applied sciences

MDPI

Article

Vision–Language Model (VLM)-Based Interactive Visual Querying for Construction Ergonomic Risk Reasoning

Chao Fan [1], Qipei Mei [2], Xiaonan Wang [3] and Xinming Li [1,*]

[1] Department of Mechanical Engineering, University of Alberta, 9211 116 St NW, Edmonton, AB T6G 1H9, Canada; cfan3@ualberta.ca
[2] Department of Civil and Environmental Engineering, University of Alberta, 9211 116 St NW, Edmonton, AB T6G 1H9, Canada; qipei.mei@ualberta.ca
[3] Department of Electrical and Computer Engineering, University of Alberta, 9211 116 St NW, Edmonton, AB T6G 1H9, Canada; xiaonan6@ualberta.ca
* Correspondence: xinming.li@ualberta.ca

Featured Application

The proposed health and safety domain-specific system can support construction safety practitioners by analyzing workplace images, answering questions about workers' postural ergonomic risks, and generating clear natural-language descriptions of identified hazards. It may be applied to early risk detection, ergonomic inspections, safety documentation, and personnel training, particularly when access to ergonomics experts is limited.

Abstract

Construction workers are frequently exposed to postural ergonomic risks, yet conventional AI-based ergonomic risk assessment (ERA) systems generally provide risk classifications or scores rather than interpretable and interactive explanations. This study investigates whether domain adaptation can improve a generic vision–language model's ability to identify and explain construction ergonomic risks from images. A dataset with 1900 construction image–text pairs was curated. The domain-adapted model was compared with the baseline model. Performance was evaluated using visual question answering (VQA) accuracy, nine image-captioning metrics, and a blinded human evaluation involving 50 participants with varying levels of ergonomics knowledge. ErgoChat improved performance across the nine caption-evaluation metrics and VQA. In the human evaluation, ErgoChat-generated descriptions were selected as more accurate in 81.69% of comparisons. Limitations include dataset size and class imbalance, potential VLM hallucination, and visual-perception errors. Future research will expand and balance real-site data, evaluate prompt robustness, and strengthen integration with structured ergonomic assessment procedures. The primary contribution is therefore a construction-ergonomics-specific domain-adaptation and evaluation framework that enables interactive VQA and natural-language ergonomic risk reasoning.

Keywords: AI; VLM; safety; ergonomic risk assessment; dependability; reliability; construction; ergonomics

Check for updates

Academic Editor: Andrea Prati

Received: 10 August 2026
Revised: 30 September 2026
Accepted: 1 October 2026
Published: 6 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In the construction industry, workers often perform repetitive movements, intense physical labor, and tasks that require awkward postures, all of which can damage health. Therefore, timely identification and elimination of these risks is extremely important for ensuring safety, maintaining health, and increasing productivity [1]. Statistics from the

United States Bureau of Labor Statistics [1] show that construction ranked fifth among industries for work-related musculoskeletal disorders (WMSD)-related injuries and illnesses. Moreover, in the same year, WMSDs represented 30% of all occupational injuries and illnesses [1]. According to the Association of Workers' Compensation Boards of Canada, the manufacturing and construction sectors reported the second and third-highest rates of lost-time injury claims in 2021, representing 13.6% and 10.4% of claims, respectively [2]. European Agency for Safety and Health at Work reported that the construction and manufacturing sectors reported the highest sick leave rates due to WMSDs [3]. Therefore, early identification and proactive prevention of WMSDs caused by postural ergonomic risks are crucial for companies and construction industry workers.

Conventional postural ergonomic risk assessment techniques include self-reporting, observational assessment, and direct measurement approaches [4–11]. Self-reporting and observational methods are limited by the need for manual data collection, which includes the significant time required by ergonomic experts and inconsistencies in data interpretation. Alternatively, the direct measurement method employs sensors or markers attached to workers and uses a motion capture system to evaluate their postural ergonomic risks. While this type of method is more time-efficient and objective compared to self-reporting and observational techniques, the attached sensors or markers can create psychological and physical burdens on the workers, potentially impacting their work performance. Overall, these limitations reduce the effectiveness of traditional ERA methods.

To overcome the challenges of subjectivity and the time-intensive nature of traditional ERA methods, predictive artificial intelligence (AI) has been widely used to address these limitations. Traditional machine learning (ML) encompasses AI systems that employ statistical analysis and algorithms to assess risk exposure. In the context of ERA, traditional ML typically involves the application of these trained models to interpret data gathered from sensors worn on the body; these trained models are used to determine whether workers are exposed to ergonomic risks based on the data received from the sensors on the workers [12]. However, it is crucial to acknowledge that while these sensor-based or marker-based traditional ML methods address the issues of low time efficiency, subjectivity, and inconsistency inherent in traditional approaches, they do not resolve the problem of intrusiveness [10]. On the other hand, computer vision (CV)-based traditional ML methods [5–9,13] can avoid introducing intrusiveness into ERA while overcoming the drawbacks of subjectivity, inconsistency, and time consumption. However, an unavoidable drawback of CV-based ERA methods is that occlusions in images or videos can significantly impact accuracy [13].

Unlike traditional ML, transformer-based large language models (LLMs) have motivated research across many domains [14]. This surge in transformer-based LLMs highlights research opportunities in ERA. Existing ERA methods face two main issues. Firstly, current ergonomic VLMs are limited by scarce domain data and training resources. Secondly, AI-based predictive ERA can identify risks but struggles to provide clear, actionable natural language explanations.

This study focuses on the construction industry, which is highly sensitive to WMSDs. To achieve high accuracy in postural ERA, the study adopts an image-based approach and uses the vision transformer (ViT) to process images. This approach was selected because postural ergonomic risk assessment requires visual information, while transformer-based language models enable the generation of natural-language interpretations of the identified risks. This approach provides functionalities that go beyond traditional ML-based ERA: (1) answering visual questions related to postural risks; and (2) producing captions that describe workers' actions and the related postural risks. It is noteworthy to emphasize the scarcity of publicly available datasets explicitly tailored for training or evaluating VLMs

with construction ergonomic risk knowledge. Therefore, this study has also created such a dataset. As a result, an interactive postural ERA approach, ErgoChat, was developed to fulfill the two functions mentioned above.

The proposed interactive visual query system brings knowledge into the application by turning ergonomic assessment principles into computational models that can interpret and analyze visual data. ERA requires precise analysis of subtle posture characteristics, task context, and expertise-scoring logic, and generic VLMs struggle to capture these aspects of expertise. For this purpose, this research combed through the ergonomic knowledge base, transformed it into the data resources needed for multimodal instruction-tuning, and constructed a specialized annotation scheme. Comparative experiments indicated that the generic VLM achieved lower performance than the domain-adapted model on the evaluated tasks. Although ErgoChat is based on MiniGPT-v2, the contribution of this work lies in the creation of a domain-customized multimodal reasoning framework dedicated to ERA. This framework is not focused on innovation in the VLM backbone, but rather on converting expert rules into executable reasoning capabilities and paving the way for a deeper integration of ERA and AI.

Generic VLMs still struggle to accurately translate visual information into ergonomic risk judgment. Without targeted training or fine-tuning, these models struggle to achieve performance improvement in generating structured ergonomic explanations. Although ErgoChat is built on an existing VLM, its research focus is on bridging capability gaps in professional scenarios and driving strong alignment between technology and actual needs. This study makes several contributions, including (1) a domain-adapted multimodal framework that uses VLMs to generate natural-language explanations of postural ergonomic risks. (2) a dataset consisting of 1900 different image–text pairs designed explicitly for identifying postural ergonomic risks among construction workers, (3) a specialized instruction tuning strategy that converts ergonomic evaluation rules into a form that VLMs can learn, enabling the model to generate domain-specific ergonomic risk judgments and explanations, and (4) a quantitative and qualitative evaluation using nine metrics and feedback from 50 ergonomic evaluators.

2. Related Work

Construction safety has been extensively studied. Although injuries and deaths resulting from safety accidents are more catastrophic, the damage to workers' physical health caused by WMSDs due to ergonomic risks is a widespread issue that should not be overlooked [15]. As researchers and safety practitioners become increasingly aware of the prevalence of WMSDs caused by ergonomic risks in the workplace, more postural ERA tools have been proposed [16]. The widely recognized classification categorizes these postural ERA tools into self-reporting, manual observation, direct measurement, and more advanced technological methods [13]. Self-report, observation, and direct measurement approaches that do not incorporate AI are collectively referred to as traditional methods in the following text. ERA tools that employ traditional ML, such as those that use CV to analyze joint angles, are referred to as traditional ML-based methods. This section examines commonly used ERA tools in the construction sector.

2.1. Traditional ERA

Self-report methods require researchers to collect ergonomic risk information from workers through surveys, interviews, or worker diaries [17]. The Standardized Nordic Questionnaire is an example of a self-report-based assessment method [18]. A detailed version of the Standardized Nordic Questionnaire has been shown to be highly reliable for identifying WMSD in both occupational and general population health [19]. Researchers

and safety practitioners have developed questionnaires for specific industries. Despite optimizing many self-report methods, they share drawbacks due to inconsistent perception, comprehension, and interpretation among respondents [19]. Other disadvantages include being time-consuming and prone to errors [20,21].

Observational methods evaluate ergonomic risks associated with workers' postures using established tools, following the guidelines provided by those tools. Observational methods can be implemented through direct observation of workers or through analysis of video recordings. The most common observation method-based assessment tools include Rapid Entire Body Assessment (REBA), Rapid Upper Limb Assessment (RULA), and the revised National Institute for Occupational Safety & Health (NIOSH) lifting equation [22,23]. REBA is an ergonomic assessment tool used to evaluate workers' risk of whole-body WMSDs [24]. RULA is an ergonomic evaluation tool designed to assess workers' risk of upper-body WMSDs [25]. The revised NIOSH lifting equation is an observational tool used to assess the physical requirements for manual weight-lifting tasks and to determine the potential risks of WMSDs. The recommended weight limit defines the load an average worker can safely handle over a certain period [26]. Like self-report methods, observational approaches also face challenges that can lead to inconsistencies, be time-consuming, and be prone to errors [13,20].

Direct measurement methods rely on devices to monitor workers' postures, movements, level of surface activation, and forces exerted. These devices include, but are not limited to, lumbar motion monitors, optical and non-optical motion capture systems, electronic goniometers, and heart rate monitors, among others [23,27–30]. However, the main drawbacks of optical systems are their dependence on high-quality ambient lighting and the potential for errors or loss of motion capture data when obstructions are present between markers and cameras [13]. Sensor-based direct measurement methods are intrusive, while marker-based methods impose strict environmental requirements on workers. Some researchers use simulation or modeling approaches [31,32] to assess ergonomic risk, but these require inputting worker data, making evaluations vulnerable to observer influence.

Traditional methods of postural ERA, such as self-reporting, observation, direct measurements, and simulation, are often constrained by their reliance on manual data collection [4]. This process is time-consuming and often leads to inconsistent data interpretation among ergonomists.

