Assisting in the identification of ergonomic risks for workers: a large vision-language model approach

Chao Fan[1], Qipei Mei[2], and Xinming Li[1]

[1]Department of Mechanical Engineering, University of Alberta, Canada
[2] Department of Civil Engineering, University of Alberta, Canada
cfan3@ualberta.ca, qipei@ualberta.ca, xinming1@ualberta.ca

Abstract –

In the construction industry, due to workers frequently engaging in highly physically demanding tasks and using various tools, workers are often exposed to ergonomic risks and safety hazards. Various observation-based traditional or computer vision-based artificial intelligence methods have been applied in the field of construction to assess ergonomic risks. However, the method of assessing ergonomic risks using Generative Pretrained Transformers (GPT) based visual language models has not been thoroughly explored. This study explores its unique ability in visual-text interaction to extract ergonomic risk information from images and generate corresponding human-like language descriptions. To test the feasibility and performance of the proposed method, two datasets were created. Each dataset contained 100 different scenarios with ergonomic risk information for finetuning and testing. Performance after finetuning the vision-language model with the finetuning dataset outperformed the model before finetuning; the results showed that the fine-tuned model achieved an accuracy of 81%, while the model before finetuning only achieved 28% accuracy. Therefore, the proposed method offers an automated, real-time, non-traditional artificial intelligence approach for identifying ergonomic risks and providing human-like language descriptions. This expands the perspective of health and safety-related problem-solving and promotes the prevention of work-related musculoskeletal disorders (WMSD) in the construction industry.

Keywords –

Ergonomic Risks Identification; Work Safety; GPT; Vision Language Model; Construction Safety

1 Introduction

The timely recognition and mitigation of workplace hazards play a pivotal role in safeguarding the safety, health, and productivity of the working environment. Job tasks in construction often expose workers to conditions that entail repetitive movements, strenuous labor, and awkward body postures, subtly but adversely affecting their well-being. The consequential impact of these operations frequently manifests in the development of WMSD. Based on data from the European Agency for Safety and Health at Work, WMSDs exhibit the highest prevalence within the construction sector when compared to other industries. [1] In 2018, 52% of individuals employed in construction reported instances of backaches, while 54% experienced WMSD in the upper limbs and 41% in the lower limbs. The construction and manufacturing sectors demonstrate a high percentage of sick leave attributed to WMSD due to the physically demanding nature of the work. [1] Notably, the United States Bureau of Labor Statistics reported that WMSD accounted for 30% of occupational injuries and illnesses in 2018. Moreover, injuries and illnesses related to WMSD in the construction industry ranked fifth highest among all industries. [2] Consequently, proactive identification and prevention of WMSDs and associated health risks prove to be highly constructive.

Conventional approaches to managing worker safety and health heavily depend on human observation, self-reporting, and direct measurements. [3]-[6] In these traditional methods, ergonomists utilize manual procedures to detect ergonomic risks, frequently incorporating the observation and interviewing of workers. In essence, the subjective and time-consuming aspects inherent in the traditional identification of ergonomic risks impede its efficacy in promptly analyzing and preventing such risks. [7], [8] Consequently, it is imperative to explore methodologies that can objectively and swiftly pinpoint ergonomic risks.

To address the inherent subjectivity and time-consuming limitations associated with traditional ergonomic risk identification, researchers have shown a growing interest in automated approaches. Current research in automated ergonomic risk identification predominantly centers around computer vision-based methods [4], [9]-[14], with vision-language models yet to receive widespread attention. These computer vision-based methods directly train on visual data to learn

patterns, particularly focusing on human joints for ergonomic risk identification. For instance, these computer vision methods implement ergonomic risk identification by training on visual data related to human joints. Thus, they predict ergonomic risks by estimating the joints of workers and calculating joint angles based on these estimations.

