REBA-VLM: Enhanced risk mitigation in construction via Pose guided visual instruction tuning

Mohammad Daniyalur Rahman[1], Likhith Kumara[1], Sivakumar S. K.[2] and Nikhil Bugalia[2]

[1]Tata Consultancy Services
[2]Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, India
mddaniyalur.rahman@tcs.com, likhith.kumara@tcs.com, ce24d040@smail.iitm.ac.in, nbugalia@civil.iitm.ac.in

Abstract –

Ergonomic risk assessment helps improve safety and reduce injuries in construction work. The Rapid Entire Body Assessment (REBA) method is widely used, but it is usually done manually. This makes it slow, subjective, and hard to apply, especially on large construction sites. Though vision-based methods do exist, many tend to yield inaccurate assessments, especially incorrect REBA scores, and do not explain why the posture causes the risk. This work proposes an automated, human-understandable REBA assessment method based on a vision–language model. The system takes a single image of workers and produces REBA score, problem identification, and actionable suggestions in simple language. We fine-tuned an existing vision-language model (VLM), namely Qwen2-VL-2B, using Low-Rank Adaptation (LoRA) to efficiently utilize limited computational resources. During dataset preparation, we chose REBAPose (a vision-based method built on new annotations of three keypoints to get more accurate REBA score), and applied on Moving Objects in Construction Sites (MOCS) Dataset to obtain REBA scores and joint angles. Gemini 2.5 Flash generated clean, consistent textual labels and explanations from posture-related cues to prepare the dataset. Then, we created a fine-tuned Qwen VLM model using that dataset. We tested the method on unseen images and compared it with common baselines. We reported score agreements and whether explanation points to the same body parts that mainly drive the REBA score. As dataset is trained on MOCS, our method is robust to construction scenes with cluttered backgrounds and different camera views, making it useful in real-world conditions and helping site engineers.

Keywords –

Ergonomics, Musculoskeletal Disorder, Construction Safety, REBA, vision-language model, Qwen2-VL-2B, LoRA.

1 Introduction

Construction work often involves awkward postures, heavy loads, and repetitive movements. These factors increase the risk of work-related musculoskeletal disorders, which are a major cause of long-term injury and reduced productivity in the construction industry. To reduce these risks, safety engineers commonly use ergonomic assessment methods to evaluate worker posture during different tasks. The Rapid Entire Body Assessment (REBA) [1] method is one of the most widely used tools for this purpose. REBA evaluates the posture of the neck, trunk, legs, and upper limbs, along with load and activity factors, to produce a risk score. The REBA score ranges from 1 to 15, with scores indicating risk levels from low (1-4) to very high (11-15), requiring immediate action. A score of 1 signifies minimal risk, while scores of 5-7, 8-10, and above 10 suggest increasing levels of concern, prompting investigation and necessary ergonomic changes to prevent musculoskeletal disorders (MSDs). While REBA is practical and well accepted, it is usually performed by trained experts through manual observation. This process is time-consuming, depends on the observer's experience, and is difficult to apply continuously on busy construction sites. In recent years, computer vision techniques have been explored to automate ergonomic assessment [2–5]. Many existing approaches rely on pose estimation models to extract body joint positions [6] and then apply rule-based calculations to obtain REBA or similar scores. These methods can reduce manual effort, but they usually keep the full pipeline in numeric form. Vision–language models offer a different way to support ergonomic assessment [7]. Instead of only working with joint coordinates and rules, they can look at the worker image and generate text that follows how safety staff think and talk about posture. This allows linking posture, REBA concepts, and simple explanations in a single step. A model of this kind can still respect the structure of REBA, but at the same time describe in plain language which body regions are under strain and what simple

changes could reduce the risk.