2.2. Traditional ML-Based ERA

Traditional Machine Learning (ML)-based ERA uses various AI algorithms to learn from past data and analyze future data based on the knowledge acquired to perform risk assessments. Traditional ML-based ERA involves three steps: collecting relevant data to train the AI, training the AI to improve accuracy, and deploying the model to predict outcomes based on learned patterns. Previous studies on Traditional ML-based ERA have commonly used AI techniques, including Support Vector Machines, Decision Trees, Long Short-Term Memory, Linear Regression, Artificial Neural Networks, Recurrent Neural Networks, and Convolutional Neural Networks, among others [12,33]. Significant potential exists for deploying AI-based ERA methods in construction, though more research has been done on AI in ERA generally. These general methods can also be adapted for construction with modifications. This study only discusses construction industry-specific methods.

The majority of traditional ML-based ERA methods use sensors attached to the human body. These methods all involve intrusiveness and inevitably increase workers' mental and physical burden. While both approaches depend on equipment data collection, the data gathered for traditional ML is used to train the model, which can then be reused to automate the ERA process. This is fundamentally different from direct measurement methods. In contrast, direct measurement methods rely on human evaluation for ERA. Approaches

that use only video as input to AI algorithms offer an alternative to sensor-based methods, eliminating the need to attach sensors or markers to workers. For instance, Fan et al. [13] developed a CNN-based CV algorithm for the real-time tracking of workers' movements. This algorithm is trained on a previously collected dataset of 3D models of construction workers. Following this, the REBA method is applied to evaluate ergonomic risks from the predicted pose data obtained through the CV algorithm. The technique developed by Cai et al. [34] leverages a CV-based pose-tracking algorithm to track workers' movements. The next step is to use an LSTM model to classify ergonomic poses, culminating in assessing their ergonomic risks using the OWAS method. This methodology also demonstrates a non-intrusive way of conducting ergonomic assessments.

2.3. Potential Applications of Generative AI in ERA

Generative AI can create new content, such as text and images, while traditional ML focuses on categorization or prediction based on input data. Modern multi-modal LLMs blur the boundary between traditional ML and generative AI, but there is a fundamental functional difference between these two approaches [35,36]. Generative AI's ability to create text allows innovative applications in ERA. It can respond with yes/no or detailed descriptions about ergonomic risks based on prompts. However, traditional ML-driven ERA, focusing on risk levels through variable detection, is limited in identifying root issues and suggesting corrections [37].

The application of generative AI to address ergonomic risks in the construction sector is underexplored, in contrast to its emerging use for safety risks. Searching 'vision language' AND 'ergonomics' on Scopus yields only two search results, both authored by the same group of authors. Both of these articles propose a VLM capable of identifying ergonomic problems and their solutions [37,38]. The difference between them lies in the use of different VLM architectures.

In contrast to construction ergonomics, several generative AI methodologies have been proposed to address construction safety issues unrelated to ergonomics. Smetana et al. [39] developed a method for text-based incident analysis utilizing the GPT-3.5 model from OpenAI. They applied GPT-3.5 and various natural language processing strategies to the OSHA severe injury reports database to pinpoint the leading causes of highway construction accidents. In an independent study [40], Uddin et al. found that ChatGPT significantly improved architecture students' ability to identify dangers. Yoo et al. [41] developed a model based on transfer learning and fine-tuned GPT to detect six types of accidents; they also proposed a technique to enhance the interpretability of the model by visualizing significant portions of text input. Although these methods utilize LLMs for text analysis and generation, they cannot generate text responses from images or videos. In contrast to these generative AI methods that cannot process image inputs, Chen et al. [42] have proposed a visual construction safety query system that integrates with head-mounted virtual reality devices. This system uses a VLM to generate image captions, facilitate safety-relevant VQA, and perform image–text retrieval.

Notably, existing methods [37,38] employing generative AI to tackle ergonomics-related challenges exclusively rely on Bootstrapped Language-Image Pretraining (BLIP) [43] as the foundational VLM. However, MiniGPT-v2 outperforms BLIP in VQA [44]. Using ChatGPT for this purpose presents multiple challenges. It not only lacks specific domain knowledge but also heavily relies on the quality of the data input. Furthermore, it is unable to recognize errors, lacks true creativity, and has a very limited understanding of ethical issues [45]. Additionally, both studies employed the Bilingual Evaluation Understudy (BLEU) [46] metric to evaluate the performance. Therefore, this study proposes an interactive method that supports explainable image captioning (IC) for ergonomic-risk description and VQA for ergonomic-risk recognition.

IC aims to generate syntactically correct, semantically coherent descriptions of an image's visual content [47]. Traditionally, this task was viewed as a process of converting images into text sequences. In the visual encoding phase, the input pixels are converted into feature vectors, which are then fed to the language model to generate descriptions. In recent years, the design and performance of IC models have significantly improved with the development of transformer architecture, self-attention mechanisms, and the introduction of methods such as single-stream BERT [47].

3. Materials and Methods

This study employs a controlled comparative experimental methodology to investigate whether construction-ergonomics-specific domain adaptation improves the ergonomic reasoning capability of a generic VLM. The same pretrained MiniGPT-v2 is used for both experimental and comparison models to control model architecture. The research procedure consists of five stages: (1) establishing the generic pretrained VLM as the comparison model; (2) curating and annotating a construction-specific ergonomic image–text dataset; (3) fine-tuning the same VLM using the domain-specific training partition; (4) evaluating both models on the identical held-out testing partition using classification performance for binary VQA and nine caption-quality metrics; and (5) performing a blinded human evaluation of the generated ergonomic descriptions. This experimental design isolates the effect of construction-ergonomics-specific fine-tuning while maintaining the same underlying VLM architecture. Figure 1 first presents the overall experimental methodology, including generic model selection, construction-ergonomics dataset development, domain-specific fine-tuning, and model evaluation. Figure 2 then provides a detailed view of the ErgoChat architecture and illustrates how REBA-informed ergonomic knowledge developed during dataset annotation is incorporated through multimodal instruction tuning.

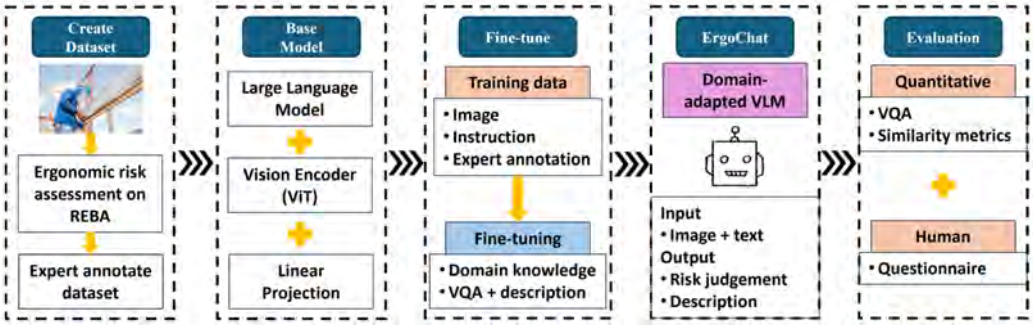

Figure 1. Overall research and evaluation workflow of the study.

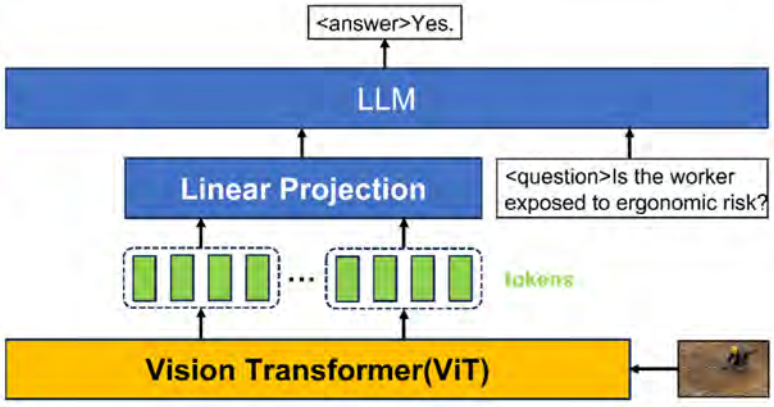

Figure 2. Architecture and domain-adaptation mechanism of ErgoChat. The LLaMA2-7B is used as the LLM backbone, and visual tokens from the ViT are projected into the LLM feature space via the linear projection layer.

The proposed ErgoChat is an interactive chatbot capable of generating textual descriptions of ergonomic risks and answering ergonomic risk–related visual questions. The proposed ErgoChat can achieve VQA and IC regarding ergonomic risks for construction scenarios. Its architecture is adopted from MiniGPT-v2, comprising a visual backbone, a linear projection layer, and an LLM. The parts can be summarized as follows.

1. The visual backbone of ErgoChat utilizes the EVA ViT [48]. A ViT employs a self-attention mechanism for image processing;
2. The role of the linear projection layer in ErgoChat is to project visual tokens [48] from the frozen visual encoder into the feature space of the LLM [49];
3. ErgoChat adopts the open-source LLaMA2-7B [49] as the backbone for its LLM. The base model LLaMA2-7B has approximately 6.74 billion trainable parameters, and the fine-tuned model has approximately 0.5% more trainable parameters.

The underlying VLM architecture of ErgoChat is adopted from MiniGPT-v2 and does not contain a separate rule-based REBA module. Domain-specific ergonomic knowledge is introduced during dataset construction and fine-tuning. Specifically, REBA principles are used by ergonomics experts to determine postural risk labels and construct natural-language descriptions of risk-related postures. These image–text and instruction pairs are subsequently used to fine-tune the VLM, thereby incorporating construction-ergonomics-specific information into the model parameters. Based on this architecture, domain-specific fine-tuning was conducted to enhance ErgoChat's construction-ergonomics-specific capabilities. This study utilized publicly available image–text pair datasets, although not specifically tailored for ergonomic risk, for the initial training of ErgoChat. The datasets used for pre-training and fine-tuning are listed in Table 1. To improve ErgoChat's ability to recognize ergonomic risks, this study specifically built a new dataset for fine-tuning VLMs. After fine-tuning was completed, a quantitative assessment of ErgoChat's performance was conducted using multiple metrics on the test set proposed in this study. Additionally, questionnaires were distributed to human evaluators to qualitatively and quantitatively evaluate the IC's performance on the testing dataset. Table 2 provides the training hyperparameters and environment for the models.

Table 1. The pre-training and fine-tuning stages and datasets in each stage. Checkmarks indicate that the dataset was used at the corresponding stage, whereas crosses indicate that it was not used.

Dataset	Dataset Type	Pre-Train 1	Pre-Train 2	Pre-Train 3	Fine-Tune
GRIT-20M [50], LAION [51], CC3M [52], SBU [53]	Weakly labeled	✓	✘	✘	✘
GRIT-20M [50]	Grounded caption	✓	✘	✘	✘
COCO caption [54], Text Captions [55]	Caption	✓	✓	✓	✘
RefCOCO [56], RefCOCO+ [57], RefCOCOg [58], Visual Genome [59]	REC	✓	✓	✓	✘
RefCOCO [56], RefCOCO+ [57], RefCOCOg [58]	REG	✓	✓	✓	✘
GQA [60], VQAv2 [61], OCR-VQA [62], OK-VQA [63], AOK-VQA [64]	VQA	✓	✓	✓	✘
LLaVA [65], Flickr30k [66], Multi-task conversation [44]	Multimodal instruction	✘	✘	✓	✘
Unnatural Instructions [67]	Language dataset	✘	✘	✓	✘
Our fine-tuning dataset	Caption, VQA	✘	✘	✘	✓

Table 2. The training hyperparameters and environment of the models.