On the other hand, vision-language model approaches, employing large language models (LLM) and 'zero-shot learning,' are gaining recognition for their ability to generate human-like descriptions. [15] They are pre-trained on LLM/large text corpora and subsequently finetuned for vision tasks using image-text pairs [15]-[17]. Unlike conventional computer vision methods, vision-language model-based approaches offer a unique perspective by generating human-like descriptions instead of numerical representations, thereby providing a more comprehensive understanding of ergonomic risk identification. GPT-4V [18] by OpenAI and MiniGPT-4 [15] are two approaches based on vision-language models. Despite both incorporating LLM for initial pretraining and refining their models through subsequent finetuning with image-text pairs, it is noteworthy that the specific LLM and image-text pairs employed in each approach are unique. Furthermore, the unavailability of the source code and model for GPT-4V poses a limitation, as it hinders the ability to finetune its model using personalized image-text pairs for the identification of ergonomic risks. This lack of accessibility underscores the challenges associated with adapting the method for specific applications when critical components are not openly accessible. This study employed the MiniGPT-4, incorporating a vision encoder with a pre-trained Vision Transformer and Q-Former, a single linear projection layer, and the Vicuna LLM [15]. It is important to highlight that MiniGPT4 incorporates image-text pairs for finetuning pre-trained models, but the images within these pairs are not related to construction workers. Furthermore, the text captions associated with these images do not provide any descriptions of ergonomic risks related to the individuals depicted in the images. To equip the vision-language model finetuned with pertinent visual knowledge related to ergonomic risks, the image-text pair dataset offered by MiniGPT-4, limited to daily scenes or common objects, proves inadequate for effectively addressing scenarios associated with the identification of ergonomic risks in the context of workers. To address this challenge, this study curates datasets containing images of construction workers along with corresponding text descriptions pertaining to ergonomic risks.

During the quest for engineering-related studies, an inquiry into the terms "large language model" and "construction" on Scopus produced 25 results, yet merely 3 were pertinent to applications in the construction domain. A similar search combining "vision language model" and "construction" revealed only 1 relevant outcome within the construction field. Regrettably, no results were obtained for the query pairing "large language model" and "ergonomic" in the Scopus database. Prior studies have utilized GPT models for tasks such as question answering, extracting information from Building Information Modeling (BIM) datasets, and optimizing scheduling and sequencing in engineering contexts. Zheng et al. (2023) presented a prompt-driven virtual assistant framework aimed at bolstering natural language-centric BIM search by integrating GPT technologies. This framework autonomously interprets users' natural language inquiries, retrieves pertinent information, and provides succinct natural language responses alongside corresponding 3D visualizations via a user interface. You et al. (2023) introduced a methodology harnessing the capabilities of ChatGPT to realize automated sequence planning in robotic assembly for construction. The efficacy of this method was demonstrated through its ability to decrease reliance on manual intervention, shorten planning durations, and enhance the overall efficiency of robot-driven assembly processes within the construction sector. Prieto et al. (2023) assessed the suitability of GPT for aiding in the creation of an automated construction schedule using prompts expressed in natural language. Chen et al. (2024) research presents an interactive query system based on Augmented Reality and Deep Learning for delivering real-time safety information through the analysis of on-site images. Table **1** shows the search results on Scopus of relevant applications in the construction sector.

Table 1. Relevant applications in the construction sector

Study	Application
Zheng et al. [19]	Natural language-based BIM information retrieval
You et al. [20]	Automated sequence planning in robotic assembly for construction tasks
Prieto et al. [21]	Automated construction schedule
Chen et al. [22]	Augmented reality-based safety information retrieval

Current vision-language model methodologies in construction predominantly concentrate on tasks such as BIM information retrieval, automated construction schedule generation, sequence planning, and safety information querying. To the best of our knowledge, even with the presence of vision-language models such as GPT-4V and MiniGPT-4, there is currently no

specialized vision-language model approach explicitly developed for the identification of ergonomic risks. In other words, without the presence of ergonomic experts, accomplishing real-time identification of ergonomic risks and generating human-like descriptions of the risks for timely prevention of WMSDs on construction sites or in front of surveillance screens becomes challenging. To address the identified gap of lacking methods for identifying and generating human-like language descriptions of the ergonomic risks, this study aims to introduce a vision-language model-driven approach. This method facilitates users without specialized expertise to obtain human-like language descriptions of ergonomic risks occurring on construction sites, with the objective of mitigating WMSDs. Additionally, the study seeks to investigate the feasibility of implementing such methodologies. The code, finetuned model, pre-trained model, and the text descriptions in the image-text pair datasets can be accessed here:
https://github.com/xinmingliUofA/ErgoGPT

2 Methodology

This study proposes a large vision-language model-based ergonomic risk identification and seeks to assess the viability and performance of employing the vision-language model approach in the domain of ergonomics risk identification. To fill the research gap of lacking image-text pair datasets for ergonomic risk identification, finetuning and testing datasets comprising image-text pairs that portray scenarios relevant to ergonomic risks faced by construction workers were generated. In the first step, we conducted finetuning on the pre-existing Vicuna LLM using the finetuning dataset we curated, consisting of image-text pairs depicting ergonomic risks. Subsequently, we evaluated the performance of the finetuned vision-language model using the distinct dataset of image-text pairs designed specifically for testing its capability in identifying ergonomics risks. Subsequently, this study conducted a performance comparison between the model finetuned using the curated data for ergonomic risk identification and the model finetuned using generic data supplied by MiniGPT-4.

