. MAE (~69%) or CLIP (~59%) in classifying hand configuration (Fig. 6).

Precision, Recall, and F1-score: There was a significant interaction effect of *Model* × *Hand Configuration* on all three metrics (Table A.3). CLIP generally showed poorer performance across all hand configurations vs. DINO or MAE (Precision, Recall, and F1-score: ~76–82%; Fig. 7, and A.9). While differences between DINO and MAE were not statistically significant, DINO generally yielded better mean Precision, Recall, and F1-score vs. MAE (~86–89% and ~84–87%, respectively). Across all models, classification performance for the one-hand

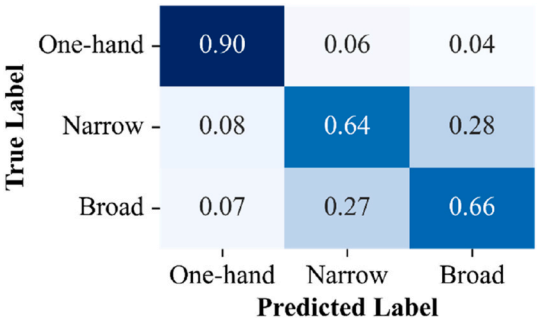


Fig. 6. Overall (mean) normalized confusion matrix across 36 folds for the DINO model, showing classification results for three hand configurations.

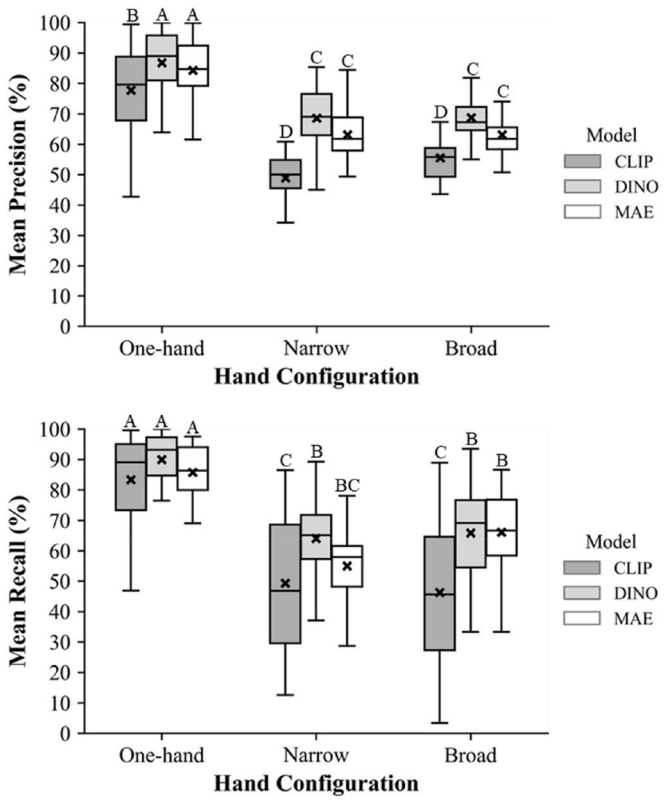


Fig. 7. Significant interaction effect of *Model* × *Hand Configuration* on Precision (top) and Recall (bottom).

configuration was consistently larger vs. the narrow or broad configurations (Fig. 7, A.6, and A.9).

3.3. Classifying Lift Origin for Task 1

Accuracy, Precision, Recall, and F1-score: There was a significant main effect of *Model* on all four metrics (Tables A.1 and A.4). DINO achieved significantly better performance (accuracy: ~98%, Precision: ~99%, Recall: ~98%, F1-score: ~98%) vs. MAE (~96% for all metrics) or CLIP (~94–95% across metrics, Figure A.7).

3.4. Temporally segmenting MMH tasks

Mean absolute start and end time error: There were significant main effects of *Model* and *MMH Task* (Table A.5). For both metrics, DINO achieved significantly smaller absolute errors (start time: ~0.45 s, end time: ~0.35 s) vs. MAE (~0.52 s and ~0.40 s, respectively) or CLIP (~0.54 s and ~0.40 s, respectively; Fig. 8).

There were significantly smaller errors in estimating start time for Task 5 vs. all other tasks, while errors were significantly larger for Tasks 6, 7, and 1 vs. Tasks 3 and 4 (Fig. 9). For end time, there were smaller errors for Tasks 3 and 4 vs. Tasks 1, 2, 6, 7, and 8 (Fig. 9).

IoU: There were significant main effects of *Model* and *MMH Task* on IoU (Table A.5). DINO achieved significantly larger IoU vs. MAE or CLIP (Fig. 10). For *MMH Task*, Task 3 had significantly larger IoU vs. all other tasks (Fig. 10). Tasks 1, 2, 4, and 5 had significantly smaller IoU vs. all other tasks.