Hyperparameters/Environment	Value
GPU	2 NVIDIA GA102GL RTX A6000
CPU	Intel® Xeon(R) Gold 6242R
Operating system	Ubuntu 22.04 64-bit
GPU driver	NVIDIA 555.42 with CUDA 12.5
Programming language	Python 3.9
Optimizer	AdamW
Initial learning rate	10^{-5}
Minimum learning rate	10^{-6}
Warmup learning rate (1000 steps)	10^{-6}
Learning-rate scheduler	Linear warmup + cosine decay
Weight decay	0.05
Epochs	50
Random seed	42
Input image resolution	448 × 448 pixels
Precision	Automatic mixed precision (AMP)
LLM adaptation	LoRA
LoRA rank	64
LoRA scaling factor	16
LoRA dropout	0.05
Beam size	1
Top-p	0.9
Repetition penalty	1.05
Maximum context length	2000

User interaction with ErgoChat follows a standard browser-based client-server workflow. A user submits a workplace image together with a textual ergonomic-risk query through the browser interface. The image and query are transmitted through HTTP to a Python server, which invokes the fine-tuned VLM on a local GPU. During inference, the pipeline loads the fine-tuned model, processes the input, and generates a text response. The generated response is first returned to the Python server, which forwards it to the browser, and is finally presented to the user in the browser. This study followed five steps to develop and evaluate ErgoChat for ergonomic risk identification and description. Ethical approval has been obtained from the University of Alberta (ethics approval: Pro00111404). All experiments followed the research protocol approved by the Declaration of Helsinki. Oral informed consent was obtained from all participants, and completing the questionnaire was considered consent.

3.1. Pre-Training a Model with Generic Image–Text Pair Datasets

The base model was trained with publicly available generic image–text pair datasets not explicitly designed for ergonomic risk identification/knowledge, resulting in a large VLM with generic visual knowledge. The training using the generic dataset was divided into three pre-training stages (pre-train 1–3); the model underwent three pre-training sessions using different combinations of those datasets. These datasets encompassed a broad range of content, including humans, other animals, plants, architectural structures, vehicles, and landscapes. Pre-train 1 involves equipping the base model with broad, generic vision–language knowledge by leveraging a combination of numerous weakly labeled image–text datasets and high-quality, fine-grained vision–language annotation datasets. Pre-train 2 involves refining the model using only fine-grained data for various tasks, such as VQA and IC. The rationale is to improve the model's performance in tasks such as VQA and IC. Therefore, the previously used weakly supervised dataset was not used again in this training step. In phase 3 of pre-training, the model is further refined

using additional multimodal instructions and datasets. This process enhances its ability to understand and respond to various multimodal instructions, enabling it to function more effectively as a multimodal chatbot. The fine-tuning enables the model to gain more robust reasoning abilities while also addressing the issue where, after training on the GRIT-20M dataset (which contains very few grounded visual objects in its captions), the model could only visually recognize a limited range of objects [44]. In this study, ergonomic risk reasoning refers to the model's ability to associate visual posture information with an ergonomic-risk judgment and to generate a natural-language explanation of the posture or action contributing to that risk. The fine-tuning annotations therefore include not only risk labels but also descriptions of the visual factors underlying the identified ergonomic risk. MiniGPT-v2 provides a base model that has already undergone the three pre-training stages. The base model is further fine-tuned on ergonomics-specific text-image data to improve its ability to recognize and describe work-related ergonomic risks. The resulting multimodal vision–language model serves as the core model within ErgoChat. Table 1 shows the datasets used for pre-training and fine-tuning. Fine-tuning utilizes the dataset proposed in this study and is discussed in the following sections.

3.2. Dataset Curation for Fine-Tuning and Testing

An image–text pair dataset was curated to fine-tune and test the VLM obtained in the previous step. When creating the dataset, the fine-tuning and testing partitions were screened for duplicate content. Images identified as already existing in the dataset were discarded. The fine-tuning partition contained 1700 distinct image–text pairs, and the performance-testing partition contained 200. The images in both fine-tuning and testing partitions were obtained by searching for 'construction works' in online resources with free-to-use licenses. In the testing partition, 5% of the images depicted workers who were not exposed to ergonomic risks, whereas the remaining 95% depicted workers who were. In the training partition, 90.59% of the images showed that workers were at ergonomic risk, and only 9.41% showed no risk. To better reflect the real construction environment, the dataset also included situations in which some workers were obscured, thereby better representing the complexity of the construction site.

All collected images were manually inspected before dataset partitioning. Exact duplicates, cropped or otherwise modified versions of the same image, and images belonging to the same image sequence or depicting essentially the same moment were identified and excluded from being distributed across the fine-tuning and testing partitions. In rare cases, images could originate from the same general construction scenario; however, these represented distinct observations rather than adjacent or near-duplicate frames. Systematic grouping based on individual workers, construction sites, metadata, or original online sources was not performed.

Descriptions of ergonomic risks in the text part of image–text pairs were based on REBA scoring. The REBA employs a systematic approach to assess the upper and lower segments of the musculoskeletal system and identify risks of WMSDs associated with a work task. The REBA classifies risk levels into five categories, ranging from negligible to very high [24]. Ergonomic risks are identified based on whether any worker shown in the images has a REBA risk level that is medium or higher. If the REBA risk level meets or exceeds the medium risk threshold, the textual description will state that the depicted worker is exposed to ergonomic risk; conversely, if the REBA risk level falls below the medium risk threshold, the textual description will indicate no exposure to ergonomic risk. The threshold was set to prioritize interventions for medium or higher REBA risks, as including low-risk cases leads to more false positives and additional downstream workload for users. Also, ergonomic practices usually focus on addressing medium-to-high-risk

postures, so in this study, the classification threshold was set at medium risk, corresponding to a REBA score of 4 or higher. This aligns with the REBA guidelines, which indicate that posture scores of 4–7 require further review and timely correction. An analysis of threshold sensitivity is not within the scope of the study, but future work will explore the effects of different threshold choices on VQA system performance. Alongside determining whether the workers depicted in the images are exposed to ergonomic risk, the text also describes the actions and tasks that put them at risk. A Python program was developed to convert these textual descriptions into a JavaScript Object Notation (JSON) file. The construction worker images and the JSON-formatted annotation file constitute this image–text pair dataset, which was used for fine-tuning and testing of ErgoChat.

Data annotation was performed independently by an ergonomic expert with more than 5 years of professional experience. REBA scoring, with clear numeric guidelines, reduces variation from personal judgment. To improve annotation consistency and reduce bias, a two-stage review was conducted with another ergonomic expert reviewing 500 samples of the annotated dataset. Agreement between the two annotators was evaluated based on the identified ergonomic risk-factor categories, including Forceful Exertions, Awkward Postures, Repetitive Motions, Static/Stationary Postures, Vibration, and Contact Stress. Cohen's kappa was 0.83, indicating strong agreement and reliability.

3.3. Fine-Tuning ErgoChat with the Specific Dataset for Ergonomic Risk Identification

The large VLM obtained in step 1 was fine-tuned using the fine-tuning dataset created in step 2. Using the ergonomic-specific dataset for construction workers proposed in this study, fine-tuning equips ErgoChat with knowledge of the ergonomic risks commonly faced by construction workers, thereby enhancing its performance in VQA and IC for ergonomic risks in the construction industry. This fine-tuning process enhanced the VLM's domain-specific capability to identify and describe construction-related ergonomic risks. In this step, ErgoChat was pre-trained through pre-training 1 to 3 and underwent fine-tuning with the ergonomic risk-specific dataset. The dataset utilized in fine-tuning was precisely the fine-tuning partition created in step 2. In addition to the datasets used for pre-training 1–3, Table 1 shows the dataset used for fine-tuning. The fine-tuned ErgoChat was subsequently evaluated to determine whether domain-specific training improved its performance on ergonomic-risk-related VQA and IC tasks. Figure 3 shows an example of the fine-tuning dataset, and the corresponding ground-truth text by the ergonomic expert states the following: 'The image shows a worker in a safety vest standing on a brown wooden pallet and cutting with a power tool. The worker is bending forward while working. The worker is exposed to an awkward working posture, and it is an ergonomic risk. The awkward posture may lead to the development of work-related musculoskeletal disorders.'



Figure 3. An image from the fine-tuning part of the dataset [68]. A man in a green and yellow safety vest and blue denim jeans standing on a brown wooden pallet.

3.4. Performance Evaluation of the Fine-Tuned ErgoChat

The performance was measured from two perspectives: (1) accuracy in VQA related to ergonomic risk identification and (2) the ability to generate textual descriptions concerning ergonomic risk identification for workers.

3.4.1. Accuracy of VQA

The large VLM obtained after fine-tuning was tested with the testing dataset curated in step 2. The model obtained without fine-tuning (after pre-training 1 to 3) was also tested on the same test partition, allowing the performance of ErgoChat before and after fine-tuning to be compared. To alleviate ambiguity across various queries, the proposed dataset includes task-identifier tokens for IC and VQA. Task identifiers were added at the beginning of prompts to reduce ambiguity. Among them, [caption] was used for IC tasks, while [vqa] was used for VQA tasks. For VQA, researchers compared the results before and after fine-tuning against actual values. For example, the VQA prompt for the test image in Figure 4 was '[vqa] Is the worker exposed to postural ergonomic risks?' and both before and after fine-tuning, ErgoChat's responses were 'yes.' The same prompt '[vqa] Is the worker exposed to postural ergonomic risks?' was used for all the cases in the testing partition. Since the ground-truth value for the VQA in Figure 4 was 'yes,' the responses from ErgoChat both before and after fine-tuning were correct. Finally, the accuracy of VQA that the two models can achieve for the 200 test data points (percentage of correct answers out of the total) was calculated. Equation (1) was used to calculate the percentage of correctly identified risks using VQA.

$$\text{percentage of correct identication} = \frac{\text{number of correct identification}}{200} \times 100\% \quad (1)$$

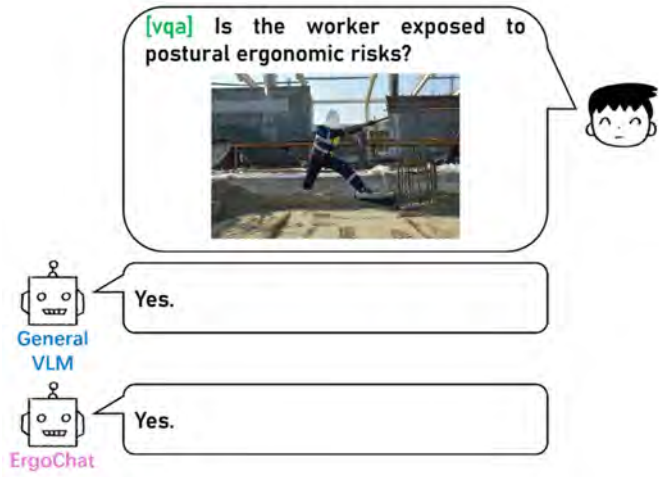

Figure 4. The VQA question and answers from 2 models in an example. The 2 VLMs both answered 'Yes' to the question.