The MiniGPT-4 establishes a connection between the visual encoder and the LLM through the integration of a linear projection layer. MiniGPT-4 employs the open-sourced Vicuna as its language decoder, an LLM built on the foundation of LLaMA, capable of executing diverse and intricate linguistic tasks. For visual perception, it utilizes the open-sourced visual encoder in BLIP-2, incorporating a ViT backbone paired with their pre-trained Q-Former. The pre-trained model obtained through MiniGPT-4 is employed to acquire vision-language knowledge from an extensive dataset of image-text pairs. The pretraining process involved the utilization of approximately 5 million image-text pairs sourced from the Conceptual Caption, SBU, and LAION datasets. [15]

The methodology section of this study is structured into three main steps. Initially, an image-text pair dataset was curated by a professional ergonomist, who provided text descriptions highlighting ergonomic risks associated with each photograph depicting construction workers' activities. These descriptions were subsequently transformed into an annotation file using a Python script. Secondly, the study involved finetuning a Vicuna LLM vision-language model utilizing the image-text pair dataset and a framework built upon MiniGPT. Additionally, a separate vision-language model exclusively employing Vicuna LLM was trained. Thirdly, the performance of both models derived from the second step was assessed. Specifically, the text descriptions generated by each model were compared to the ground truth using the testing dataset.

2.1 Generating alignment data for the vision-language approach

As models undergo pre-training in the initial phase, they sometimes face challenges in producing coherent human language output. Consequently, additional finetuning is necessary to address issues like repetitive phrases, fragmented sentences, and irrelevant content that may arise in models that are solely pre-trained. Similar issues are present in other approaches utilizing large language models, like GPT-3. GPT-3.5 mitigates these issues through a combination of instruction finetuning and reinforcement learning. [15], [23] Hence, to make the output results fluent and more akin to human language, finetuning the pre-trained model is indispensable.

Datasets for finetuning instruction and conversations are abundant in the field of natural language processing/large language models, yet they remain scarce in the realm of vision language, particularly in the context of ergonomic risk identification within ergonomic risk identification. As the primary objective of this study is to investigate the feasibility and proficiency of large vision-language models in responding to questions related to ergonomic risk assessment, the datasets are rooted in real-world situations involving construction workers. The dataset employed for finetuning incorporated 100 images featuring real-world construction workers engaged in various tasks, accompanied by captions detailing ergonomic risks associated with the depicted scenarios. Likewise, the dataset used for testing the finetuned vision-language model comprised similar images and captions focusing on ergonomics risks. The models used for testing include one finetuned with generic image-text pairs unrelated to ergonomic risks,

provided by MiniGPT-4, and the other finetuned using the dataset created in this study for finetuning purposes.

The image component of the generic data supplied by MiniGPT-4 for the finetuning of models does not pertain to construction workers or the identification of ergonomic risks—many of the images solely depict mere objects. Furthermore, the text component in the data provided by MiniGPT-4 lacks descriptions related to ergonomic risks or workers' postures that might cause ergonomic issues. Consequently, the approach employed in this study for the generation of image-text alignment data differs from that of MiniGPT-4. In the process of generating alignment data, MiniGPT-4 generates the description of chosen images using a pre-trained model and subsequently employs ChatGPT to refine these descriptions, addressing issues such as repetitive words and fragmented sentences. In contrast to MiniGPT-4, the text component of the datasets in this study was curated with input from an experienced ergonomist. In particular, the text descriptions aligned with the images in the image-text pairs of the finetuning and testing datasets were created by an ergonomist with three years of professional expertise. The internet-sourced images portraying construction workers in action, utilized as the dataset for finetuning and testing, were supplied to the ergonomist. Subsequently, the ergonomist assessed each image, responding to the question 'Describe the workers and their postures in the image and tell me if they are exposed to ergonomic risks due to their postures?' The ergonomist's responses were documented in a CSV file as ground truth.

