

Skeleton-enhanced vision-language framework for intent recognition of construction workers

Wenting Mo¹, Fanfan Meng¹, Mi Pan¹, and Mirosław J. Skibniewski²

¹Department of Civil and Environmental Engineering, University of Macau, Macao SAR, China

²Department of Civil & Environmental Engineering, University of Maryland, College Park, 20742, MD, USA

wentingmo@um.edu.mo, yc47936@um.edu.mo, mipan@um.edu.mo, mirek@umd.edu

Abstract -

Human-robot collaboration (HRC) in construction requires robots to understand worker intent and predict subsequent actions for proactive assistance. However, existing methods are primarily limited by closed-vocabulary classification, hindering their adaptability to diverse construction tasks. These approaches are also easily compromised by site environmental factors like poor lighting and severe occlusions. This paper proposes a skeleton-enhanced vision-language framework for intent recognition and action prediction of HRC in construction. The pipeline analyzes posture and environment for open-vocabulary intent recognition and top-3 next-action prediction, integrates with a teacher-student instruction-tuning strategy, and uses Gemini 3 Pro-generated annotations with human-in-the-loop validation to fine-tune a deployable Qwen3-VL-8B model via full-parameter supervised fine-tuning (SFT). Experiments on a dataset of 1,819 expert-verified construction images demonstrate that the proposed framework achieves 71.43% intent recognition accuracy, a 28.03% improvement over the zero-shot baseline. These findings confirm the effectiveness of skeleton enhancement and structured reasoning in improving the understanding of construction activities.

Keywords -

Vision-Language Model; Construction; Human-Robot Collaboration; Skeleton-Enhanced Visual Representation

1 Introduction

Human-robot collaboration (HRC) in construction is attracting significant attention due to its potential to assist workers in hazardous or physically demanding tasks. Compared with full automation, collaborative assistance is often preferred in construction as workflows are highly adaptive and non-linear [1]. However, current HRC systems still lack autonomous decision-making capabilities in complex environments [2]. A critical bottleneck lies in real-time perception for worker intent recognition and proactive action prediction. While vision-based human activity recognition has achieved promising performance in controlled settings, its effectiveness degrades at construc-

tion sites that suffer from severe illumination variations, frequent occlusions, and dynamic environmental contexts. Human skeleton information offers robustness to environmental noise [3], yet its potential to enhance semantic visual understanding in construction remains unexplored.

Leveraging the cross-modal alignment capabilities of vision-language models (VLMs) presents a viable path for interpreting such complex human behaviors. Recent VLMs have shown promise in behavioral intent reasoning [4]. Nevertheless, several critical obstacles remain: (1) existing approaches mainly rely on closed-set classification with limited categories and fail to capture the open-vocabulary nature and causal intent-to-action reasoning of complex tasks; (2) conventional skeleton fusion strategies require architectural modifications that hinder direct reuse of pretrained encoders and sacrifice spatial correspondence between pose and visual context; (3) in the low-resource construction domain, the scarcity of high-quality Chain-of-Thought (CoT) annotations constrains the reasoning performance of lightweight models.

To overcome these limitations, this paper proposes a skeleton-enhanced vision-language framework. As a simple yet efficient way to enhance visual representation, 2D skeletal information is overlaid onto original images to serve as the model's visual input. The task is formulated as two-stage reasoning: stage 1 generates natural language descriptions of current activities, and stage 2 predicts subsequent actions with workflow rationales. CoT prompts guide the model through structured reasoning steps from observations to predictions. Furthermore, a lightweight student model is fine-tuned on expert-verified, CoT-augmented data from a teacher model via full-parameter supervised fine-tuning (SFT). The practical significance of this work lies in the framework itself: by coupling domain-specific geometric guidance with reasoning logic in a Teacher-Student paradigm, it enables edge-friendly models to explicitly produce a reliable rationale of *why* a worker is moving, shifting from bounded classification labels to proactive human-robot collaboration. The main contributions of this work are summarized in four aspects. (1) An intent-guided action prediction task for construction scenarios is proposed, formulated

as two-stage reasoning with open-vocabulary generation to enable explicit intent-to-action causal modeling and CoT-based interpretability. (2) A skeleton overlay strategy achieves pose-context fusion without architectural modification, preserving pretrained generalization while enabling geometry-guided attention. (3) A human-in-the-loop data annotation pipeline combining teacher generation and expert verification is developed to acquire high-quality CoT annotations under resource constraints. (4) A dataset with 1,819 expert-verified samples is provided, covering diverse construction scenarios such as rebar tying, concrete pouring, and material handling.