4. Discussion

Our goal here was to evaluate the performance of VLMs in classifying diverse MMH tasks and specific task conditions, as well as detecting the temporal boundaries between tasks. We used an approach that

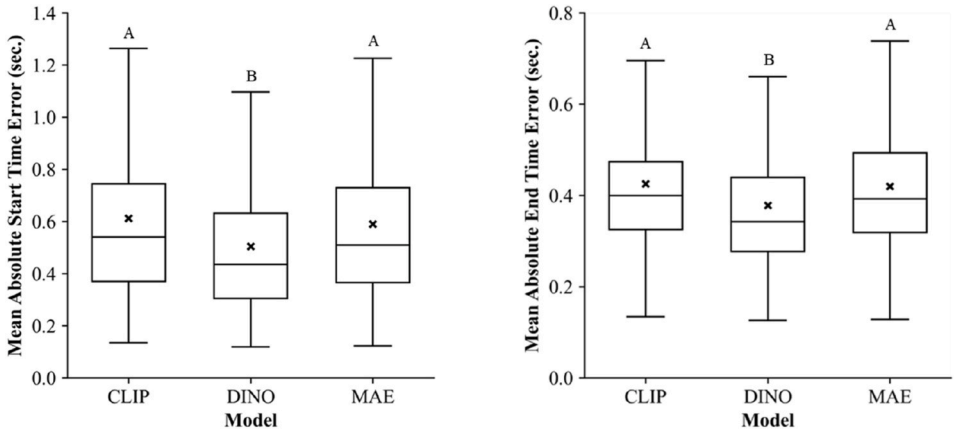


Fig. 8. Significant main effects of the *Model* on mean absolute errors in estimated start time (left) and end time (right).

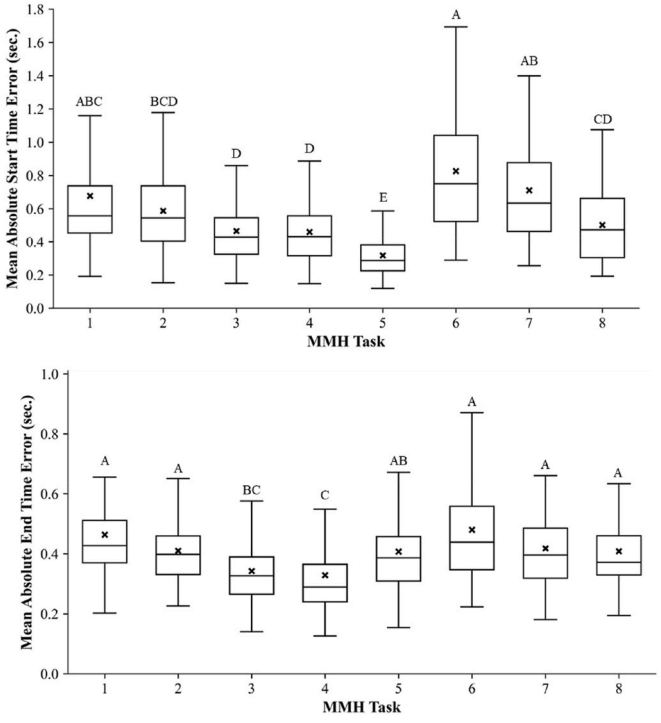


Fig. 9. Significant main effects of *MMH Task* on mean absolute errors in estimated start time (top) and end time (bottom).

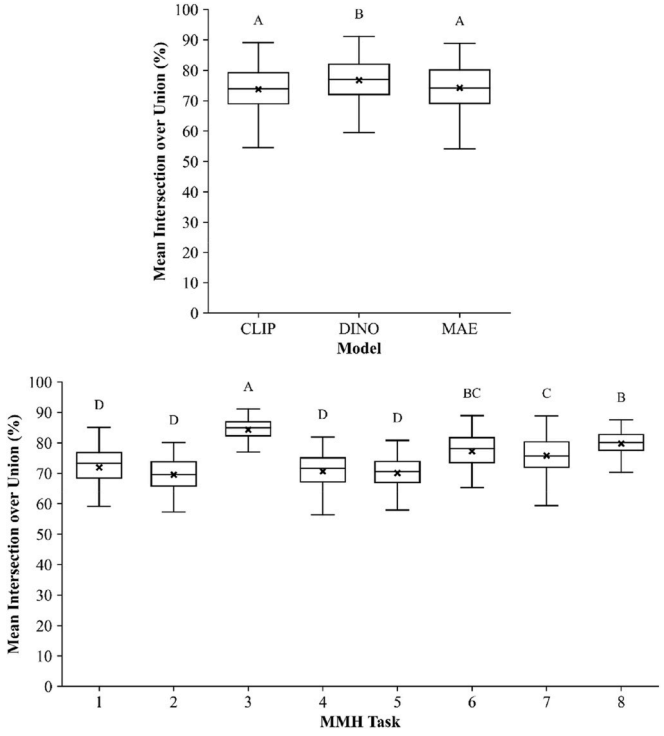


Fig. 10. Significant main effects of *Model* (top) and *MMH Task* (bottom) on IoU.

integrates ViTs with a GPT-based model. Across the three VLMs pipelines, classification of general MMH tasks was achieved with a mean accuracy of ~82–85%, and temporal segmentation yielded a mean IoU of ~74–77% and mean absolute errors in start and end times of ~0.50s and ~0.38s, respectively. Classifying lift origin for Task 1 yielded relatively large accuracy (~94–98%), whereas classifying hand configuration resulted in lesser accuracy (~59–73%). The following discussion addresses these findings in more detail, including model-specific differences and task-level variations in classification and temporal segmentation performance.