3.4.2. Quality of Generated Text Descriptions

The textual descriptions of ergonomic risks generated by ErgoChat were compared with the ground truth before and after fine-tuning. The prompt used to create text descriptions was '[caption] Describe the workers and their postures in the image and tell me if they are exposed to ergonomic risks due to their postures?' The same prompt was used for all the cases in the testing partition. ErgoChat's performance was measured by calculating the similarity between generated and ground-truth descriptions. The perplexity [69] was used for VQA. A total of 9 similarity metrics were used to calculate the similarity between the generated and ground-truth texts for each image in the test dataset. The similarity metrics

were ROUGE_r, ROUGE_p, ROUGE_f [70], BLEU [46], NIST [71], cosine similarity [72], Euclidean distance [72], METEOR [73], and SPICE [74]. Except for perplexity, NIST, cosine similarity, and Euclidean distance, metric values range from 0 to 1. All these metrics, except perplexity and Euclidean distance, indicate better performance with higher scores. Table 3 provides more detailed information about these metrics. The 9 similarity metrics were used to evaluate IC performance. Perplexity was used as an auxiliary measure of relative model preference between two candidate risk responses rather than as a probability of ergonomic-risk correctness. For each test case, separate perplexity scores were calculated for the candidate labels 'Yes' and 'No.' Because lower perplexity indicates that a sequence is more compatible with the model, the candidate with the lower perplexity was treated as the model's perplexity-based label preference. This preferred label was subsequently compared with the ground-truth ergonomic-risk label. The resulting measure is therefore referred to as perplexity-based label-preference agreement and should not be interpreted as a calibrated probability of ergonomic risk or as direct classification accuracy.

Table 3. The similarity metrics used in this study to evaluate performance.

Metric	Range	Description
Perplexity [69]	1 to ∞	Assessment of how effectively the model has learned the distribution of the text on which it was trained.
ROUGE_r [70]	0 to 1	Measurement of the completeness of the information captured by a generated caption.
ROUGE_p [70]	0 to 1	Indicator of the accuracy of the system-generated caption.
ROUGE_f [70]	0 to 1	A combination of ROUGE_r and ROUGE_p.
BLEU [46]	0 to 1	Measurement of how closely the machine-generated text resembles a set of high-quality reference texts.
NIST [71]	0 to ∞	Evaluation of machine translation that uses n-gram precision to calculate the similarity between a machine translation and a reference translation.
Cosine similarity [72]	−1 to 1	A measurement of how similar the documents are, regardless of their size.
Euclidean distance [73]	0 to ∞	A measurement of the similarity between vectors, such as word embeddings or document vectors.
METEOR [74]	0 to 1	A measurement of the alignment between the generated text and the reference text, which assesses the quality of the generated text.
SPICE [75]	0 to 1	Evaluation of IC quality by analyzing the semantic similarity between reference and generated captions.

The perplexity was used to test the probability that the generated textual descriptions for ergonomic risk identification were correct. If the generated description indicates that the worker is performing an ergonomically unsafe action, the corresponding score for identifying the ergonomic risk will be lower. A lower perplexity score indicates a higher probability that the description concludes the worker is exposed to ergonomic risk. A more detailed description of perplexity can be found in [69]. A text description with a high perplexity score indicating ergonomic risk and matching ground truth suggests the VLM provides a more accurate description. ROUGE, meant for evaluating summaries, can also assess machine translation. The similarity metrics compare a generated summary or translation against high-quality, human-produced reference summaries or translations [69]. ROUGE_r was computed by dividing shared unigrams between the expert and ErgoChat descriptions by the total unigrams in the expert description. ROUGE_p can be computed by taking the ratio of the number of unigrams in the ErgoChat-generated description that also appear in the ergonomics expert-generated description to the total number of unigrams in

the ErgoChat-generated description. The ROUGE_f was calculated using ROUGE_r and ROUGE_p with Equation (2):

$$\text{ROUGE}_f = \frac{2 \times (\text{ROUGE_r} \times \text{ROUGE_p})}{\text{ROUGE_r} + \text{ROUGE_p}} \quad (2)$$

BLEU is a metric for comparing a generated translation to one or more reference translations. While initially developed for translation, it is also applicable for evaluating other tasks, such as paraphrasing and text summarization [46]. For further details on the other similarity metrics used in this study, readers can refer to [46,71–75].

Due to disparities in the amount of ergonomic risk knowledge, the similarity between the textual descriptions produced by ErgoChat and the ground truth descriptions was expected to vary before and after fine-tuning. The nine previously mentioned similarity metrics for IC were employed to evaluate the similarity between the generated textual descriptions and the ground truth descriptions for both general VLM (pre-trained 1 to 3) and fine-tuned (ErgoChat) models. Each of the 200 descriptions generated by the fine-tuned ErgoChat underwent evaluation for the nine metrics, yielding a total of 1800 computations. Similarly, the 200 descriptions generated by the model before fine-tuning underwent the same evaluation. For each metric, the similarity scores of the general VLM were subtracted from those of ErgoChat for each image, and the average of the differences for all 200 data points was determined by using Equation (3):

$$\text{average of difference} = \frac{\sum_{i=0}^{200}(f_{m,i} - u_{m,i})}{200} \quad (3)$$

where $f_{m,i}$ denotes the fine-tuned metric score for the metric m and image i. Similarly, $u_{m,i}$ represents the metric score of the model without fine-tuning for the metric m and image i.

Subsequently, the average improvement in similarity across the nine similarity metrics for the fine-tuned ErgoChat was computed. For each metric, the percentage of testing data that exhibited improvement with the fine-tuned ErgoChat was also determined using Equation (4):

$$\%\text{ of data that indicated improvement} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[s_m(f_{m,i} - u_{m,i}) > 0] \times 100\% \quad (4)$$

where N = 200 is the number of test images, $f_{m,i}$ denotes the score of the fine-tuned model for metric m and image i, $u_{m,i}$ denotes the corresponding score of the general VLM, $\mathbb{I}$ is the indicator function, and s_m accounts for the direction of improvement. Specifically, $s_m = 1$ for metrics for which a larger value represents better performance and $s_m = -1$ for metrics, such as Euclidean distance, for which a smaller value represents better performance.

Additionally, for the applicable metrics, the absolute difference between the fine-tuned and baseline scores was normalized to the metric scale and expressed as a percentage. This quantity is referred to as the normalized absolute score difference and was determined using Equation (5):

$$\text{average improvement percentage} = \frac{\sum_{i=0}^{200}(f_{m,i} - u_{m,i})}{200} \times \frac{1}{m_{max}} * 100\% \quad (5)$$

where $f_{m,i}$ denotes the fine-tuned metric score for the metric m and image i. Similarly, $u_{m,i}$ denotes the metric score of the model without fine-tuning for the metric m and image i. m_{max} represents the maximum value/upper bound of the metric m.

The similarity metrics used in this study quantify different aspects of similarity between model-generated descriptions and expert-authored reference descriptions; however,

they should not be interpreted as direct measures of ergonomic correctness. Ergonomically equivalent statements may use different vocabulary or sentence structures and therefore receive relatively low lexical-similarity scores, while high textual similarity does not necessarily ensure that all ergonomically important information is correct. Accordingly, the nine metrics are treated as indicators of reference-caption similarity rather than definitive measures of ergonomic accuracy. The blinded human evaluation was included to provide a domain-informed assessment of whether the generated descriptions accurately represented the ergonomic risks shown in the images. Because the similarity metrics and human evaluation assess different constructs, their results are used as complementary evidence rather than as interchangeable measures.

3.5. Human Evaluation of Generated Textual Descriptions

In addition to the nine similarity metrics, human evaluation was used to assess the improvement in fine-tuning on ErgoChat for text description generation related to ergonomic risk identification. A questionnaire was created to gather feedback from ergonomics evaluators on the accuracy of descriptions from fine-tuned ErgoChat compared to the original model before fine-tuning on the test dataset. A total of 50 participants were recruited to complete the questionnaire. Participants were recruited through two channels. First, the researchers' university distributed recruitment emails to former students who had completed graduate-level ergonomics courses. Second, the researchers circulated recruitment information through their professional and academic networks, including university professors and professionals working in industry. To be eligible for participation, individuals were required to have either professional work experience related to ergonomics or prior completion of graduate-level ergonomics coursework. To characterize the evaluation sample, participants reported their education level, employment sector, level of ergonomics knowledge, and age. The survey was conducted anonymously by not asking for personal information. The answers were collected through an online survey platform.

54% of the questionnaire participants are Ph.D. degree holders or students; 40% are master's degree holders or students; 6% are undergraduate students or degree holders. Among the human evaluators, 56% work in manufacturing, construction, or agriculture-related sectors, and 34% work in the public sector or education. 10% of the evaluators consider themselves experts in ergonomics; 16% consider themselves advanced; 28% consider themselves intermediate; 24% consider themselves novices; 22% consider themselves to have fundamental awareness; None of them consider themselves to have no knowledge of ergonomics. 18% of the evaluators are 26 years old; 24% are 24, 25, or 29; 18% are 28, 32, or 34; the other ages do not comprise a significant portion of the evaluators. The ages of the evaluators range from 23 to 54.

There were 13 versions of the questionnaire, with 8 containing 15 images each and the remaining 5 containing 16 images each. Together, these 13 versions covered 200 images in the test dataset. Each image was accompanied by two survey questions. 50 participants provided 770 pairwise caption evaluations. Thirty participants evaluated 15 images each, while 20 participants evaluated 16 images each. Across the 200 test images, each image received a mean of 3.85 evaluations per image. Because observations were clustered within both participants and images, the fine-tuned-model preference rate was analyzed using an intercept-only logistic regression with two-way cluster-robust standard errors for participant and image. The null hypothesis corresponded to an equal probability (50%) of selecting either model. The evaluators were blinded to the source of the generated captions, and the presentation order of the two captions was randomized. For each evaluated image, participants first selected the description that more accurately represented the ergonomic risks shown in the image and then indicated the perceived magnitude of the difference

using a six-level scale of 0%, 20%, 40%, 60%, 80%, or 100%. Because the 13 questionnaire versions contained different subsets of the 200 test images, not every participant evaluated every test image. Therefore, the reported choice rates and relative-accuracy ratings are interpreted primarily as descriptive measures of participant preference rather than as fully paired participant-level inferential statistics. Future studies will use a more balanced evaluation design or statistical models that explicitly account for participant- and image-level variability.

In the first question, the participants were asked to select the more accurate description of ergonomic risks faced by the construction worker depicted in the image from two provided text descriptions. The second question asked participants to rate how much more accurate their choice was than the other one. These two text descriptions were generated by the model without fine-tuning and the fine-tuned ErgoChat, respectively, and were identical to those in Section 3.4. The first question was formulated: 'Which description is more accurate to describe the ergonomic risks in the image?' The choices for the first question were randomized for all the testing cases. Figure 5 shows a question from the questionnaire.