This study chooses to adopt a methodology different from that of MiniGPT-4 for generating text descriptions corresponding to images in the dataset. The rationale lies in the fact that the generic datasets (Conceptual Caption, SBU, and LAION) employed for model pretraining were not crafted by ergonomists and lacked comprehensive ergonomic-related descriptions for the images within the dataset. Simply put, in light of the requirement for domain-specific knowledge in ergonomics within the pre-trained model, rather than employing the pre-trained model method in MiniGPT-4 to generate the text component for images, this study opts for input from ergonomists to guarantee the reliability of the relevant knowledge. Consequently, employing a pre-trained model to generate descriptions for selected images related to ergonomic risks is not deemed an optimal approach in this study.

The dataset creation process in this study unfolds through three stages. Initially, images are procured by conducting a search for 'construction worker' in a search engine, and relevant images depicting workers in action are selected from the search results. In this process, a total of 200 images were carefully chosen, with 100 allocated to the dataset later employed for finetuning the pre-trained model and the remaining 100 set aside for evaluating the models' performance. Subsequently, text descriptions corresponding to the images are composed in the second step, a task accomplished by an experienced ergonomist. These text descriptions encompass an assessment of whether the workers depicted in the images are exposed to ergonomic risks, along with the factors contributing to these risks. The final step involves the construction of image-text pairs, wherein images are named in an incremental numerical format with a specific file extension (e.g., .jpg). Notably, a Python program was devised in this study to automatically integrate the text descriptions created by the ergonomist into an annotation file, denoted by the .json file extension.

Figure 1 shows an image sample from the finetuning dataset, and the ground truth text description of this image from the ergonomist is 'The image shows a worker wearing a white safety hat pouring cement on the road. The worker is exposed to awkward working posture because of the leaning forward posture, which is an ergonomic risk. The awkward posture may lead to the development of work-related musculoskeletal disorders.' Figure 2 shows an image sample from the testing dataset, and the ground truth text description of this image from the ergonomist is 'The image shows a worker standing in an awkward working posture on a black metal frame or a scaffold. The worker is exposed to ergonomic risks due to his overhead work posture, which may lead to work-related musculoskeletal disorders. The worker may be exposed to fall hazards, and it is recommended to have a safety harness on him.'

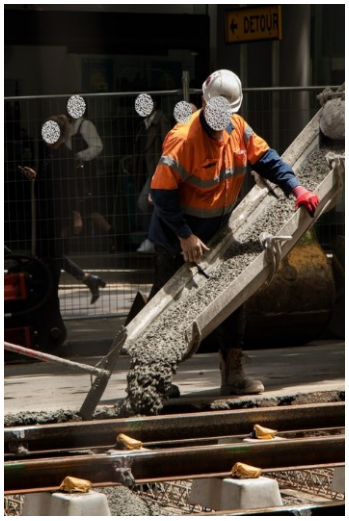

Figure 1. An image sample from the finetuning dataset.

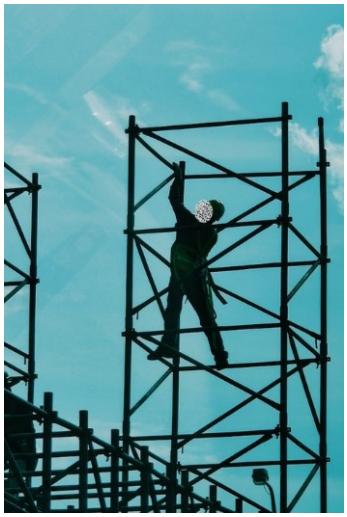

Figure 2. An image sample from the testing dataset.

2.2 Finetuning the pre-trained vision-language model

Following the initial phase, the pre-trained model underwent additional finetuning in the second phase to achieve coherent linguistic output. The finetuning phase involved the application of the Python script provided by MiniGPT-4, which was specifically designed for finetuning pre-trained models. The workstation's configuration for finetuning comprised an Intel Xeon Gold 6242 CPU, NVIDIA RTX A6000 GPU, and 128 GB of DDR4 memory. The workstation operated on the Ubuntu 22.04.1 LTS 64-bit operating system.