4.1. MMH task-specific differences in classification performance

General MMH task classification performance differed—and in some cases, substantially—depending on the task type. These differences are not surprising, since earlier studies have also shown that classification models often misclassify material handling and construction-related tasks with comparable body kinematics (e.g., Barazandeh et al., 2018; Khosrowpour et al., 2014; Kim and Nussbaum, 2014; Ojelade et al., 2025; Porta et al., 2021; Sochopoulos et al., 2025; Yang et al., 2016). These earlier studies, though, used different sensing modalities or modeling approaches. For example, Kim and Nussbaum (2014) employed multiple wearable inertial measurement unit (IMU) sensors, Porta et al. (2021) used a single wearable IMU sensor, Ojelade et al. (2025) applied Kinect-based markerless motion capture with recurrent neural networks, and Sochopoulos et al. (2025) implemented deep learning methods using RGB and depth-based computer vision data.

Consistent with those earlier findings, tasks involving a smaller range of movement variability and more distinct motion patterns generally yielded larger classification performance. For instance, among the MMH tasks studied here, box carrying had the largest classification Precision (Fig. 4). This classification performance is comparable to values reported by Ojelade et al. (2025) using Kinect-based MMC, likely due to this task having a continuous and distinct motion pattern combined with a stable object-handling posture. By contrast, tasks such as asymmetric

lifting and overhead lifting were frequently misclassified with one another (Fig. 3 and A.5), similar to the misclassification patterns reported by Ojelade et al. (2025). We suspect that these misclassifications resulted primarily from the substantial kinematic similarity between the tasks, as both involve bending and box-handling postures that differ mainly in torso rotation and the vertical height of the lift.

The two least frequent tasks—box pushing (Task 4) and box pulling (Task 5)—were most often misclassified with other tasks, whereas box carrying (Task 3), the most frequent task, was often classified more accurately (Figs. 3 and 4, and A.5). Such differences can be largely attributed to class imbalance, a well-recognized challenge in human activity recognition that often leads to lower performance for underrepresented tasks (Alani et al., 2020; Alharbi et al., 2022; Singh et al., 2025). This challenge was also evident in our dataset: box pushing and box pulling had fewer sequences, while box carrying had the most, which may have biased the model toward the more frequent classes. This data imbalance results in a limited temporal context for underrepresented classes, making it challenging for VLMs to learn frame-level patterns and ultimately increasing the possibility of misclassification (Alharbi et al., 2022; De Giorgio et al., 2023; Hamad et al., 2020). Although we applied two common strategies to mitigate this—weighted random sampling during training and weighted loss functions—an imbalanced representation still appeared to affect classification performance (Figs. 3 and 4, and A.5). Others have also found that such strategies may not fully address class imbalance in deep learning models (Dablain et al., 2024; Johnson and Khoshgoftaar, 2019; Pouyanfar et al., 2018). Thus, class imbalance remains a critical limitation in applying VLMs to the classification of diverse MMH tasks and warrants attention in future experimental designs and model development.

4.2. Effects of different VLMs on MMH task classification and temporal segmentation performance

MMH task classification and temporal segmentation performance differed across three VLMs. DINO consistently outperformed MAE and CLIP for both general MMH tasks (Figs. 3 and 4, A.5, and A.8) and task conditions (Fig. 7, A.6, and A.9), although the magnitude of differences was small in some cases. Similar patterns have been reported in prior studies on general video or image datasets, in which DINO yielded stronger performance on fine-grained spatial features in human–object interaction tasks (e.g., hand placement, body posture, and object location) vs. MAE or CLIP (Bertasius et al., 2021; Caron et al., 2021; Charisoudis et al., 2023; Jiang et al., 2024; Vanyan et al., 2024). These differences may be attributed to the underlying learning objectives of such models and how those objectives affect spatial feature extraction in MMH tasks. Prior work has shown that self-distillation improves spatial representations for human–object interaction recognition (Bertasius et al., 2021; Caron et al., 2021). Extending these findings to MMH tasks, our results show that DINO extracted more effective features for both classification and temporal segmentation vs. MAE or CLIP. MAE achieved comparable classification performance but was less reliable for detailed MMH task conditions such as hand configuration, consistent with earlier reports that its learned features do not generalize as well across different tasks (Charisoudis et al., 2023; Vanyan et al., 2024). CLIP was the least effective model here, likely because it prioritizes global semantic features over localized task-relevant cues, a limitation also noted in prior work (Jiang et al., 2024; Radford et al., 2019). Thus, careful model selection is critical in applying VLMs to effectively classify and segment MMH tasks.

Temporal segmentation performance in our study was relatively lower than that reported in some earlier work, but the comparisons are not direct because those studies often used different modalities and task types. For example, Porta et al. (2021) used wearable sensors and achieved relatively large accuracy (~97–99%) in estimating task duration and frequency. Romeo et al. (2025) reported frame-wise accuracies of ~91–94% for industrial assembly tasks when combining RGB, depth, and skeletal data, with performance dropping to ~78–81% under cross-location validation. Other work has integrated segmentation with ergonomics risk prediction (Parsa and Banerjee, 2020) or graph-based activity modeling (Benmessabih et al., 2025; Huang et al., 2020), typically requiring multimodal sensing inputs. In contrast, our approach shows that segmentation of diverse MMH tasks is feasible *using only regular RGB video streams*, suggesting a potentially practical and scalable alternative for real environments.