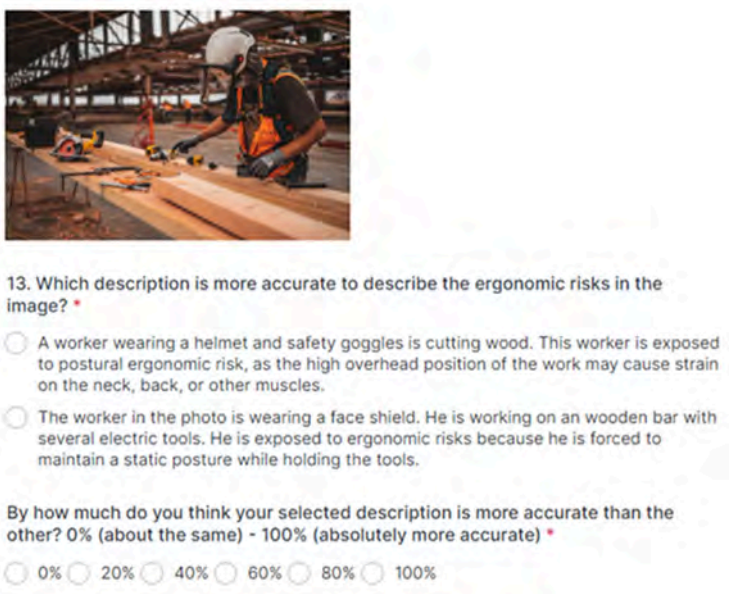

Figure 5. Example of the human-evaluation questionnaire. The 13 questionnaire versions collectively covered the 200 test images, and each image was evaluated using the same two questions in this figure regarding relative caption accuracy and the perceived magnitude of the difference. Red asterisk means the question is required. All questions in the questionaire are required.

This study analyzes responses to the first question to gather human evaluations of ErgoChat's accuracy in describing ergonomic risks. It compares human assessments with metric evaluations from Section 3.4 to assess consistency. For example, suppose the majority of human evaluators' responses to the first question in the questionnaire suggest that descriptions generated by the fine-tuned ErgoChat are more accurate, and simultaneously, results from the nine similarity metrics indicate that these descriptions closely align with the ground truth. In that case, this provides additional evidence that the fine-tuned ErgoChat generated descriptions that were preferred more frequently by human evaluators. Together, these results provide evidence supporting the feasibility of the proposed methodology for identifying and describing ergonomic risks. The process of interpreting human evaluation results is shown in Figure 6.

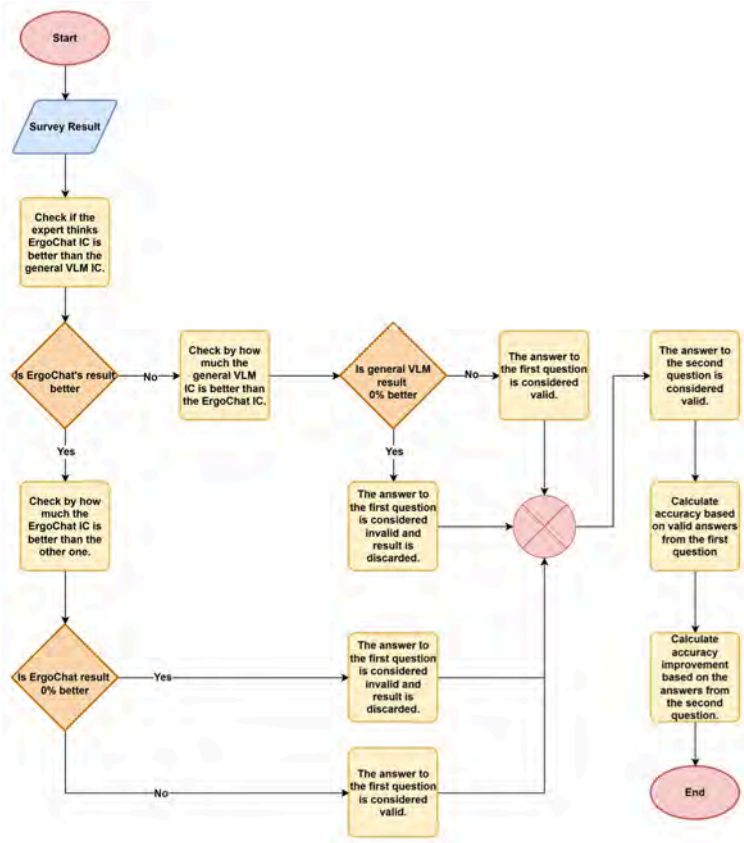

Figure 6. The process of handling human evaluation data.

The fine-tuned choice rate represents the proportion of valid pairwise comparisons in which participants selected the description generated by fine-tuned ErgoChat as the more accurate description. The second questionnaire item measured the perceived magnitude of the difference between the selected and non-selected descriptions using a six-level scale ranging from 0% to 100%. The phrasing of this question was, 'By how much do you think your selected description is more accurate than the other? 0% (about the same)–100% (absolutely more accurate).' The resulting 69.7% value therefore represents a mean participant-rated relative accuracy difference; this value should be interpreted as a subjective comparative judgment from the evaluators. Concurrently, based on which ErgoChat model the most responses belong to, this study calculated the percentage by which one model is more accurate.

4. Results

4.1. Performance of VQA

The testing dataset was utilized to assess the performance of the proposed ErgoChat in ergonomic risk identification in VQA. To assess the performance of ErgoChat and the general VLM in ergonomic risk identification, this study calculated accuracy using perplexity scores and VQA results from the descriptions. The results were obtained with the testing partition containing 200 data samples. Table 4 presents the VQA classification performance of ErgoChat and the general VLM on the test dataset. The results were obtained using Equation (1). Since ErgoChat's perplexity score indicates the probability of ergonomic risk, the perplexity scores from ErgoChat and the general VLM were calculated and compared to the ground truth of VQA. The perplexity-based label preference agreed with the ground-truth risk label for 172 of 200 test cases (86.0%) for ErgoChat and 127 of 200 cases (63.5%) for the general VLM. These results indicate that the fine-tuned model's relative preference

between the two candidate responses was more frequently consistent with the ground-truth label. Because perplexity is not a calibrated probability of ergonomic-risk correctness, these values are reported only as an auxiliary agreement measure. Direct VQA classification metrics are used as the primary evaluation of ergonomic-risk identification. ErgoChat achieved an accuracy of 96.5% in responding to the prompt '[vqa] Is the worker exposed to postural ergonomic risks?' using only images from the test dataset. In comparison, the other VLM achieved 95% accuracy on the same task. Medium-to-high-risk scenarios play a central role in preventing WMSDs, but are also the most susceptible to recognition bias in generic VLMs. ErgoChat achieved a classification performance for binary VQA of 96.5% (193/200; 95% Wilson CI: 92.95–98.29%), compared with 95.0% (190/200; 95% Wilson CI: 91.04–97.26%) for the general VLM. The class imbalance in the testing dataset should be considered when interpreting classification performance for binary VQA. Because both models were evaluated on the same images, an exact McNemar's test was conducted on the paired binary outcomes. The fine-tuned model was correct while the general VLM was incorrect for five cases, whereas the reverse occurred for two cases. The difference was not statistically significant (exact McNemar's test, $p = 0.453$). Of the 200 test images, 190 were risk-positive, and 10 were risk-negative; therefore, a classifier predicting 'risk' for every image would achieve 95% accuracy. To provide a more informative evaluation, the confusion matrix and class-sensitive performance measures were calculated. ErgoChat produced 189 true positives, 4 true negatives, 6 false positives, and 1 false negative, corresponding to 99.47% sensitivity, 40.00% specificity, 96.92% precision, a 98.18% F1-score, 69.74% balanced accuracy, and a Matthews correlation coefficient (MCC) of 0.551. The general VLM produced 188 true positives, 2 true negatives, 8 false positives, and 2 false negatives, corresponding to 98.95% sensitivity, 20.00% specificity, 95.92% precision, a 97.41% F1-score, 59.47% balanced accuracy, and an MCC of 0.295. These results indicate that the fine-tuned model performed better than the general VLM on the present test set, including improved identification of the minority no-risk class. Nevertheless, the relatively low specificity of both models, particularly given the small number of no-risk samples, demonstrates that the high overall accuracy could be influenced by class imbalance and should therefore be interpreted cautiously. Accordingly, the VQA results indicate a slight improvement.

Table 4. VQA classification performance of ErgoChat and the general VLM on the test dataset. Note: The reported model results were obtained from a single fine-tuning and evaluation run; therefore, between-run standard deviations are not available.

Metric	Correct Identification	
	ErgoChat	General VLM
True positive	189	188
True negative	4	2
False positive	6	8
False negative	1	2
Sensitivity/Recall	99.47%	98.95%
Specificity	40.00%	20.00%
Precision	96.92%	95.92%
F1-score	98.18%	97.41%
Balanced accuracy	69.74%	59.47%
MCC	0.551	0.295
Classification performance for binary VQA	96.5%	95%

Overall, ErgoChat achieved slightly higher classification performance for binary VQA than the corresponding general VLM. It is worth noting that although the VLM used for comparison did not undergo fine-tuning, it demonstrated strong performance in VQA

for ergonomic risk identification. These findings suggest that the method, pre-trained on generic datasets, can detect ergonomic risks in images. Nevertheless, its capacity to generate natural-language descriptions of ergonomic risks and elucidate actions leading to such risks remains more limited in the evaluated caption-generation task.

4.2. Performance of the IC of Ergonomic Risk-Related Descriptions

To evaluate the performance of fine-tuned ErgoChat and the general VLM in generating textual descriptions of ergonomic risks from images, this study used multiple similarity metrics to measure the similarity between the generated text and the ground truth descriptions in the dataset. Table 5 shows the difference (improvement) in the average scores obtained by ErgoChat compared to the general VLM on the testing partition using 9 similarity metrics and the improvement represented by these differences. These two indicator types can be obtained using Equations (3) and (5). Additionally, Table 5 also includes the percentage of data on the entire testing partition for which improvement can be achieved through fine-tuning ErgoChat, and this can be obtained by using Equation (4). N/A indicates that the normalized absolute score difference was not calculated for NIST or Euclidean distance because these metrics do not have a fixed finite upper bound required for normalization. In addition, a lower Euclidean distance indicates greater similarity. Table 6 shows the average metric scores for fine-tuned ErgoChat and the general VLM. It is worth mentioning that among these similarity metrics, only the Euclidean distance indicates lower similarity as the score increases.

Table 5. Similarity metric results of IC for the general VLM and ErgoChat. Note: The reported model results were obtained from a single fine-tuning and evaluation run; therefore, between-run standard deviations are not available.

Metric	Average of Difference (Equation (3))	Percentage of Data with Improvement (Equation (4))	Average Improvement (Equation (5))
ROUGE_r	0.26	97.0%	25.53%
ROUGE_p	0.07	70.5%	7.45%
ROUGE_f	0.21	95.5%	20.74%
BLEU	0.14	95.5%	14.44%
NIST	1.45	97.0%	N/A
cos_similarity	0.22	93.5%	22.27%
Euc_distance	−0.42	65.0%	N/A
METEOR	0.25	96.0%	25.33%
SPICE	0.16	93.5%	16.30%

Table 6. Paired statistical comparison of image-captioning metrics between ErgoChat and the general VLM. Note: The reported model results were obtained from a single fine-tuning and evaluation run; therefore, between-run standard deviations are not available.