This paper proposes an automated and explainable REBA assessment framework based on such a vision-language model. Specifically, we used REBAPose, a novel pose-based REBA scoring framework developed by two of the authors which is built on a new annotated dataset containing 149,813 persons, to get accurate REBA scores and joint angles as inputs to train our model. The system analyzes a single image of a construction worker and outputs REBA body-part scores [9], the overall risk level, and a short explanation that is aligned with REBA rules. The goal is to support safety engineers and site managers with a practical tool that improves consistency, reduces manual scoring effort, and can later be extended toward real-time and edge-based deployment in construction environments.

2 Related Work

Ergonomic risk assessment for construction workers has traditionally relied on observation-based tools such as REBA, Rapid Upper Limb assessment (RULA) and Ovaco Working Posture Analysis system (OWAS) [8]. These tools are widely accepted in practice, but they require trained experts, are time-consuming, and are hard to apply often on busy sites. Because of this, many recent studies try to automate parts of the process using movement data or images.

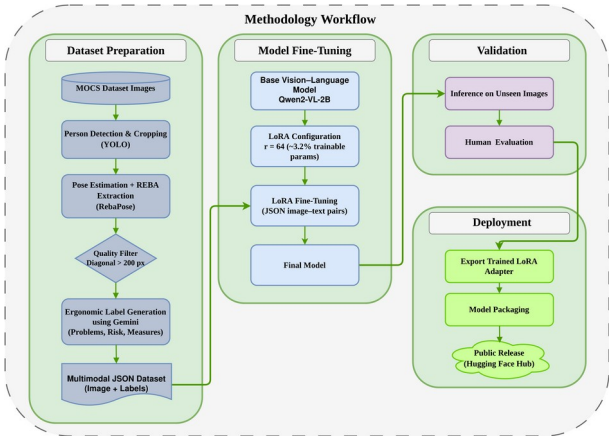
REBAPose [9] is a computer-vision-based framework that aims to build a fully automatic REBA risk calculator. The focus of that work is on improving the pose information needed for REBA by adding three new key-point annotations to the Common Objects in Context (COCO) dataset and training pose models on this extended set. The authors show that these new points can be detected with better accuracy and that the resulting joint angles make REBA scoring more reliable than other models which use only standard COCO joints. However, REBAPose only produces numeric scores and does not generate text-based explanation of the risk or suggestions for corrective actions for the safety managers.

A recent study [10] looks at question answering on building regulations. They build a retrieval-augmented system where a large language model is fine-tuned in several stages using Low-Rank Adaptation (LoRA) [11] and a “capability distillation” scheme within an open-book exam framework. Each stage handles a specific sub-task such as document retrieval, reasoning or answer writing, and the output of one stage becomes the input to the next. This design helps engineers cope with the “document skipping” problem across many linked regulations and makes the reasoning path easier to trace. Their work shows that frontier models need domain-specific fine-tuning and that traceability of the reasoning matters. However, their system works only with regulated text and legal questions. It does not handle worker images, does not estimate ergonomic risk, and does not link the output to a standard ergonomic scale such as REBA, which is the focus of our study. A second line of work [12] uses images directly and proposes a vision-language approach to assist in identifying ergonomic risks for workers. Their system uses construction images and text-based supervision to generate ergonomic descriptions and risk-related responses, showing that domain-specific training can improve performance over models trained only on generic image-text data. This demonstrates the potential of vision-language methods for supporting ergonomic screening in construction. However, the output remains free-form and is not tied to a structured REBA assessment that can be directly recorded in site documents.

ErgoChat [12] goes further by framing ergonomic risk assessment as a visual query system. The system supports both image captioning and visual question answering about workers’ postural risks, and the authors also introduce a dedicated dataset for training and testing. Users can ask their own questions and receive text answers about posture and risk level. Their results show strong performance, especially for question answering, but the outputs are still free-form text rather than a fixed ergonomic scoring format.