2.3 Testing the performance of the finetuned vision-language model

To validate the proposed approach in this study for identifying ergonomic risks using a vision-language model and to evaluate the performance of the finetuned models, the test dataset outlined in Section 3.1 was employed. Initially, the finetuned models were configured to serve as the vision-language models utilized by MiniGPT-4 for executing vision-language tasks. Subsequently, following the instructions provided by MiniGPT-4, the program was executed, and the finetuned models generated text descriptions corresponding to the images in the test dataset. The prompt used to generate text descriptions was 'Describe the workers and their postures in the image and tell me if they are exposed to ergonomic risks due to their postures?'

This study utilized a uniform prompt across the entire test dataset to maintain consistency in both the format and content of text descriptions across all test cases. Employing a uniform prompt for each test case also ensures fairness in the evaluations. Of greater importance, the ground truth text descriptions of the test dataset comprise descriptions of workers' actions and whether the workers are exposed to ergonomic risks based on these actions. Hence, the prompt is designed to focus explicitly on capturing descriptions of workers' actions and whether they are exposed to ergonomic risks. This uniform prompt aims to minimize unrelated factors when conducting a comparison between the ground truth text descriptions and the text descriptions generated by the finetuned model.

To quantitatively assess the enhancement attained by the model trained on the ergonomic risk dataset proposed in this study, this study additionally finetuned a model using the generic image-text pairs provided by MiniGPT-4. Following that, this study evaluated the performance of this model using the test dataset and the identical prompt. Subsequently, the accuracy of the two models was compared.

The prompt is framed as a yes-or-no question, inquiring whether the workers depicted in the images are subject to ergonomic risks. Considering that the objective of this study does not involve assessing or improving the accuracy of computer vision algorithms/models for human pose estimation, only the performance of whether the finetuned models can correctly identify ergonomic risks in the images was tested. The finetuned model's judgment is considered correct only when there is an agreement between ground truth and the finetuned model's judgment regarding whether workers in the images are exposed to ergonomic risks. The accuracy of both models on the test dataset can be expressed by equation 1.

$$Accuracy = \frac{number\ of\ correct\ cases}{number\ of\ all\ cases} * 100\% \quad (1)$$

3 Results

In contrast to the model finetuned with generic data from MiniGPT-4 (referred to as Model 1), the model finetuned with the dataset tailored for ergonomic risk identification among construction workers proposed in this study (referred to as Model 2) exhibited superior accuracy in recognizing ergonomic risks within the images of the test dataset. In terms of ergonomic risk identification, model 1 achieved an accuracy of 28% in generating text descriptions for the 100 test cases within the test dataset, whereas Model 2 achieved an accuracy of 81% for the same test cases, indicating a noteworthy accuracy improvement of 53%. Table 2 presents the accuracy obtained by testing Models 1 and 2 using the test dataset. Figure 3 illustrates an actual model 1 test case in the testing dataset featuring only one worker.

Figure 5 shows the result from model 2 for the same test case. The ground truth of this test case is 'The image shows a worker standing on a ladder using a power tool to fix pipes on the roof of a pavilion. He is exposed to safety hazards because he is standing on the ladder without wearing a safety harness. He is exposed to ergonomic risk because he is leaning his head forward for a long time, which can strain the neck and shoulders, leading to discomfort and potential neck pain.' Meanwhile, Figure 4 illustrates an actual model 1 test case in the test dataset featuring multiple workers. Figure 6 shows the result from model 2 for the same test case. The ground truth of this multi-worker test case is 'The image shows workers laying bricks. They are wearing hard hats and gloves. They are exposed to ergonomic risks because they are squatting and bending forward, and their squatting and bending forward postures may lead to the development of work-related musculoskeletal disorders.'

Table 2. Obtained accuracy by testing Models 1 and 2 using the test dataset.

Model	Accuracy (%)
Model finetuned with generic data	28
Model finetuned with our data	81