4.3. Classification of specific MMH task conditions

To our knowledge, no directly comparable studies have examined automated classification of MMH task conditions (i.e., lift origin and hand configuration), aside from the markerless motion capture–based work by Ojelade et al. (2025). Compared to that earlier work, our VLM-based approach exhibited noticeable advantages in classifying specific MMH task conditions. Specifically, our models achieved larger mean classification accuracies (~94–98%) for lift origin vs. the ~80–84% reported by Ojelade et al. (2025), despite their use of kinematic data from depth-enabled MMC systems and recurrent neural networks. This improvement likely reflects the ability of the VLMs to extract detailed spatial information from RGB video and to detect task-relevant features for lifting, such as lifting posture and box location. Yet, classifying hand configuration remained more challenging in both studies. Ojelade et al. (2025) reported the largest F1-scores for the one-hand configuration (~97%), with lower performance for narrow and broad configurations, which align closely with the performance patterns observed here (F1-scores = ~87–84%; Fig. 7 and A.9). This limitation is likely driven primarily by sensing constraints; low-resolution or distant RGB video may lack the fidelity to capture subtle differences in hand placement and grip width, while the use of frozen ViT features could further limit sensitivity to such fine-grained visual cues.

4.4. Differences related to biological sex in classifying MMH tasks

Classification Precision was generally greater among males vs. females (Fig. 5), whereas Ojelade et al. (2025) reported lower Precision among males. This difference may be related to both methodological and sample differences between the two studies. Prior research has shown sex-related variations in lifting and posture, with females often using greater lower-limb flexion, smaller trunk flexion, and different spinal loading patterns compared to males (Marras et al., 2003; Marras and Davis, 1998; Plamondon et al., 2014, 2017). However, part of the observed difference may also be due to sample imbalance in our dataset (22 males vs. 14 females), which can bias model learning toward the majority group (Alani et al., 2020; Singh et al., 2025). In addition, recent work has shown that VLMs can inherit biological sex–activity binding bias, where certain activities are more strongly associated with one sex, which can reduce performance when the performer's sex differs from the models learned associations (Abdollahi et al., 2024). These findings together suggest that both biomechanical and data-related factors may have contributed to the sex-related variability observed here. We recommend future work to clarify sex-related effects on MMH task classification.

4.5. Implications of the leave-one-subject-out validation method

Several earlier, analogous studies used different approaches for quantifying the classification performance of machine learning models (e.g., random splits or other conventional validation methods). We adopted a more challenging "leave-one-subject-out approach" (LOSO) and still achieved classification performance comparable to prior work (e.g., Bayat et al., 2014; Chen et al., 2017; Cleland et al., 2013; Kim and Nussbaum, 2014). Although this approach has been used in some earlier studies (e.g., Ojelade et al., 2025; Porta et al., 2021), it often results in

high variance in accuracy because different participants can perform the same tasks in different ways (Li et al., 2020). Accordingly, the relatively large variance we found can be attributed primarily to the LOSO validation method. But as we noted earlier, we believe this validation approach better represents likely performance in actually application, in which a model is trained offline on known participants and applied to unseen ones—an essential condition for assessing generalizability (Jordao et al., 2018; Li et al., 2020).

4.6. Limitations

Several limitations of our study should be mentioned. All MMH tasks were performed by a relatively young (i.e., 18–39 years old) and healthy group under controlled laboratory conditions (e.g., participant clothing, lighting conditions, and camera positioning). In real environments, factors such as visual occlusion, background clutter, the use of personal protective equipment, and greater lighting variability could adversely affect vision-based model performance, although the magnitude of these effects is currently unknown. Thus, caution should be taken in generalizing the findings to other populations and settings. Similarly, the use of a three-camera setup may limit scalability, and our results may represent an upper bound on performance compared to fewer (or one) cameras. There could be value in exploring the performance of VLMs across more diverse populations and in real environments using fewer cameras under less controlled conditions.

We also used ViTs pretrained on general-purpose datasets with frozen parameters. We did not fine-tune the ViTs on MMH tasks, which may have limited the ability of the model to extract the most effective features for efficient MMH task classification (e.g., Bertasius et al., 2021; Dosovitskiy et al., 2020). We also used the GPT-based temporal model with fixed-length (100-frame) sequences; this design choice can truncate longer tasks or pad shorter ones, potentially reducing classification and segmentation performance for variable-duration tasks. We suggest that future work explore fine-tuning ViTs on MMH-specific data and developing adaptive sequence models that better handle variable task lengths and temporal dependencies.

4.7. Practical applications of our findings

One advantage of our approach is that it relies solely on regular RGB video. While the overall computational cost of the method remains to be determined, using regular RGB video itself offers a non-contact and relatively cost-effective alternative to several previous physical exposure assessment approaches—including self-reports, observations, and wearable sensors—the limitations of which were summarized earlier. This advantage might enhance scalability and usability in real environments, since regular RGB cameras are widely available, relatively inexpensive, and already commonly installed in many workplaces.

Further, our VLM-based approach achieved satisfactory performance in classifying and temporally segmenting MMH tasks, comparable to—and in some cases exceeding—performance reported in prior studies (e.g., Ojelade et al., 2025). Importantly, our frame-wise classification approach enables more fine-grained temporal segmentation compared to sequence-to-sequence models (e.g., Ojelade et al., 2025) or sample-by-sample approaches (Yin et al., 2017). Enhanced temporal resolution could prove critical for assessing physical exposures in dynamic work environments (e.g., order picking) wherein rapid task changes can occur frequently.