3 Methodology

This section describes the end-to-end workflow of our method, from raw images to a deployed model. Figure 1 shows an overview with four main stages: dataset preparation, model training, validation and deployment. The idea is to take a single construction image that contains one or more workers and, for each worker, produce a REBA-based assessment. For every worker, the system returns the REBA score, the risk level and a short explanation written in simple language. To reach this goal, we first prepare a labelled dataset from construction images, then fine-tune a compact vision-language model on these labels using LoRA, and finally check the outputs manually before packaging the model for release. The pipeline is designed to be modular, allowing individual stages to be updated without affecting the overall system. It also supports processing multiple workers within the same image in a consistent and scalable manner.

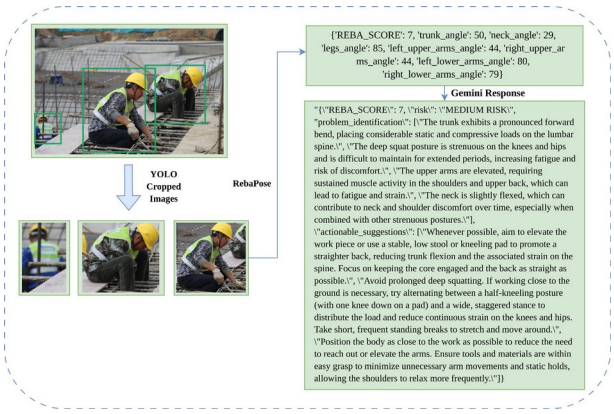


Methodology

Figure 1. Overview of the proposed REBA assessment workflow, from dataset preparation (worker crops and REBA labels) through LoRA fine-tuning of Qwen2-VL-2B, human validation on unseen images, and packaging of the trained adapter for deployment.

3.1 Dataset Preparation

We started from the MOCS construction motion dataset [13], which contains site images with multiple workers. Since REBA is defined at worker level, each person is processed separately. We used 6,872 filtered images for training. An example of this worker-level preparation for one image is shown in Figure 2.



Example Transformation

Figure 2. Example of dataset preparation: YOLO first detects and crops each worker, REBAPose outputs the REBA score and joint angles, and Gemini converts these posture cues into structured text labels with risk, problem identification, and actionable suggestions.

3.1.1 Person cropping (YOLO)

Since REBA is defined for individual workers, the first step is to isolate each person. A You Only Look Once (YOLO)-based detector [14] is used to find workers in the image. For every detected person, we crop a tight region around the body. This makes the posture clearer and ensures that later labels always refer to a single worker.

3.1.2 REBA labels and posture cues (REBAPose)

Each cropped image is then passed to REBAPose [9], which estimates body joint positions, joint angles and a REBA score for that posture. This gives us both numeric posture cues and an initial REBA label. Very small crops are removed: if the diagonal of the crop is less than 200 pixels, details of the posture cannot be seen well and joint estimates are unreliable, so these samples are discarded.

3.1.3 Text labels for risk and guidance (Gemini 2.5 Flash)

Gemini 2.5 Flash is used only in this offline labeling stage. Given the REBA score and joint angles from REBAPose, we prompt Gemini to write short, structured text in three parts:

1. Risk level / risk summary.
2. Problem identification (which body parts and posture issues increase the risk).
3. Suggested measures (simple corrective actions).

3.1.4 JSON creation

For training, each sample is stored as a JSON record linking the cropped image to REBA score, posture cues, and the three text fields we discussed in the last section above. This gives us a dataset where each worker image has both numeric and human-readable labels that follow REBA concepts.

3.2 Model Training

We used Qwen2-VL-2B [15] as the base model and adapted it to the REBA task using the 6,872 filtered training samples described above. To keep the number of trainable parameters small, we used LoRA [11]. In practice, LoRA adds small low-rank adapter matrices to selected projection layers of Qwen2-VL-2B, including the attention and feed-forward blocks. The original weights of these layers stay fixed, and only the LoRA adapter weights are updated during training. The adapter has rank 64, which means that roughly 3.2% of the parameters are updated while more than 96% of the model is reused as it is. This keeps memory and computation cost low and also makes it easy to store

and share the task-specific part of the model as a small adapter file.