Metric	General VLM Mean	ErgoChat Mean	Mean Paired Difference	95% Bootstrap CI
ROUGE_r	0.15	0.40	0.255	[0.235, 0.276]
ROUGE_p	0.35	0.42	0.075	[0.052, 0.097]
ROUGE_f	0.20	0.40	0.207	[0.190, 0.225]
BLEU	0.01	0.16	0.144	[0.131, 0.158]
NIST	0.17	1.61	1.446	[1.326, 1.564]
cos_similarity	0.17	0.39	0.223	[0.201, 0.245]
Euc_distance	8.97	8.55	−0.417	[−0.621, −0.205]
METEOR	0.11	0.36	0.253	[0.233, 0.274]
SPICE	0.10	0.26	0.163	[0.147, 0.179]

Since cosine similarity ranges from −1 to 1, the values obtained from Equations (3)–(5) are not particularly informative. Moreover, a smaller Euclidean distance signifies higher accuracy in the generated text. Consequently, the data presented in Tables 5 and 6 demonstrate that, apart from cosine similarity, other similarity metrics indicate that ErgoChat is more accurate than the general VLM in delivering ergonomic risk-related information to construction workers. The nine similarity metrics indicated greater similarity between ErgoChat-generated descriptions and the expert-authored reference descriptions than between the general VLM outputs and the references. Across the entire testing dataset, the average similarity scores obtained when evaluating the 9 similarity metrics against the ground truth are higher for the proposed ErgoChat than for the same architecture VLM. Specifically, the average ROUGE_r score of the proposed method surpasses that of the alternative method by 0.25531, corresponding to a 25.53% improvement, as the ROUGE_r metric ranges from 0 to 1. Moreover, the proposed method demonstrates higher similarity scores for 97% of the data in the entire testing dataset than the alternative model. Because both models generated descriptions for the same 200 test images, the captioning metric scores were analyzed as paired observations. For each metric, the mean paired difference between ErgoChat and the general VLM was calculated, and 95% confidence intervals were estimated using paired bootstrap resampling. Two-sided paired permutation tests were used to evaluate the differences, with Holm adjustment applied for multiple comparisons across the nine similarity metrics. The 95% confidence intervals excluded zero for all nine metrics, and all differences remained statistically significant after Holm adjustment (adjusted $p < 0.001$). ErgoChat achieved higher scores for ROUGE_r, ROUGE_p, ROUGE_f, BLEU, NIST, cosine similarity, METEOR, and SPICE, and a lower Euclidean distance than the general VLM. Corresponding results for other similarity metrics are detailed in Table 5.

Ergonomic annotation requires domain-specific expertise, and some variability may arise when experts interpret fine-grained posture characteristics. This subjectivity limits text-based metrics, so despite a 25.53% ROUGE_r improvement, the small absolute increase is expected. This difference should be interpreted in the context of the variability inherent in expert-authored ergonomic descriptions, whose value lies in driving task optimization, rather than in measuring whether the annotations are perfect.

For the evaluation of generated textual descriptions, in addition to the metric evaluation results in Tables 5 and 6, the human evaluation results are summarized in Table 7. The process for handling human evaluation results is illustrated in Figure 6. The fine-tuned choice rate is the average percentage of correct choices (i.e., the choices generated by the fine-tuned ErgoChat). The average accuracy improvement compared to the other choice represents the average improvement rated by the survey participants, and it represents the result for the following question: 'By how much do you think your selected description is more accurate than the other? 0% (about the same)–100% (absolutely more accurate)'.

Table 7. Results for human evaluation of ErgoChat-generated ergonomic risk-related text description.

Metric	Fine-Tuned Choice Rate	Mean Participant-Rated Relative Accuracy Difference
Value	81.69%	69.7%

After domain-specific fine-tuning, ErgoChat generated ergonomic-risk descriptions that more closely matched the reference descriptions on the held-out test dataset. The proposed method exhibited improvements in ROUGE_r, ROUGE_p, ROUGE_f, BLEU, NIST, cosine similarity, Euclidean distance, METEOR, and SPICE in text generation over the VLM without fine-tuning. Human evaluators chose the more accurate description from two method outputs in questionnaires. Across the 770 pairwise evaluations, participants

selected the fine-tuned ErgoChat description in 629 cases, corresponding to a preference rate of 81.69%. After accounting for clustering by participant and image, the estimated preference rate had a 95% confidence interval of 75.08–86.85%. The fine-tuned model was selected significantly more often than the 50% chance level ($z = 7.47$, $p < 0.001$). These results indicate a consistent participant preference for the descriptions generated by the fine-tuned model under the evaluation conditions of this study. Furthermore, participants who favored the fine-tuned output rated its relative advantage at an average of 69.7% on the questionnaire scale. Based on the results above, the proposed method achieved higher average caption-evaluation scores than the model with the same architecture without domain-specific fine-tuning. Figure 7 illustrates an actual test case in the testing dataset. On the left side of the image are the questions from IC, along with the answers provided by ErgoChat, while on the right side are the questions from VQA, along with the answers provided by ErgoChat. Figure 8 shows the IC and VQA generated by the other model for the same image.

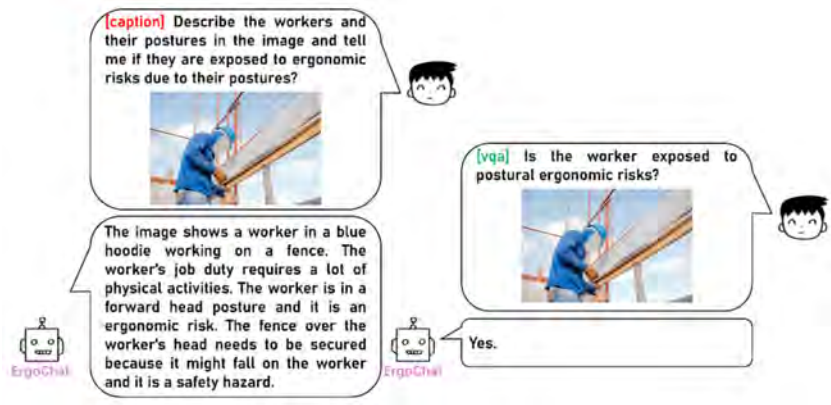

Figure 7. An example of IC and VQA generated by ErgoChat from the testing partition. It describes the posture that leads to the risk in the IC task.

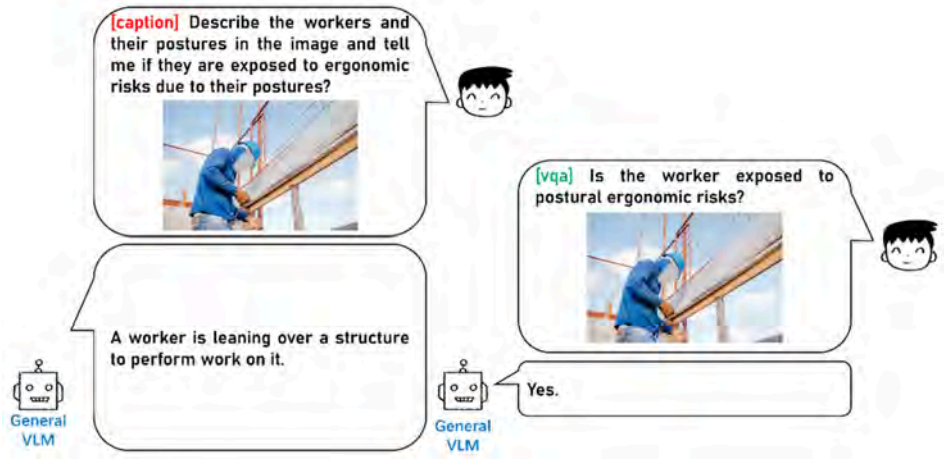

Figure 8. IC and VQA generated by the other model for the same image. It only describes posture and ignores the prompt's requirement of risk description.

The results indicate that the proposed approach can generate natural-language descriptions of image-based ergonomic risks and perform ergonomic-risk-related VQA under the experimental conditions of this study. ErgoChat achieved higher caption-similarity scores and a slightly higher VQA accuracy than the corresponding general VLM. Although the proposed method is effective, there is still room for improvement. ErgoChat's visual perception capabilities are not perfect and sometimes lead it to incorrectly describe workers' behavior during IC tasks. This error may prevent people from accurately identifying er-

gonomic risks. These problems are expected to improve with training ErgoChat on broader, more diverse image datasets.

ErgoChat was developed to automatically identify ergonomic risks associated with WMSDs to assist construction safety practitioners. It is designed to issue early warnings and address ergonomic risks and related safety issues facing construction workers. What distinguishes it from conventional AI algorithms used to identify WMSD-related ergonomic risks is its capacity to produce descriptions of ergonomic risks in human language rather than solely determining whether workers are exposed to them. ErgoChat can potentially generate ergonomic injury reports to help safety personnel identify risks and conduct safety checks. This capability is important because it provides natural-language output that is generally not produced by conventional classification-based ERA methods in completing text-heavy safety documents. Consequently, ErgoChat functions as an interactive tool for WMSD-related ergonomic hazard identification, facilitating ERA even for non-experts, and may provide a basis for future image-based ergonomic hazard-monitoring applications. ErgoChat is intended to function as an ergonomic screening and decision-support tool for safety practitioners rather than as a replacement for professional ergonomists. Its output should be reviewed by qualified personnel, particularly when consequential workplace safety decisions are made.

5. Discussion

The 9 similarity metrics and human evaluation for IC all indicate that the proposed method outperforms the method of the same architecture in terms of ergonomic IC performance. Additionally, both the perplexity and VQA results demonstrate that the proposed method surpasses the method of the same architecture in ergonomic identification VQA performance. The evaluation results support further investigation of ErgoChat as a decision-support tool for ergonomic risk identification on construction sites.

In the context of VQA evaluation, ErgoChat realizes a 1.5% increase in accuracy, achieving a total of 96.5%. The class distribution of the testing dataset should be considered when interpreting the classification performance for binary VQA. Because 95% of the test images depict workers exposed to ergonomic risks, overall accuracy can be influenced by the majority class. Therefore, the reported classification performance for binary VQA should not be interpreted as balanced discrimination between risk and no-risk cases. Future evaluation should therefore employ more balanced datasets and report class-specific measures, such as sensitivity, specificity, and balanced accuracy. However, this imbalance also reflects the nature of construction work, in which workers frequently perform physically demanding tasks and adopt awkward or high-load postures associated with WMSD risk; consequently, photos depicting construction activities are more likely to contain observable postural ergonomic risks than risk-free postures. Human evaluation results suggest that 81.69% of the text descriptions generated by ErgoChat are more accurate than those from the other model, with a mean participant-rated relative accuracy difference for these descriptions being 69.7% higher than that of the other model's outputs.