5. Conclusions

Effectively and accurately quantifying physical exposures during MMH tasks is essential for assessing and controlling the risk of WMSDs. While several assessment tools are available, they are often limited in scope, accuracy, or feasibility in real environments. We examined a VLM-based approach using three ViTs and a GPT-based model to classify eight MMH tasks and two task conditions—lift origin and hand configuration—and to temporally segment task boundaries. General MMH task classification demonstrated good performance, with a mean accuracy of ~82–85%, and with Precision, Recall, and F1-scores ranging from ~70 to 90%. Temporal segmentation performance was comparable to values reported in related work, with a mean IoU of ~74–77% and mean absolute errors of ~0.50 s and ~0.38 s for task start and end times, respectively. Classification of lift origin yielded notably large accuracy (~94–98%), while hand configuration classification showed greater variability between different hand configurations and models. Overall, our proposed approach shows promise for efficiently and effectively quantifying MMH task exposures, providing a practical balance of simplicity and non-invasiveness. However, further research is needed to evaluate its performance across diverse worker populations and in real occupational settings.

CRediT authorship contribution statement

Mohammad Sadra Rajabi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Aanuoluwapo Ojelade:** Conceptualization, Data curation, Validation, Writing – original draft, Writing – review & editing. **Sunwook Kim:** Resources, Visualization, Writing – original draft, Writing – review & editing. **Maury A. Nussbaum:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

Acknowledgements

The first author was supported by a predoctoral training program grant (T03 OH008613) from CDC/NIOSH. The current contents are solely the authors' responsibility and do not necessarily represent the official views of CPWR, NIOSH, or the CDC. The authors thank Advanced Research Computing at Virginia Tech for providing computational resources and technical support that contributed to the results reported in this paper. URL: https://arc.vt.edu/

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apergo.2026.104831.

References

Abdollahi, A., Ghaznavi, M., Nejad, M.R.K., Oriyad, A.M., Abbasi, R., Salesi, A., Behjati, M., Rohban, M.H., Baghshah, M.S., 2024. GABInsight: exploring gender-activity binding bias in vision-language models. https://doi.org/10.3233/FAIA240555.

Alani, A.A., Cosma, G., Taherkhani, A., 2020. Classifying imbalanced multi-modal sensor data for human activity recognition in a smart home using deep learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207697.

Alharbi, F., Ouarbya, L., Ward, J.A., 2022. Comparing sampling strategies for tackling imbalanced data in human activity recognition. Sensors 22 (4), 1373. https://doi.org/10.3390/s22041373.

Anacleto Filho, P.C., Colim, A., Jesus, C., Lopes, S.I., Carneiro, P., 2024. Digital and virtual technologies for work-related biomechanical risk assessment: a scoping review. Saf. Now 10 (3), 79. https://doi.org/10.3390/safety10030079.

Anam, M.K., Defit, S., Haviluddin, H., Efrizoni, L., Firdaus, M.B., 2024. Early stopping on CNN-LSTM development to improve classification performance. J. Appl. Data Sci. 5 (3), 1175–1188.

Antwi-Afari, M.F., Li, H., Yu, Y., Kong, L., 2018. Wearable insole pressure system for automated detection and classification of awkward working postures in construction workers. Autom. ConStruct. 96, 433–441. https://doi.org/10.1016/j.autcon.2018.10.004.

Barazandeh, B., Bastani, K., Rafieisakhaei, M., Kim, S., Kong, Z., Nussbaum, M.A., 2018. Robust sparse representation-based classification using online sensor data for monitoring manual material handling tasks. IEEE Trans. Autom. Sci. Eng. 15 (4), 1573–1584. https://doi.org/10.1109/TASE.2017.2729583.

Bayat, A., Pomplun, M., Tran, D.A., 2014. A study on human activity recognition using accelerometer data from smartphones. Procedia Comput. Sci. 34, 450–457. https://doi.org/10.1016/j.procs.2014.07.009.

Belcaid, M., Gonzalez Martinez, A., Leigh, J., 2022. Leveraging deep contrastive learning for semantic interaction. PeerJ Comput. Sci. 8, e925. https://doi.org/10.7717/peerj-cs.925.

Benmessabih, T., Slama, R., Havard, V., Baudry, D., 2025. Graph-based framework for temporal human action recognition and segmentation in industrial context. Eng. Appl. Artif. Intell. 159, 111710. https://doi.org/10.1016/j.engappai.2025.111710.

Bertasius, G., Wang, H., Torresani, L., 2021. Is space-time attention all you need for video understanding?, 2 (3), 4.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in self-supervised Vision Transformers, pp. 9650–9660.

Charisoudis, A., Huth, S. E. von, Jansson, E., 2023. [Re] masked autoencoders are small scale vision learners: a reproduction under resource constraints. ReScience C 9 (2), 40. http://doi.org/10.5281/zenodo.8173751.

Chen, J., Qiu, J., Ahn, C., 2017. Construction worker's awkward posture recognition through supervised motion tensor decomposition. Autom. ConStruct. 77, 67–81. https://doi.org/10.1016/j.autcon.2017.01.020.

Cleland, I., Kikhia, B., Nugent, C., Boytsov, A., Hallberg, J., Synnes, K., McClean, S., Finlay, D., 2013. Optimal placement of accelerometers for the detection of everyday activities. Sensors 13 (7), 9183–9200. https://doi.org/10.3390/s130709183.

Dablain, D., Bellinger, C., Krawczyk, B., Aha, D.W., Chawla, N., 2024. Understanding imbalanced data: XAI & interpretable ML framework. Mach. Learn. 113 (6), 3751–3769. https://doi.org/10.1007/s10994-023-06414-w.