During training, the model received a worker crop as input along with a fixed text prompt asking it to assess the ergonomic risk of the worker in the image. The model was trained to generate an output that follows the same structure as the JSON labels. The target text contained the REBA score, the risk level, a short problem description with the affected body parts, and suggested corrective measures. The training used a standard next-token prediction loss, where the model learned to produce each word of the output based on the image and the words generated so far. In this way, the model learned to connect what it sees in the image with REBA body segments and scoring rules, and to express the result in clear language.

We trained the model for 4 epochs with a batch size of 8 and a learning rate of 5e-6, using the default AdamW optimizer. The training was done on a 24GB NVIDIA H100. The model we obtained after this step was the main model of the paper: a compact vision--language model that maps a single worker image to a REBA-style assessment, with the task-specific behaviour captured mainly in the LoRA adapter.

3.3 Human Validation

After training, we tested the model on images that were never used during training or label cleaning. For each worker in these images, the model produced a REBA score and explanation. We then manually reviewed a subset of these cases. We checked whether the predicted score and risk level are reasonable for the shown posture and whether the explanation correctly points to the main body regions under strain, such as trunk, neck, or legs.

3.4 Evaluation Methodology

After training the model, we planned to evaluate it on images that were not used during training or label preparation. For these images, we applied REBA scores and joint angles from REBAPose. Then we generated outputs from three models (GPT-4o [16], Gemini 2.5-Flash [17] and Qwen-3 Max [18]) by giving them the same images and asking them to assess ergonomic risk. For each image, we therefore had four outputs to confirm or discard the findings:

- Our model REBA-VLM (Qwen2-VL-2B with LoRA).
- GPT-4o.
- Gemini.
- Qwen-3 Max.

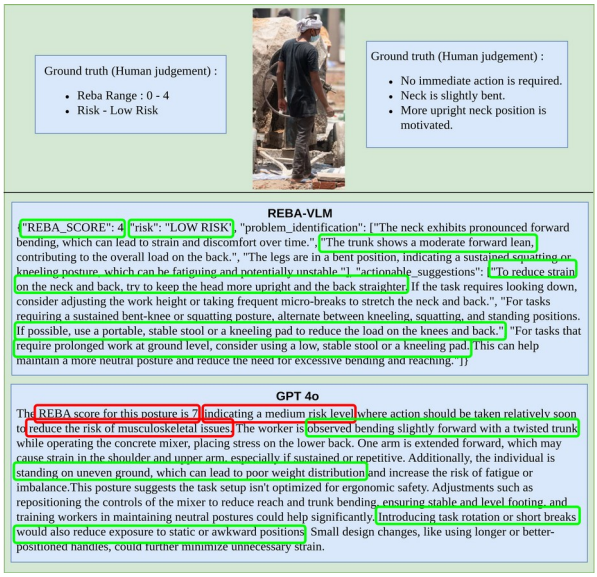
3.5 Deployment

For deployment, we export the trained LoRA adapter and combine it with the unchanged Qwen2-VL-2B base model in a simple inference script. The script takes a worker crop and returns the same structured output used during training. The adapter and configuration are uploaded to the Hugging Face Hub so that others can load the model and reproduce the inference setup. This makes it easier to move toward edge devices later, as the adapter is small and requires no runtime external services. The inference interface for the model is publicly available at [https://huggingface.co/MdDaniyalurRahman892/Reba VLM].