2 Related Work

2.1 Intent Recognition in Construction

In construction, activity recognition has been studied using sensor-based and vision-based methods [5]. Although wearable approaches provide accurate motion data, they are intrusive and have limited capability to support the real-time contextual understanding required for human-robot collaboration [6]. Worker activity recognition further involves subtle motion patterns and strong task-specific semantics, requiring more advanced model representations. Prior studies have explored multimodal fusion of wearable and visual data to improve recognition accuracy, indicating that single-modality observations are insufficient [7]. Beyond motion cues, incorporating construction context is essential for distinguishing task-specific activities. Existing semantic frameworks, such as the visual attention model proposed by Zhong et al. [8], can transform visual scenes into natural language descriptions but still formulate activity recognition as a closed-set classification problem. This formulation constrains their ability to handle long-tail activities and to link recognition with prediction, as they lack interpretable reasoning to explain intent-action causality [9]. As a result, open-vocabulary frameworks that explicitly model intent-to-action causality with interpretable reasoning remain underexplored in construction human-robot collaboration.

2.2 Skeleton-Enhanced Visual Representation

VLMs demonstrate powerful multimodal reasoning capabilities through large-scale pretraining, offering new opportunities for the comprehension of construction activities. However, real-world site environments pose significant difficulties for pure RGB-based visual recognition, including small targets, overlapping occlusions[10], and multi-viewpoint variations[11]. Recent studies show that skeleton-based methods exhibit remarkable robustness under occlusion conditions[12]. Nevertheless, existing skeleton-visual fusion approaches typically employ dual-stream architectures that separately encode skeleton

and RGB features before late fusion. This separate encoding causes spatial decoupling, where the pixel-level spatial correspondence between skeletal keypoints and visual pixels is disrupted (e.g., the hand skeleton cannot align with tool pixels), weakening geometry-guided attention. Moreover, dual-stream architectures hinder the utilization of VLMs' pretrained visual encoders, increasing training costs[13]. An efficient input-level overlay strategy is needed that seamlessly integrates skeletal geometric guidance while preserving VLM native visual encoding capabilities.

2.3 Reasoning Adaptation in Low-resource Scenarios

Intent recognition in construction environments is inherently complex, including worker postures, tool-object interactions, and site contexts. Under low-resource conditions in construction environments, limited sensing quality, computational capacity, and annotated data further hinder reliable intent inference. CoT reasoning can enhance reasoning performance by structuring sparse observations into sequential inference steps. Although large-scale models with large parameter counts demonstrate strong CoT reasoning capabilities, their deployment in low-resource domains is constrained by limited edge computing resources [14], heterogeneous and unstable network environments [15], and the trade-off between real-time responsiveness and inference latency [16]. Conversely, lightweight models are more suitable for resource-constrained deployment, but often lack sufficient reasoning capacity and rely on scarce, costly-to-annotate data in construction domains [17]. Recent studies indicate that even a small number of high-quality samples can activate complex reasoning capacity in pretrained models[18]. Related work such as Fine-tune-CoT [19] explores the use of CoT supervision for reasoning adaptation with limited data, while parameter-efficient fine-tuning methods, including LoRA [20], are widely adopted, though full-parameter fine-tuning remains advantageous for fully unlocking complex reasoning capabilities when sufficient computing resources are available for training.