David, G.C., 2005. Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. Occup. Med. 55 (3), 190–199. https://doi.org/10.1093/occmed/kqi082.

De Giorgio, A., Cola, G., Wang, L., 2023. Systematic review of class imbalance problems in manufacturing. J. Manuf. Syst. 71, 620–644. https://doi.org/10.1016/j.jmsy.2023.10.014.

Dempsey, P.G., McGorry, R.W., Maynard, W.S., 2005. A survey of tools and methods used by certified professional ergonomists. Appl. Ergon. 36 (4), 489–503. https://doi.org/10.1016/j.apergo.2005.01.007.

Deros, B.M., Daruis, D.D.I., Rosly, A.L., Abd Aziz, I., Hishamuddin, N.S., Abd Hamid, N. H., Roslin, S.M., 2017. Ergonomic risk assessment of manual material handling at an automotive manufacturing company. PressAcademia Procedia 5 (1), 317–324.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: transformers for image recognition at scale (version 2). arXiv. https://doi.org/10.48550/ARXIV.2010.11929.

Fan, C., Mei, Q., Li, X., 2024a. Assisting in the identification of ergonomic risks for workers: a large vision-language model approach. In: ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, vol. 41. SciTech Premium Collection, pp. 1010–1017 (3092410938).

Fan, C., Mei, Q., Wang, X., Li, X., 2024b. ErgoChat: a visual query system for the ergonomic risk assessment of construction workers (arXiv:2412.19954). arXiv. https://doi.org/10.48550/arXiv.2412.19954.

Ferrara, E., 2024. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: a survey of early trends, datasets, and challenges. Sensors 24 (15), 5045. https://doi.org/10.3390/s24155045.

Gholamiangonabadi, D., Kiselov, N., Grolinger, K., 2020. Deep neural networks for human activity recognition with wearable sensors: leave-one-subject-out cross-validation for model selection. IEEE Access 8, 133982–133994. https://doi.org/10.1109/ACCESS.2020.3010715.

Golabchi, A., Han, S., Fayek, A.R., 2016. A fuzzy logic approach to posture-based ergonomic analysis for field observation and assessment of construction manual operations. Can. J. Civ. Eng. 43 (4), 294–303. https://doi.org/10.1139/cjce-2015-0143.

Guner, U., Dasdemir, J., 2023. Novel self-calibration method for IMU using distributed inertial sensors. IEEE Sens. J. 23 (2), 1527–1540. https://doi.org/10.1109/JSEN.2022.3227341.

Hamad, R.A., Yang, L., Woo, W.L., Wei, B., 2020. Joint learning of temporal models to handle imbalanced data for human activity recognition. Appl. Sci. 10 (15), 5293. https://doi.org/10.3390/app10155293.

Hoozemans, M.J., Kuijer, P.P.F., Kingma, I., Van Dieën, J.H., De Vries, W.H., Van Der Woude, L.H., Veeger, D.J.H.E.J., Van Der Beek, A.J., Frings-Dresen, M.H., 2004. Mechanical loading of the low back and shoulders during pushing and pulling activities. Ergonomics 47 (1), 1–18. https://doi.org/10.1080/00140130310001593577.

Huang, Y., Sugano, Y., Sato, Y., 2020. Improving Action Segmentation via Graph-based Temporal Reasoning, pp. 14024–14034 https://openaccess.thecvf.com/content_CVPR_2020/html/Huang_Improving_Action_Segmentation_via_Graph-Based_Temporal_Reasoning_CVPR_2020_paper.html.

Jiang, D., Liu, Y., Liu, S., Zhao, J., Zhang, H., Gao, Z., Zhang, X., Li, J., Xiong, H., 2024. From CLIP to DINO: visual encoders shout in multi-modal large language models (arXiv:2310.08825). arXiv. https://doi.org/10.48550/arXiv.2310.08825.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6 (1), 27. https://doi.org/10.1186/s40537-019-0192-5.

Jordao, A., Nazare, A.C., Sena, J., Schwartz, W.R., 2018. Human activity recognition based on wearable sensor data: a standardization of the state-of-the-art (version 3). arXiv. https://doi.org/10.48550/ARXIV.1806.05226.

Kadikon, Y., Rahman, M.N.A., 2016. Manual material handling risk assessment tool for assessing exposure to risk factor of work-related musculoskeletal disorders: a review. J. Eng. Appl. Sci. 100 (10), 2226–2232.

Khosrowpour, A., Niebles, J.C., Golparvar-Fard, M., 2014. Vision-based workface assessment using depth images for activity analysis of interior construction operations. Autom. ConStruct. 48, 74–87. https://doi.org/10.1016/j.autcon.2014.08.003.

Kim, W., Sung, J., Saakes, D., Huang, C., Xiong, S., 2021. Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose). Int. J. Ind. Ergon. 84, 103164. https://doi.org/10.1016/j.ergon.2021.103164.

Kim, S., Nussbaum, M.A., 2014. An evaluation of classification algorithms for manual material handling tasks based on data obtained using wearable technologies. Ergonomics 57 (7), 1040–1051. https://doi.org/10.1080/00140139.2014.907450.

Laurençon, H., Tronchon, L., Cord, M., Sanh, V., 2024. What matters when building vision-language models? (arXiv:2405.02246). arXiv. https://doi.org/10.48550/arXiv.2405.02246.