4 Results

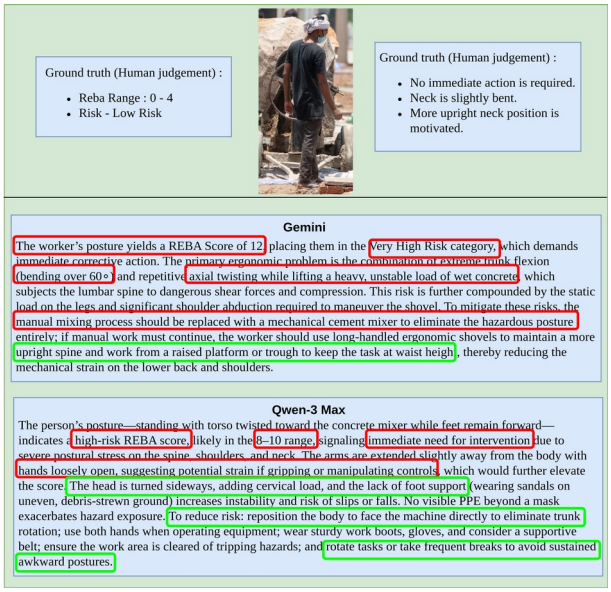
To evaluate the proposed REBA-VLM model, we tested it on unseen construction images and compared its outputs with three frontier models: GPT-4o, Gemini 2.5-Flash, and Qwen-3 Max. We reviewed 30 images and checked each output for correctness of the REBA score, risk level, identification of affected body parts, and usefulness of the suggestions. We presented detailed examples below, followed by a summary in Table 1.

4.1 Comparison with Frontier Models



Output Comparison - Part I

Figure 3. Comparison of REBA-VLM and GPT-4o outputs for a low-risk standing posture. Green highlights show correct observations, red highlights show incorrect or unsupported claims.



Output Comparison - Part II

Figure 4. Comparison of Gemini and Qwen-3 Max outputs for the same low-risk posture. Red highlights show incorrect claims including overestimated scores and assumptions not visible in the image.

Figures 3 and 4 show the outputs of all four models for the same worker image. The ground truth, based on manual assessment, is a REBA score in the range of 0–4 with low risk. The worker is mostly standing upright near a concrete mixer, and the only visible issue is that the neck is slightly bent. No immediate action is required.

Our REBA-VLM model predicts a REBA score of 4 and labels the risk as low. Its explanation correctly points to mild neck bending and forward lean of the trunk, and suggests keeping the head more upright and adjusting the work height. These observations match the ground truth and are written in short, structured text that can be used directly in site records. The model also avoids overestimating the risk, unlike other models which assign higher scores for the same posture. This shows that the model can capture small posture deviations without exaggerating their impact. Overall, the prediction remains consistent with the actual working condition shown in the image.

GPT-4o, on the other hand, assigns a REBA score of 7 and calls it a medium risk level, suggesting that action should be taken relatively soon. It describes the worker as having a "twisted trunk" and "standing on uneven ground" with "poor weight distribution." It also suggests "task rotation or short breaks," which are not relevant to this simple standing posture. These observations do not match what is visible in the image - the worker is standing upright with no trunk twist, and the footing appears stable. The highlighted sections in green and red in Figure 3 show where GPT-4o's output does not align with the actual posture.

Gemini 2.5 Flash gives the most extreme result, assigning a REBA score of 12 and labeling it as very high risk requiring immediate corrective action. It describes "extreme trunk flexion (bending over 60°)," "heavy, unstable load of wet concrete," and "axial twisting" - none of which are visible in the image. It even suggests replacing the manual process with a mechanical mixer and working from a raised platform. These suggestions assume a completely different posture and task than what the image shows.

As highlighted in Figure 4, nearly all of Gemini's output is incorrect for this case. Qwen-3 Max also overestimates the risk, placing the REBA score in the 8-10 range and calling for immediate intervention. It mentions "cervical load," "wearing sandals on uneven debris-strewn ground," and "hands loosely open" - details that are either not visible or not relevant. It also brings in personal protective equipment (PPE) concerns, which are outside the scope of REBA. Like the other frontier models, it does not follow the REBA structure and adds assumptions that go beyond what the image shows.