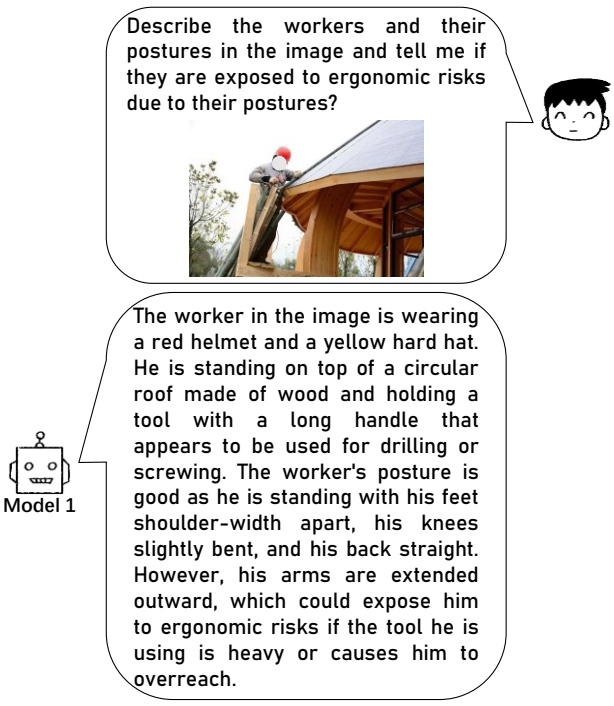

Figure 3. An example of model 1 generating a text description for an image containing a worker.

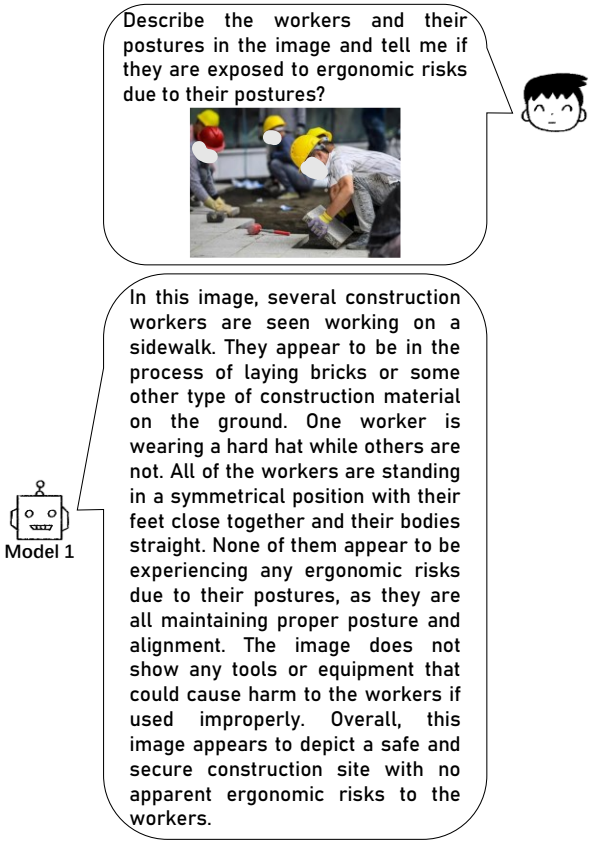

Figure 4. An example of model 1 generating text description for an image containing multiple workers.

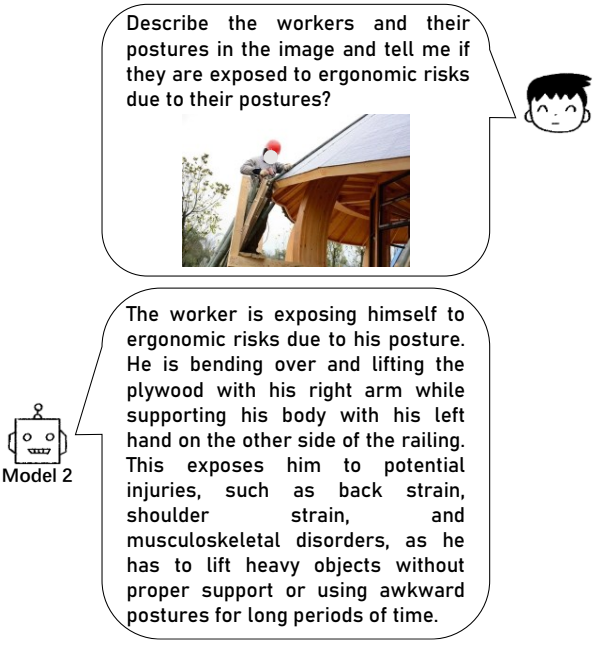

Figure 5. An example of model 2 generating text description for the same image in Figure 3.