In the IC task, ErgoChat performed better overall than the comparison model. Although the average cosine distance difference between the two models was slight, the remaining eight evaluation similarity metrics supported this conclusion. ErgoChat achieved higher scores than the comparison model for six metrics on more than 90% of the test cases. The other two metrics show improved IC for more than 65% of the test dataset. Human evaluation results demonstrated that 81.69% of the text descriptions produced by the fine-tuned ErgoChat model were more accurate than those generated by the unmodified VLM.

When traditional AI approaches are applied to WMSD- and VQA-related tasks in the construction industry, it is difficult to accurately identify underlying ergonomic problems

and propose effective corrective measures in natural language. To address this challenge, ErgoChat was developed to generate text descriptions related to these ergonomic risks. Built on the prevalent GPT architecture, similar methods still have significant room for improvement in the above functionalities, even after training solely on generic datasets. To explore the feasibility and potential for performance improvement of ErgoChat, this study introduced a dataset for fine-tuning and testing. To mitigate ambiguity across different queries and tasks, the dataset includes task identifier tokens for [caption] and [vqa]. The token [caption] is used as the task identifier for IC in the dataset, whereas the token [vqa] is utilized as the task identifier for VQA annotations.

This study makes three main academic contributions. First, the study proposes an approach based on a VLM to interactively identify ergonomic risks related to WMSD in the construction industry. Second, a construction-ergonomics-specific image–text dataset was curated, containing task identifiers and natural-language annotations describing postural ergonomic risks in construction scenarios. This dataset will be publicly available for researchers to use. Third, this study combined quantitative analysis of nine similarity metrics with qualitative evaluations from 50 ergonomic evaluators to conduct a comprehensive assessment. The principal scientific contribution is therefore not a new foundation-model architecture, but evidence that structured domain knowledge and expert-informed multimodal annotations can be used to adapt a generic VLM toward interactive, image-grounded ergonomic reasoning. This extends automated ERA beyond risk classification toward interpretable natural-language explanations. Collectively, these findings provide evidence on how domain-specific fine-tuning and evaluation can be applied to VLMs for identifying and describing WMSD-related ergonomic risks in construction environments.

The results demonstrate the feasibility of representing domain-specific ergonomic information through natural-language descriptions in addition to risk labels. ErgoChat combines visual evidence and ERA tools, leveraging an instruction-based VQA approach to provide natural-language information in addition to a risk-level judgment. The natural-language output could potentially support risk-prevention activities, safety documentation, and the development of risk-mitigation recommendations. Hence, it can reduce the burden of routine documentation when expert resources are limited. As an interactive tool, it may also be investigated as a potential educational or query-support tool for construction safety personnel. The ErgoChat proposed in this study is expected to help identify ergonomic risks for construction workers, with the intended goal of enhancing their well-being and the safety of their work environment.

Although this approach achieves accuracy improvement, it faces limitations like hallucination and perceptual issues in LLMs and VLMs. The study did not vary prompts, as prompt engineering was outside its scope, so the impact of different prompts on IC and VQA results was not explored. In addition, due to the limited number of freely available images online, the dataset still needs to be expanded. Although the test image–text data uses actual photos of construction workers, it has not yet undergone prospective longitudinal validation during routine construction-site operations. The dataset has more images with ergonomic risks than without, making it easy for models to be biased toward predicting higher risk and to overperform in low-risk scenarios. Although manual inspection was used to remove exact duplicates, cropped versions, and closely related sequential images across the fine-tuning and testing partitions, source-level independence cannot be completely guaranteed. In particular, overlap involving the same worker, construction site, or original online source was not systematically tracked. Such overlap, if present, could increase visual similarity between the fine-tuning and testing data and potentially lead to optimistic estimates of generalization.

These limitations directly affect the interpretation and generalizability of the reported performance. First, the limited size and diversity of the dataset mean that the performance may decrease for unfamiliar tasks, worker characteristics, viewpoints, severe occlusions, or substantially different site conditions. Second, because a fixed prompt was used throughout the evaluation, the reported VQA and captioning results do not reflect the performance of different prompting strategies. In addition, hallucination and visual-perception errors mean that generated descriptions may occasionally contain incorrect, omitted, or unsupported ergonomic information and should therefore be reviewed by a knowledgeable user before informing safety-related decisions. Finally, because human evaluation used different questionnaire subsets, the current findings should be interpreted as experimental evidence of feasibility rather than evidence of established field effectiveness. Accordingly, ErgoChat can currently be considered most reliable as a human-supervised screening and decision-support tool for image-based assessment, rather than as a replacement for professional ergonomic assessment.

To improve IC quality, integrate more mechanisms connecting the vision encoder to the LLM and use more consistent datasets. Future studies could collect more construction site images to expand the dataset. Although generative AI does not clearly outperform traditional ML in estimating joint positions or angles, future work might incorporate ergonomic assessment tools like REBA into training data. The general VLM baseline uses the same underlying MiniGPT-v2 architecture as ErgoChat so that the comparison isolates the effect of construction-ergonomics-specific fine-tuning while controlling model architecture. Consequently, the present comparison should not be interpreted as demonstrating superiority over other contemporary VLM architectures. Future studies will extend the evaluation to external VLM baselines, under consistent datasets and prompting conditions. Future research will unify the survey questionnaire across all evaluators, thereby supporting the implementation of paired statistical testing.

Author Contributions: Conceptualization, C.F., Q.M. and X.L.; methodology, C.F., Q.M. and X.L.; software, X.W.; validation, C.F. and X.W.; formal analysis, C.F. and X.W.; investigation, C.F.; resources, Q.M. and X.L.; data curation, X.W.; writing—original draft preparation, C.F.; writing—review and editing, X.W., Q.M. and X.L.; visualization, C.F.; supervision, Q.M. and X.L.; project administration, X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Alliance Grant with Alberta Innovates, grant number ALLRP 561120-20.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Boards of the University of Alberta (protocol code pro00111404 and date of approval 8 May 2025).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The model code and annotation metadata will be made publicly available at https://github.com/Crocodile-Chris/ErgoChat (accessed on 1 April 2026). Construction images will be redistributed only where the original source license explicitly permits redistribution; for images that cannot legally be redistributed, the dataset release will provide the corresponding annotations and source information rather than republishing the image files. Human-evaluation response data contain participant information and will be made available in de-identified form upon reasonable request, subject to the approved ethics protocol.

Acknowledgments: We are very grateful to the ergonomists who participated in this research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

VLMs	Vision–Language Models
VQA	Visual Question Answering
IC	Image Captioning
WMSD	Work-related Musculoskeletal Disorders
ERA	Ergonomic Risk Assessment
AI	Artificial Intelligence
ML	Machine Learning
CV	Computer Vision
LLMs	Large Language Models
ViT	Vision Transformer
REBA	Rapid Entire Body Assessment
RULA	Rapid Upper Limb Assessment
JSON	JavaScript Object Notation

References

1. *U.S. Bureau of Labor Statistics Injuries, Illnesses, and Fatalities*; U.S. Bureau of Labor Statistics: Washington, DC, USA, 2024.
2. National Work Injury/Disease Statistics Program (NWISP). Available online: https://awcbc.org/data-and-statistics (accessed on 10 July 2026).
3. *European Agency for Safety and Health at Work Work-Related Musculoskeletal Disorders: Prevalence, Costs and Demographics in the EU*; Publications Office of the European Union: Luxembourg, 2019.
4. David, G.C. Ergonomic Methods for Assessing Exposure to Risk Factors for Work-Related Musculoskeletal Disorders. *Occup. Med.* **2005**, *55*, 190–199. [CrossRef] [PubMed]
5. MassirisFernández, M.; Fernández, J.Á.; Bajo, J.M.; Delrieux, C.A. Ergonomic Risk Assessment Based on Computer Vision and Machine Learning. *Comput. Ind. Eng.* **2020**, *149*, 106816. [CrossRef]
6. Jeong, S.; Kook, J. CREBAS: Computer-Based REBA Evaluation System for Wood Manufacturers Using MediaPipe. *Appl. Sci.* **2023**, *13*, 938. [CrossRef]
7. Barberi, E.; Chillemi, M.; Cucinotta, F.; Milardi, D.; Raffaele, M.; Salmeri, F.; Sfravara, F. Posture Interactive Self Evaluation Algorithm Based on Computer Vision. In *Advances on Mechanics, Design Engineering and Manufacturing IV*; Gerbino, S., Lanzotti, A., Martorelli, M., Mirálbes Buil, R., Rizzi, C., Roucoules, L., Eds.; Lecture Notes in Mechanical Engineering; Springer International Publishing: Cham, Switzerland, 2023; pp. 1516–1526.
8. Nayak, G.K.; Kim, E. Development of a Fully Automated RULA Assessment System Based on Computer Vision. *Int. J. Ind. Ergon.* **2021**, *86*, 103218. [CrossRef]
9. Seo, J.; Yin, K.; Lee, S. Automated Postural Ergonomic Assessment Using a Computer Vision-Based Posture Classification. In *Proceedings of the Construction Research Congress 2016*; American Society of Civil Engineers: San Juan, Puerto Rico, 2016; pp. 809–818.
10. Plantard, P.; Shum, H.P.H.; Le Pierres, A.-S.; Multon, F. Validation of an Ergonomic Assessment Method Using Kinect Data in Real Workplace Conditions. *Appl. Ergon.* **2017**, *65*, 562–569. [CrossRef] [PubMed]
11. Vignais, N.; Bernard, F.; Touvenot, G.; Sagot, J.-C. Physical Risk Factors Identification Based on Body Sensor Network Combined to Videotaping. *Appl. Ergon.* **2017**, *65*, 410–417. [CrossRef] [PubMed]
12. Donisi, L.; Cesarelli, G.; Pisani, N.; Ponsiglione, A.M.; Ricciardi, C.; Capodaglio, E. Wearable Sensors and Artificial Intelligence for Physical Ergonomics: A Systematic Review of Literature. *Diagnostics* **2022**, *12*, 3048. [CrossRef] [PubMed]
13. Fan, C.; Mei, Q.; Li, X. 3D Pose Estimation Dataset and Deep Learning-Based Ergonomic Risk Assessment in Construction. *Autom. Constr.* **2024**, *164*, 105452. [CrossRef]
14. Yenduri, G.; Ramalingam, M.; Selvi, G.C.; Supriya, Y.; Srivastava, G.; Maddikunta, P.K.R.; Raj, G.D.; Jhaveri, R.H.; Prabadevi, B.; Wang, W.; et al. GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access* **2024**, *12*, 54608–54649. [CrossRef]
15. Aliasgari, R.; Fan, C.; Li, X.; Golabchi, A.; Hamzeh, F. MOCAP and AI-Based Automated Physical Demand Analysis for Workplace Safety. *J. Constr. Eng. Manag.* **2024**, *150*, 04024060. [CrossRef]
16. Vijayakumar, R.; Choi, J. Emerging Trends of Ergonomic Risk Assessment in Construction Safety Management: A Scientometric Visualization Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 16120. [CrossRef] [PubMed]