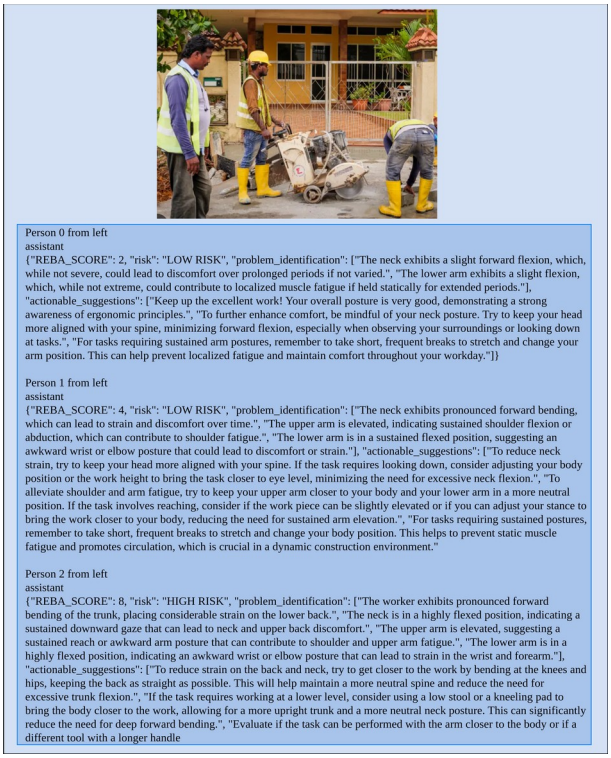
This comparison shows a clear pattern: the frontier models tend to overestimate risk, add details not visible in the image, add irrelevant information unrelated to posture, and produce long unstructured responses. Our model stays within the REBA framework and gives a short, focused output that matches the actual posture. In addition, it avoids speculative assumptions and maintains strict alignment with observable evidence. This results in more consistent scoring, clearer traceability, and a level of interpretability that is essential for ergonomic assessment workflows. Ultimately, the structured discipline of our approach helps ensure reliability while reducing noise in downstream decision-making.

In practice, this means the output from REBA-VLM can be used as-is for documentation or decision-making on site, while the outputs from the other models would need manual review and correction before they could be trusted or acted upon. For safety managers handling dozens of assessments daily, this difference matters - a model that requires constant verification offers little

advantage over doing the assessment manually in the first place.

4.2 Outputs Across Different Risk Levels

To show that the model handles different risk levels correctly, Figure 5 presents the outputs for three workers visible in a single construction site image.



REBA-VLM Outputs for Different Risk Levels.

Figure 5. REBA-VLM outputs for three workers in a single construction site image, showing how the model distinguishes between low-risk and high-risk postures through differences in REBA score, affected body parts, and suggested corrective actions.

Person 0 (from left) is standing upright with good overall posture. The model assigns a REBA score of 2 with low risk. It notes a slight forward neck flexion and a minor arm flexion, and suggests being mindful of neck posture and taking short breaks. The suggestions match the posture - the worker is in a safe position and only small improvements are needed. Person 1 (from left) is also standing but with a slightly more noticeable forward lean. The model gives a REBA score of 4, still low risk. It identifies neck forward bending, elevated upper arm, and a sustained lower arm position. The suggestions include adjusting work height and keeping the upper arm closer to the body. These observations correctly reflect the small differences between this worker and Person 0. Person 2 (from left) is clearly bending forward, which is a higher-risk posture. The model assigns a REBA score of 8 and labels it as high risk. It identifies forward bending of the trunk, a highly flexed neck with a downward gaze, elevated upper arm, and an awkward wrist position. The suggestions include bending at the knees instead of the back, using a low stool or kneeling pad, and bringing the work closer to the body. These match the visible posture and follow how a safety engineer would assess this worker.

This example shows that the model can correctly distinguish between low-risk and high-risk postures within the same image and adjust its score, problem identification, and suggestions accordingly.

4.3 Qualitative Comparison with Frontier Models

We went through 30 construction-related images and analyzed the outputs of all four models. Table 1 summarizes the pattern we observed across these images. For each model, we checked whether the output correctly produces the result.