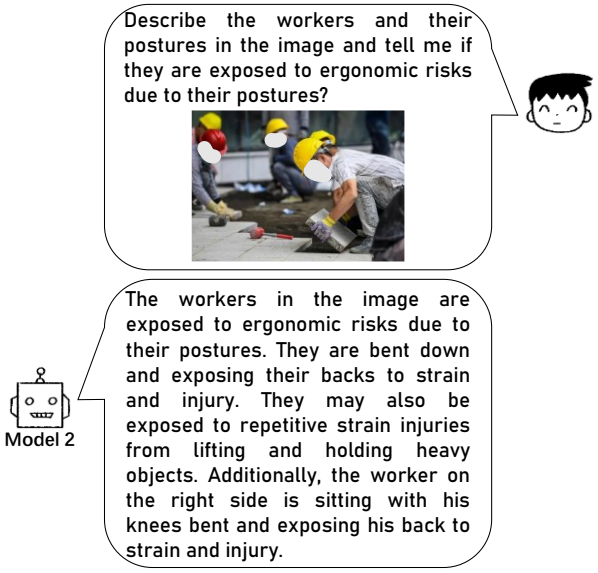

Figure 6. An example of model 2 generating text description for the same image in Figure 4.

4 Conclusions and future work

The dataset with a specific focus on identifying ergonomic risks among construction workers was created to finetune a GPT-based vision-language model. The proposed methodology showcased robust generalization capabilities when applied to the testing dataset derived from real-world scenarios. The model, which underwent finetuning using image-text pairs featuring ergonomic risk information, demonstrated an 81% accuracy in real-world test cases. This performance surpasses that of the model finetuned with generic image-text pairs lacking ergonomic risk information by 53%. These findings underscore the effectiveness of finetuning GPT-based vision-language models in achieving robust performance for the identification of ergonomic risks.

As the objective of this study is to explore the feasibility and ability of visual question answering related to ergonomic risk identification by finetuning vision-language models using data related to ergonomic risks, the correctness of the text descriptions generated by the models is entirely dependent on the correctness of the text descriptions for postures. The study's objective does not prioritize the precision of computer vision algorithms in recognizing the content of images. For instance, if the model identifies ergonomic risks for different reasons compared to the ground truth, it is considered a correct identification.

While computer vision algorithms have garnered significant attention for assessing ergonomic risks based on workers' actions in images, the unique and powerful capabilities of vision-language models in managing interactions between visual and textual elements have not been harnessed for ergonomic risk assessment. The validation experiments conducted in this study, which encompass scenarios featuring both multiple construction workers and a single construction worker, illustrate the efficacy of the proposed approach. This method, involving the finetuning of vision-language models with ergonomic-related data, demonstrates superior performance in ergonomic risk identification compared to models finetuned with generic image-text pairs data provided by MiniGPT-4. We aspire for the GPT-based vision-language approach presented in this study to play a role in identifying ergonomic risks for upcoming construction workers, thereby augmenting the well-being of workers and the safety of their work environments. Moreover, it is expected that this research will inspire further research of vision-language models for the identification of ergonomic risks and safety measures.

As interest in this field grows, the potential applications of data acquired by surveillance cameras at construction sites are expanding, moving beyond simple video storage and traditional computer vision tasks like object recognition. With large vision language models capable of extracting human-like language descriptions from surveillance data, they can aid or potentially supplant the role of ergonomists and safety experts in real-time video analysis and alerting workers or safety personnel about safety concerns on construction sites. Furthermore, individuals without specialized expertise can leverage this technology to produce injury or safety reports based on construction activities, leading to improved construction practices and streamlining the process of filing insurance claims.

5 Limitations

Despite the remarkable accuracy demonstrated by the proposed method, akin to the utilization of vision-language models in other fields, it has its limitations. These limitations revolve around issues such as language hallucination and insufficient perception capacities. As the vision-language model is constructed upon LLMs, its limitations in terms of language hallucination primarily result from inheriting the unreliable reasoning ability and hallucinating non-existent knowledge of LLMs. The inadequacy in perception capacities is predominantly linked to the limited visual perception capability of MiniGPT-4. Future research endeavors could mitigate language hallucination by further leveraging high-quality image-text pairs containing ergonomic information and refined LLMs. Addressing limited perception capacities may involve incorporating multiple layers in the projection layer, implementing a more robust visual perception model, and utilizing data that is more well-aligned.

In general, the mentioned limitations can be mitigated by integrating a wider range of diverse image-text pairs and refining the accuracy of textual descriptions detailing ergonomic risks depicted in the images. Expanding the pool of images and refining the accuracy of textual descriptions/annotations linked with these images has the potential to alleviate these limitations.

References

[1] European Agency for Safety and Health at Work. Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU. On-line: https://osha.europa.eu/en/publications/msds-facts-and-figures-overview-prevalence-costs-and-demographics-msds-europe, Accessed: 14/12/2023.