17. Prince, S.A.; Adamo, K.B.; Hamel, M.; Hardt, J.; Connor Gorber, S.; Tremblay, M. A Comparison of Direct versus Self-Report Measures for Assessing Physical Activity in Adults: A Systematic Review. *Int. J. Behav. Nutr. Phys. Act.* **2008**, *5*, 56. [CrossRef] [PubMed]
18. Crawford, J.O. The Nordic Musculoskeletal Questionnaire. *Occup. Med.* **2007**, *57*, 300–301. [CrossRef]
19. Dawson, A.P.; Steele, E.J.; Hodges, P.W.; Stewart, S. Development and Test–Retest Reliability of an Extended Version of the Nordic Musculoskeletal Questionnaire (NMQ-E): A Screening Instrument for Musculoskeletal Pain. *J. Pain* **2009**, *10*, 517–526. [CrossRef] [PubMed]
20. Li, X.; Han, S.; Gul, M.; Al-Hussein, M. Automated Ergonomic Risk Assessment Based on 3D Visualization. In *Proceedings of the International Symposium on Automation and Robotics in Construction*; IAARC Publications: Taipei, Taiwan, 2017.
21. Guo, S.Y.; Ding, L.Y.; Luo, H.B.; Jiang, X.Y. A Big-Data-Based Platform of Workers' Behavior: Observations from the Field. *Accid. Anal. Prev.* **2016**, *93*, 299–309. [CrossRef] [PubMed]
22. Marklin, R.W.; Wilzbacher, J.R. Four Assessment Tools of Ergonomics Interventions: Case Study at an Electric Utility's Warehouse System. *Am. Ind. Hyg. Assoc. J.* **1999**, *60*, 777–784. [CrossRef] [PubMed]
23. Lowe, B.D.; Dempsey, P.G.; Jones, E.M. Ergonomics Assessment Methods Used by Ergonomics Professionals. *Appl. Ergon.* **2019**, *81*, 102882. [CrossRef] [PubMed]
24. Hignett, S.; McAtamney, L. Rapid Entire Body Assessment (REBA). *Appl. Ergon.* **2000**, *31*, 201–205. [CrossRef] [PubMed]
25. McAtamney, L.; Nigel Corlett, E. RULA: A Survey Method for the Investigation of Work-Related Upper Limb Disorders. *Appl. Ergon.* **1993**, *24*, 91–99. [CrossRef] [PubMed]
26. Fox, R.R.; Lu, M.-L.; Occhipinti, E.; Jaeger, M. Understanding Outcome Metrics of the Revised NIOSH Lifting Equation. *Appl. Ergon.* **2019**, *81*, 102897. [CrossRef] [PubMed]
27. Matthews, J.D.; MacKinnon, S.N.; Albert, W.J.; Holmes, M.; Patterson, A. Effects of Moving Environments on the Physical Demands of Heavy Materials Handling Operators. *Int. J. Ind. Ergon.* **2007**, *37*, 43–50. [CrossRef]
28. Yunus, M.N.H.; Jaafar, M.H.; Mohamed, A.S.A.; Azraai, N.Z.; Hossain, M.S. Implementation of Kinetic and Kinematic Variables in Ergonomic Risk Assessment Using Motion Capture Simulation: A Review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8342. [CrossRef] [PubMed]
29. Colim, A.; Faria, C.; Braga, A.C.; Sousa, N.; Rocha, L.; Carneiro, P.; Costa, N.; Arezes, P. Towards an Ergonomic Assessment Framework for Industrial Assembly Workstations—A Case Study. *Appl. Sci.* **2020**, *10*, 3048. [CrossRef]
30. Onofrejova, D.; Balazikova, M.; Glatz, J.; Kotianova, Z.; Vaskovicova, K. Ergonomic Assessment of Physical Load in Slovak Industry Using Wearable Technologies. *Appl. Sci.* **2022**, *12*, 3607. [CrossRef]
31. Atici, H.; Gonen, D.; Oral, A.; Kaya, B. Ergonomic Analysis of an Assembly Line Using the AnyBody Modeling System. In Proceedings of the 3nd World Congress on Mechanical, Chemical, and Material Engineering, MCM 2017, Rome, Italy, 8–10 June 2017.
32. Mortensen, J.; Trkov, M.; Merryweather, A. Improved Ergonomic Risk Factor Assessment Using Opensim and Inertial Measurement Units. In *Proceedings of the 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*; ACM: Washington, DC, USA, 2018; pp. 27–28.
33. Sekkay, F. Prevention of Work-Related Musculoskeletal Disorders Supported by Artificial Intelligence. In Proceedings of the 10th International Conference on Human Interaction and Emerging Technologies (IHIET 2023), Nice, France, 22–24 August 2023.
34. Cai, J.; Li, X.; Liang, X.; Wei, W.; Li, S. Construction Worker Ergonomic Assessment via LSTM-Based Multi-Task Learning Framework. In *Proceedings of the Construction Research Congress 2022*; American Society of Civil Engineers: Arlington, VA, USA, 2022; pp. 215–224.
35. Damilos, S.; Saliakas, S.; Karasavvas, D.; Koumoulos, E.P. An Overview of Tools and Challenges for Safety Evaluation and Exposure Assessment in Industry 4.0. *Appl. Sci.* **2024**, *14*, 4207. [CrossRef]
36. Rane, N.; Choudhary, S.; Rane, J. Integrating ChatGPT, Bard, and Leading-Edge Generative Artificial Intelligence in Building and Construction Industry: Applications, Framework, Challenges, and Future Scope. *SSRN J.* **2023**. [CrossRef]
37. Yong, G.; Liu, M.; Lee, S. Automated Captioning for Ergonomic Problem and Solution Identification in Construction Using a Vision-Language Model and Caption Augmentation. In *Proceedings of the Construction Research Congress 2024*; American Society of Civil Engineers: Des Moines, IA, USA, 2024; pp. 709–718.
38. Yong, G.; Liu, M.; Lee, S. Explainable Image Captioning to Identify Ergonomic Problems and Solutions for Construction Workers. *J. Comput. Civ. Eng.* **2024**, *38*, 04024022. [CrossRef]
39. Smetana, M.; Salles De Salles, L.; Sukharev, I.; Khazanovich, L. Highway Construction Safety Analysis Using Large Language Models. *Appl. Sci.* **2024**, *14*, 1352. [CrossRef]
40. Uddin, S.M.J.; Albert, A.; Ovid, A.; Alsharef, A. Leveraging ChatGPT to Aid Construction Hazard Recognition and Support Safety Education and Training. *Sustainability* **2023**, *15*, 7121. [CrossRef]
41. Yoo, B.; Kim, J.; Park, S.; Ahn, C.R.; Oh, T. Harnessing Generative Pre-Trained Transformers for Construction Accident Prediction with Saliency Visualization. *Appl. Sci.* **2024**, *14*, 664. [CrossRef]

42. Chen, H.; Hou, L.; Wu, S.; Zhang, G.; Zou, Y.; Moon, S.; Bhuiyan, M. Augmented Reality, Deep Learning and Vision-Language Query System for Construction Worker Safety. *Autom. Constr.* **2024**, *157*, 105158. [CrossRef]
43. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*; MLResearchPress: Brookline, MA, USA, 2022.
44. Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; Elhoseiny, M. MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-Task Learning. *arXiv* **2023**, arXiv:2310.09478. [CrossRef]
45. Lecler, A.; Duron, L.; Soyer, P. Revolutionizing Radiology with GPT-Based Models: Current Applications, Future Possibilities and Limitations of ChatGPT. *Diagn. Interv. Imaging* **2023**, *104*, 269–274. [CrossRef] [PubMed]
46. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL '02*; Association for Computational Linguistics: Philadelphia, PA, USA, 2001; p. 311.
47. Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; Cucchiara, R. From Show to Tell: A Survey on Deep Learning-Based Image Captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 539–559. [CrossRef] [PubMed]
48. Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Vancouver, BC, Canada, 2023; pp. 19358–19369.
49. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
50. Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; Wei, F. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv* **2023**, arXiv:2306.14824.
51. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv* **2021**, arXiv:2111.02114.
52. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 2556–2565.
53. Shawe-Taylor, J.; Neural Information Processing Systems Foundation (Eds.) Advances in Neural Information Processing Systems 24. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems 2011, Granada, Spain, 12–15 December 2011*; Curran: Red Hook, NY, USA, 2012.
54. Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; Zitnick, C.L. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv* **2015**, arXiv:1504.00325.
55. Sidorov, O.; Hu, R.; Rohrbach, M.; Singh, A. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12347, pp. 742–758.
56. Kazemzadeh, S.; Ordonez, V.; Matten, M.; Berg, T. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 787–798.
57. Yu, L.; Poirson, P.; Yang, S.; Berg, A.C.; Berg, T.L. Modeling Context in Referring Expressions. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9906, pp. 69–85.
58. Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; Murphy, K. Generation and Comprehension of Unambiguous Object Descriptions. *arXiv* **2015**, arXiv:1511.02283.
59. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
60. Hudson, D.A.; Manning, C.D. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Long Beach, CA, USA, 2019; pp. 6693–6702.
61. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Honolulu, HI, USA, 2017; pp. 6325–6334.
62. Mishra, A.; Shekhar, S.; Singh, A.K.; Chakraborty, A. OCR-VQA: Visual Question Answering by Reading Text in Images. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*; IEEE: Sydney, Australia, 2019; pp. 947–952.

63. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Long Beach, CA, USA, 2019; pp. 3190–3199.
64. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *Computer Vision—ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2022; Volume 13668, pp. 146–162.
65. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
66. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Santiago, Chile, 2015; pp. 2641–2649.
67. Honovich, O.; Scialom, T.; Levy, O.; Schick, T. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 14409–14428.
68. Houghton, E. Man in Green and Yellow Safety Vest and Blue Denim Jeans Standing on Brown Wooden Pallet. 2021. Available online: https://unsplash.com/photos/man-in-green-and-yellow-safety-vest-and-blue-denim-jeans-standing-on-brown-wooden-pallet-EixJzIdl4bc (accessed on 1 April 2026).
69. Jelinek, F.; Mercer, R.L.; Bahl, L.R.; Baker, J.K. Perplexity—A Measure of the Difficulty of Speech Recognition Tasks. *J. Acoust. Soc. Am.* **1977**, *62*, S63. [CrossRef]
70. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; Volume Text Summarization Branches Out.
71. Doddington, G. Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*; Association for Computational Linguistics: San Diego, CA, USA, 2002; p. 138.
72. Pal, S.; Chang, M.; Iriarte, M.F. Summary Generation Using Natural Language Processing Techniques and Cosine Similarity. In *Intelligent Systems Design and Applications*; Abraham, A., Gandhi, N., Hanne, T., Hong, T.-P., Nogueira Rios, T., Ding, W., Eds.; Lecture Notes in Networks and Systems; Springer International Publishing: Cham, Switzerland, 2022; Volume 418, pp. 508–517.
73. Salihu, S.A.; Onyekwere, I.P.; Mabayoje, M.A.; Mojeed, H.A. Performance Evaluation of Manhattan and Euclidean Distance Measures for Clustering Based Automatic Text Summarization. *FUOYE J. Eng. Technol.* **2019**, *4*, 135–139. [CrossRef]
74. Lavie, A.; Agarwal, A. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation—StatMT '07*; Association for Computational Linguistics: Prague, Czech Republic, 2007; pp. 228–231.
75. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In *Computer Vision–ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9909, pp. 382–398.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.