Table 1. Comparison of key elements in the ergonomic reports generated by each model (✓ consistently correct across cases, ± partially correct or inconsistent, × largely incorrect or not relevant. Based on manual review of 30 images.).

Model	REBA-VLM	ChatGPT 4o	Gemini	Qwen 3 Max
REBA Score	✓	±	±	±
Risk Level	✓	✘	✓	✘
Body Parts Affected	✓	✓	✓	✓
Focused Suggestions	✓	✓	✓	✘

Across these images, REBA-VLM consistently produced correct REBA scores and risk levels, pointed to the right body parts, and gave useful suggestions. The frontier models sometimes identified the affected body

parts correctly, but their REBA scores and risk levels were often too high, and their suggestions were either too general or based on details not visible in the image. This pattern was consistent with what we showed in the detailed examples in Figures 3, 4, and 5.

In instances where our model did not perform well, the disagreement was usually small (for example, between two neighboring REBA risk levels) and could be traced back to difficult viewing conditions, such as partial occlusion or very small worker size in the original image.

5 Discussion

These qualitative results suggest that a small, task-specific model that is trained directly on REBA-style outputs can give more useful feedback than frontier models, even if those models are stronger in a broad sense. Because our model is forced to produce a fixed structure and to stay within the REBA framework, it tends to stay on topic and avoids talking about factors that are not visible in the image. This aligns with the method's goal of helping safety engineers quickly document posture-related risk consistently.

One reason the frontier models tend to overestimate risk is that they are not trained on REBA-specific data. When asked about ergonomic risk, they fall back on broad safety knowledge and often assume the worst case - for example, assuming heavy loads or unstable footing even when the image shows a worker standing upright with no visible load. They also tend to produce long responses that mix posture assessment with unrelated advice about personal protective equipment or general hazard awareness. In contrast, our model stays focused on the body parts and posture factors that REBA actually measures, because it was trained on outputs that follow that structure.

The use of LoRA also has practical benefits beyond training efficiency. Since the task-specific knowledge is captured in a small adapter file (roughly 3.2% of the total parameters), the adapter can be easily shared, versioned, and updated without retraining the full model. This also makes it easier to move toward edge deployment, as the base Qwen2-VL-2B model is already compact and the adapter adds very little overhead. The training dataset, built from MOCS construction images [13], also plays an important role. Since MOCS contains real construction site images with cluttered backgrounds, varying camera angles, and multiple workers in a single frame, the model learned to handle the visual conditions that are common on actual sites. This is different from models trained on clean, controlled images where the worker is clearly visible and isolated.

In cases where our model did not perform well, the errors were usually small - for example, a disagreement between two neighboring risk levels - and could be traced back to difficult viewing conditions such as partial occlusion or very small worker size in the original image. These are cases where even a human observer would find it hard to assess the posture accurately from a single image. Despite these limitations, the study shows that it is feasible to combine REBA structure, image understanding, and plain-language explanation in a single model. In the reviewed examples, the method produced structured initial assessments that were generally aligned with REBA concepts. However, its practical usefulness for supporting safety engineers should be validated through larger-scale studies and direct comparison with expert assessment.

6 Conclusion and Future Work

This work presented a method for automated REBA-based ergonomic assessment of construction workers using a single image. We prepared a worker-level dataset by cropping individuals from MOCS images, using REBAPose to obtain REBA scores and joint angles, and using Gemini 2.5 flash to write short, structured explanations in a fixed format. We then fine-tuned Qwen2-VL-2B with a LoRA adapter so that, given a worker crop, it outputs REBA scores and a short explanation that names the main risky body parts and suggests simple corrective actions.

This study demonstrates the feasibility of a compact vision-language model for generating REBA-style ergonomic assessments from single construction images. The main contribution is not a claim of expert-level performance, but a proof-of-concept that links image understanding, structured REBA outputs, and short natural-language explanations in one model. At the same time, the current evaluation is limited in scale and qualitative in nature, so broader quantitative validation against expert assessment is needed before practical use can be established.