[2] U.S. Bureau of Labor Statistics. Injuries, Illnesses, and Fatalities. On-line: https://www.bls.gov/iif/factsheets/msds.htm, Accessed: 03/12/2023.

[3] David G. C. Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. *Occupational Medicine (Chic Ill)*, 55(3):190-9, 2005.

[4] MassirisFernández M., Fernández J. Á., Bajo J. M., and Delrieux C. A. Ergonomic risk assessment based on computer vision and machine learning. *Computers & Industrial Engineering,* 149:106816, 2020.

[5] Plantard P., Shum H. P. H., Le Pierres A.-S., and Multon F. Validation of an ergonomic assessment method using Kinect data in real workplace conditions. *Applied ergonomics,* 65:562-9, 2017.

[6] Vignais N., Bernard F., Touvenot G., and Sagot J.-C. Physical risk factors identification based on body sensor network combined to videotaping. *Applied ergonomics,* 65:410-7, 2017.

[7] Li X., Han S., Gul M., and Al-Hussein M. Automated Ergonomic Risk Assessment based on 3D Visualization. In *34th International Symposium on Automation and Robotics in Construction (ISARC 2017),* Vol. 34, Taipei, Taiwan, 2017.

[8] Guo S. Y., Ding L. Y., Luo H. B., and Jiang X. Y. A Big-Data-based platform of workers' behavior: Observations from the field. *Accident Analysis & Prevention,* 93:299-309, 2016.

[9] Fan C., Mei Q., Yang Q., and Li X. Computer-vision based rapid entire body analysis (REBA) estimation. In *Modular and Offsite Construction (MOC) Summit Proceedings*, pages 90–97, Edmonton, Canada, 2022.

[10] Jeong S. and Kook J. CREBAS: Computer-Based REBA Evaluation System for Wood Manufacturers Using MediaPipe. *Applied Sciences*, 13(2):938, 2023.

[11] Fang W., Ding L., Love P.E., Luo H., Li H., Pena-Mora F., Zhong B., and Zhou C. Computer vision applications in construction safety assurance. *Automation in Construction*, 110:103013, 2020.

[12] Barberi E., Chillemi M., Cucinotta F., Milardi D., Raffaele M., Salmeri F., and Sfravara F. Posture Interactive Self Evaluation Algorithm Based on Computer Vision. In *International Joint Conference on Mechanics, Design Engineering & Advanced Manufacturing*, pages 1516-1526, 2022.

[13] Nayak G. K. and Kim E. Development of a fully automated RULA assessment system based on computer vision. *International Journal of Industrial Ergonomics,* 86:103218, 2021.

[14] Seo J., Yin K., and Lee S. Automated Postural Ergonomic Assessment Using a Computer Vision-Based Posture Classification. In *Construction Research Congress 2016,* pages 809–818, Reston, United States, 2016.

[15] Zhu D., Chen J., Shen X., Li X., and Elhoseiny M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint*, arXiv:2304.10592, 2023.

[16] Driessen T., Dodou D., Bazilinskyy P., and de Winter J. Putting ChatGPT Vision (GPT-4V) to the test: Risk perception in traffic images. 2023.

[17] Yang Z., Li L., Lin K., Wang J., Lin CC., Liu Z., and Wang L. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint,* arXiv:2309.17421, 2023.

[18] OpenAI. ChatGPT can now see, hear, and speak. On-line: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak, Accessed: 15/12/2023.

[19] Zheng J. and Fischer M. Dynamic prompt-based virtual assistant framework for BIM information search. *Automation in Construction*, 1;155:105067, 2023.

[20] You H., Ye Y., Zhou T., Zhu Q., and Du J. Robot-Enabled Construction Assembly with Automated Sequence Planning Based on ChatGPT: RoboGPT. *Buildings*, 13(7), 2023.

[21] Prieto S. A., Mengiste E. T., and García de Soto B. Investigating the Use of ChatGPT for the Scheduling of Construction Projects. *Buildings*, 13(4):857, 2023.

[22] Chen H., Hou L., Wu S., Zhang G., Zou Y., Moon S., and Bhuiyan M. Augmented reality, deep learning and vision-language query system for construction worker safety. *Automation in Construction*, 157:105158, 2024.

[23] See A., Pappu A., Saxena R., Yerukola A., and Manning CD. Do massively pretrained language models make better storytellers?. *arXiv preprint,* arXiv:1909.10705, 2019.