Leggieri, S., Fanti, V., Caldwell, D.G., Di Natali, C., 2023. Online ergonomic evaluation in realistic manual material handling task: proof of concept. Bioengineering 11 (1), 14. https://doi.org/10.3390/bioengineering11010014.

Li, H., Shrestha, A., Heidari, H., Le Kernec, J., Fioranelli, F., 2020. Bi-LSTM network for multimodal continuous human activity recognition and fall detection. IEEE Sens. J. 20 (3), 1191–1201. https://doi.org/10.1109/JSEN.2019.2946095.

Li, G., Buckle, P., 1999. Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture-based methods. Ergonomics. https://doi.org/10.1080/001401399185388 (world).

Lind, C.M., Abtahi, F., Forsman, M., 2023. Wearable motion capture devices for the prevention of work-related musculoskeletal disorders in ergonomics—an overview of current applications, challenges, and future opportunities. Sensors 23 (9), 4259. https://doi.org/10.3390/s23094259.

Marras, W.S., Davis, K.G., Jorgensen, M., 2003. Gender influences on spine loads during complex lifting. Spine J. 3 (2), 93–99. https://doi.org/10.1016/S1529-9430(02)00570-3.

Marras, W.S., Davis, K.G., 1998. Spine loading during asymmetric lifting using one versus two hands. Ergonomics 41 (6), 817–834. https://doi.org/10.1080/001401398186667.

Menanno, M., Riccio, C., Benedetto, V., Gissi, F., Savino, M.M., Troiano, L., 2024. An ergonomic risk assessment system based on 3D human pose estimation and collaborative robot. Appl. Sci. 14 (11), 4823. https://doi.org/10.3390/app14114823.

Mudiyanselage, S.E., Nguyen, P.H.D., Rajabi, M.S., Akhavian, R., 2021. Automated workers' ergonomic risk assessment in manual material handling using sEMG wearable sensors and machine learning. Electronics 10 (20), 2558. https://doi.org/10.3390/electronics10202558.

Nath, N.D., Chaspari, T., Behzadan, A.H., 2018. Automated ergonomic risk monitoring using body-mounted sensors and machine learning. Adv. Eng. Inform. 38, 514–526. https://doi.org/10.1016/j.aei.2018.08.020.

Ojelade, A., Rajabi, M.S., Kim, S., Nussbaum, M.A., 2025. A data-driven approach to classifying manual material handling tasks using markerless motion capture and recurrent neural networks. Int. J. Ind. Ergon. https://doi.org/10.1016/j.ergon.2025.103755.

Parsa, B., Banerjee, A.G., 2020. A multi-task learning approach for human activity segmentation and ergonomics risk assessment (arXiv:2008.03014). arXiv. https://doi.org/10.48550/arXiv.2008.03014.

Plamondon, A., Larivière, C., Denis, D., Mecheri, H., Nastasia, I., 2017. Difference between male and female workers lifting the same relative load when palletizing boxes. Appl. Ergon. 60, 93–102. https://doi.org/10.1016/j.apergo.2016.10.014.

Plamondon, A., Larivière, C., Denis, D., St-Vincent, M., Delisle, A., 2014. Sex differences in lifting strategies during a repetitive palletizing task. Appl. Ergon. 45 (6), 1558–1569. https://doi.org/10.1016/j.apergo.2014.05.005.

Porta, M., Kim, S., Pau, M., Nussbaum, M.A., 2021. Classifying diverse manual material handling tasks using a single wearable sensor. Appl. Ergon. 93, 103386. https://doi.org/10.1016/j.apergo.2021.103386.

Pouyanfar, S., Tao, Y., Mohan, A., Tian, H., Kaseb, A.S., Gauen, K., Dailey, R., Aghajanzadeh, S., Lu, Y.-H., Chen, S.-C., Shyu, M.-L., 2018. Dynamic sampling in convolutional neural networks for imbalanced data classification. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 112–117. https://doi.org/10.1109/MIPR.2018.00027.

Punnett, L., Wegman, D.H., 2004. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. J. Electromyogr. Kinesiol. 14 (1), 13–23. https://doi.org/10.1016/j.jelekin.2003.09.015.

Qin, R., Qiao, K., Wang, L., Zeng, L., Chen, J., Yan, B., 2018. Weighted focal loss: an effective loss function to overcome unbalance problem of chest X-ray14. IOP Conf. Ser. Mater. Sci. Eng. 428, 012022. https://doi.org/10.1088/1757-899X/428/1/012022.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.

Rajabi, M.S., Ojelade, A., Kim, S., Nussbaum, M.A., 2025. Classifying diverse manual material handling tasks using vision transformers and recurrent neural networks. Proc. Hum. Factors Ergon. Soc. Annu. Meet., 10711813251367009 https://doi.org/10.1177/10711813251367009.

Ren, S., Yao, L., Li, S., Sun, X., Hou, L., 2024. TimeChat: a time-sensitive multimodal large language model for long video understanding (arXiv:2312.02051). arXiv. http://arxiv.org/abs/2312.02051.

Rezaei-Dastjerdehei, M.R., Mijani, A., Fatemizadeh, E., 2020. Addressing imbalance in multi-label classification using weighted cross entropy loss function. In: 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), pp. 333–338. https://doi.org/10.1109/ICBME51989.2020.9319440.

Romeo, L., Marani, R., Perri, A.G., Gall, J., 2025. Multi-modal temporal action segmentation for manufacturing scenarios. Eng. Appl. Artif. Intell. 148, 110320. https://doi.org/10.1016/j.engappai.2025.110320.