There are several directions for future work. First, the dataset can be expanded to include more trades, tools and viewpoints, and to cover a wider range of REBA scores. Second, the current model works with static images; many construction tasks involve movement over time, so extending the approach to short video clips would enable the analysis of repetitive actions and dynamic loads. Third, adding task context

such as task type, duration and load weight would make the assessment more complete, since REBA is only one part of ergonomic risk. Finally, we plan to study deployment on edge devices and integration with existing safety workflows, for example, as a support tool for safety walks or as part of a monitoring system that helps site managers focus their attention on the most critical postures.

References

[1] S. Hignett and L. McAtamney. Rapid entire body assessment (REBA). Applied Ergonomics, 31(2):201–205, 2000. doi: 10.1016/S0003-6870(99)00039-3.

[2] X. Yan, H. Li, C. Wang, J. Seo, H. Zhang, and H. Wang. Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. Advanced Engineering Informatics, 34:152–163, 2017. doi: 10.1016/j.aei.2017.11.001.

[3] X. Luo, H. Li, F. Wang, Y. Fang, and M. Cao. Automated postural ergonomic risk assessment using vision-based posture classification. Automation in Construction, 128:103725, 2021. doi: 10.1016/j.autcon.2021.103725.

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

[5] X. Luo, H. Li, Y. Fang, and M. Cao. Data-driven ergonomic assessment of construction workers. Automation in Construction, 165:105561, 2024. doi: 10.1016/j.autcon.2024.105561.

[6] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1):172–186, 2021. doi: 10.1109/TPAMI.2019.2929257.

[7] J. Liu, Q. Zhang, and J. Han. Vision–language models for human activity understanding: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. doi: 10.1109/TPAMI.2024.3369699.

[8] D. Kee. Systematic comparison of OWAS, RULA, and REBA based on a literature review. International Journal of Environmental Research and Public Health, 19(1), 2022. doi: 10.3390/ijerph19010595.

[9] K. S. Sivakumar, N. Bugalia, and B. Raphael. REBAPose – A computer vision based musculoskeletal disorder risk assessment framework. In Proceedings of the 41st International Symposium on Automation and Robotics in Construction (ISARC 2024), pages 607–614, Lille, France, 2024. doi: 10.22260/ISARC2024/0079.

[10] Z. Jiang, G. Chen, and Z. Xu. Building regulation question-answering system using retrieval-augmented generation with dual-stage fine-tuned large language model. Advanced Engineering Informatics, 69:104089, 2026. doi: 10.1016/j.aei.2025.104089.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), 2022. doi: 10.48550/arXiv.2106.09685.

[12] C. Fan, Q. Mei, X. Wang, and X. Li. ErgoChat – a visual query system for the ergonomic risk assessment of construction workers. arXiv preprint arXiv:2412.19954, 2024. doi: 10.48550/arXiv.2412.19954.

[13] X. An, Z. Li, Z. Liu, C. Wang, P. Li, and Z. Li. Dataset and benchmark for detecting moving objects in construction sites. Automation in Construction, 122:103482, 2021. doi: 10.1016/j.autcon.2020.103482.

[14] J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. doi: 10.48550/arXiv.1804.02767.

[15] Qwen Team. Qwen2-VL model card. On-line: https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct, Accessed: 18/12/2025.

[16] OpenAI. GPT-4o system card. arXiv preprint arXiv:2410.21276, 2024. doi: 10.48550/arXiv.2410.21276.

[17] Google DeepMind. Gemini 2.5 Flash model card. On-line: https://ai.google.dev/gemini-api/docs/models, Accessed: 08/01/2026.

[18] Qwen Team. Qwen3-Max model card. On-line: https://qwen.ai/blog?id=qwen3-max, Accessed: 08/01/2026.