Sabiri, B., Khtira, A., El Asri, B., Rhanoui, M., 2024. Investigating contrastive pair learning's frontiers in supervised, semisupervised, and self-supervised learning. J. Imag. 10 (8), 196. https://doi.org/10.3390/jimaging10080196.

Santiago-Colón, A., Rocheleau, C.M., Bertke, S., Christianson, A., Collins, D.T., Trester-Wilson, E., Sanderson, W., Waters, M.A., Reefhuis, J., National Birth Defects Prevention Study, 2021. Testing and validating semi-automated approaches to the occupational exposure assessment of polycyclic aromatic hydrocarbons. Ann. Work Exposures Health 65 (6), 682–693. https://doi.org/10.1093/annweh/wxab002.

Scabini, L., Sacilotti, A., Zielinski, K.M., Ribas, L.C., Baets, B.D., Bruno, O.M., 2024. A comparative survey of vision transformers for feature extraction in texture analysis (arXiv:2406.06136). arXiv. https://doi.org/10.48550/arXiv.2406.06136.

Schall, M.C., Sesek, R.F., Cavuoto, L.A., 2018. Barriers to the adoption of wearable sensors in the workplace: a survey of occupational safety and health professionals. Hum. Factors: J. Hum. Factors Ergon. Soc. 60 (3), 351–362. https://doi.org/10.1177/0018720817753907.

Seo, J., Lee, S., 2021. Automated postural ergonomic risk assessment using vision-based posture classification. Autom. ConStruct. 128, 103725. https://doi.org/10.1016/j.autcon.2021.103725.

Singh, D., Merdivan, E., Kropf, J., Holzinger, A., 2025. Class imbalance in multi-resident activity recognition: an evaluative study on explainability of deep learning approaches. Univers. Access Inf. Soc. 24 (2), 1173–1191. https://doi.org/10.1007/s10209-024-01123-0.

Sochopoulos, A., Poliero, T., Ahmad, J., Caldwell, D.G., Di Natali, C., 2025. Human activity recognition algorithms for manual material handling activities. Sci. Rep. 15 (1), 10954. https://doi.org/10.1038/s41598-024-81312-2.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45 (4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Spielholz, P., Silverstein, B., Morgan, M., Checkoway, H., Kaufman, J., 2001. Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors. Ergonomics 44 (6), 588–613. https://doi.org/10.1080/00140130118050.

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Zhang, J., Luo, P., Luo, J., Xu, C., 2024. Video understanding with large language models: a survey (arXiv:2312.17432). arXiv. http://arxiv.org/abs/2312.17432.

Tong, Z., Song, Y., Wang, J., Wang, L., 2022. Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. Adv. Neural Inf. Process. Syst. 35, 10078–10093.

Trubakov, A., Trubakova, A., 2020. Method for Improving an Image Segment in a Video Stream Using Median Filtering and Blind Deconvolution Based on Evolutionary Algorithms, 2744, pp. 1–11.

Tzelepi, M., Mezaris, V., 2024. Exploiting LMM-based knowledge for image classification tasks. In: Iliadis, L., Maglogiannis, I., Papaleonidas, A., Pimenidis, E., Jayne, C. (Eds.), Engineering Applications of Neural Networks, 2141. Springer, Nature Switzerland, pp. 166–177. https://doi.org/10.1007/978-3-031-62495-7_13.

Vanyan, A., Barseghyan, A., Tamazyan, H., Huroyan, V., Khachatrian, H., Danelljan, M., 2024. Analyzing local representations of self-supervised vision transformers (arXiv:2401.00463). arXiv. https://doi.org/10.48550/arXiv.2401.00463.

Wang, J., Kou, L., Kwan, M.-P., Shakespeare, R.M., Lee, K., Park, Y.M., 2021. An integrated individual environmental exposure assessment system for real-time Mobile sensing in environmental health studies. Sensors 21 (12), 4039. https://doi.org/10.3390/s21124039.

Yang, J., Shi, Z., Wu, Z., 2016. Vision-based action recognition of construction workers using dense trajectories. Adv. Eng. Inform. 30 (3), 327–336. https://doi.org/10.1016/j.aei.2016.04.009.

Yin, W., Kann, K., Yu, M., Schütze, H., 2017. Comparative study of CNN and RNN for natural language processing (arXiv:1702.01923). arXiv. https://doi.org/10.48550/arXiv.1702.01923.

Yong, G., Liu, M., Lee, S., 2024. Automated captioning for ergonomic problem and solution identification in construction using a vision-language model and caption augmentation. Constr. Res. Congr. 709–718. https://doi.org/10.1061/9780784485293.071, 2024.

Zhang, H., Li, X., Bing, L., 2023. Video-LLaMA: an instruction-tuned audio-visual language model for video understanding (arXiv:2306.02858). arXiv. https://doi.org/10.48550/arXiv.2306.02858.

Zhang, J., Huang, J., Jin, S., Lu, S., 2024. Vision-language models for vision tasks: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 46 (8), 5625–5644. https://doi.org/10.1109/TPAMI.2024.3369699.

Zhao, Y.S., Jaafar, M.H., Mohamed, A.S.A., Azraai, N.Z., Amil, N., 2022. Ergonomics risk assessment for manual material handling of warehouse activities involving high shelf and low shelf binning processes: application of marker-based motion capture. Sustainability 14 (10), 5767. https://doi.org/10.3390/su14105767.