3 Methodology

This paper proposes a skeleton-enhanced vision-language framework for robust activity understanding in construction HRC. As illustrated in Figure 1, the framework consists of four components: (1) human skeleton keypoints extracted by RTMPose [21] are overlaid onto RGB images to create pose-enhanced visual inputs (§3.1); (2) a two-stage CoT reasoning pipeline guides the model to recognize current construction intent and predict top-3 next actions (§3.2); (3) an efficient data annotation pipeline combines teacher model generation with expert review,

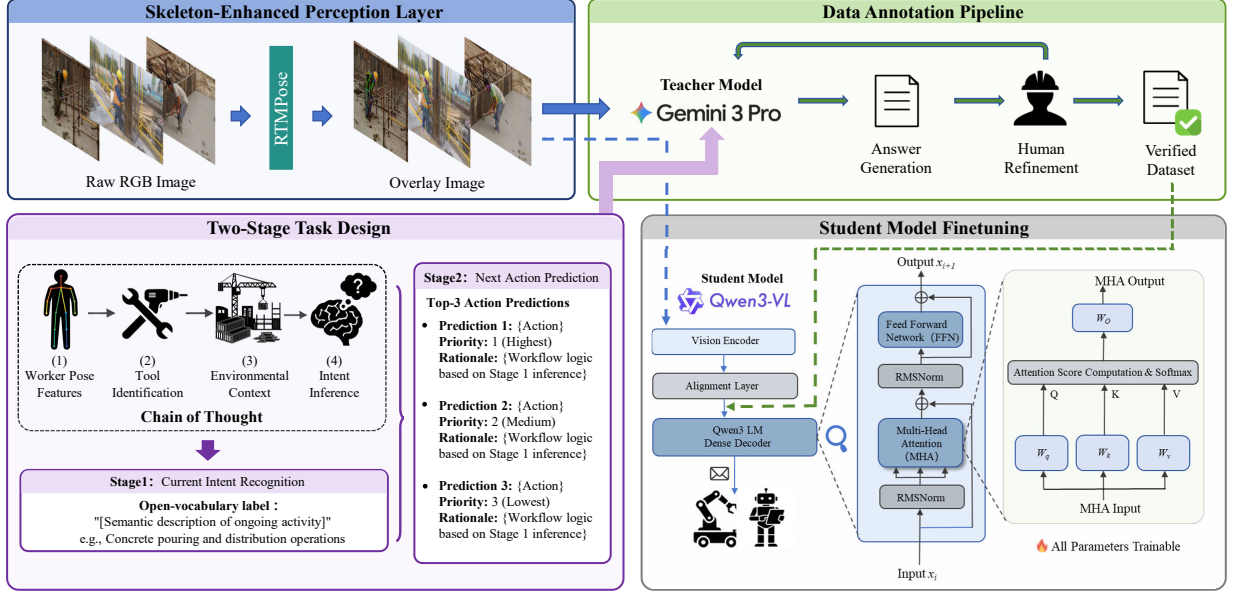


Figure 1. Architecture of the proposed framework.

achieving speedup while maintaining high quality (§3.3); (4) a teacher-student distillation strategy transfers knowledge from the powerful teacher model (Gemini 3 Pro) to a lightweight student model (Qwen3-VL-8B) through full-parameter SFT on the annotated data (§3.4).

3.1 Skeleton-Enhanced Visual Representation

Traditional RGB-based activity recognition methods struggle with varying lighting, occlusions, and cluttered backgrounds on construction sites. While pure skeleton-based methods offer robustness, they discard critical environmental context such as construction tools and materials. To address both limitations, a skeleton-enhanced visual representation is proposed that overlays human pose keypoints onto RGB images. Human skeleton extraction is performed using RTMPose, which detects 17 COCO-format keypoints in real-time. The extracted skeleton is rendered as semi-transparent green lines ($\alpha = 0.6$) with red joint markers overlaid on the original RGB image. This design preserves both explicit pose information and environmental context in a single unified input. In this representation, the skeleton provides robustness against lighting and occlusions, while the RGB background enables environmental understanding of tools, materials, and workspace layout. This aligns with the requirements of robot-assisted construction scenarios, where robot-mounted cameras typically recognize worker intent from dynamic on-site viewpoints. The collected image dataset was captured from human perspectives at construction sites, consistent with such mobile viewpoints and reflecting realistic HRC conditions.

3.2 Two-Stage Chain-of-Thought Reasoning

To support real-time robot assistance in construction environments, a two-stage reasoning pipeline is designed that leverages CoT prompting to guide the VLM through structured analysis. The designed prompt template instructs the model to perform current intent recognition and next-action prediction sequentially.

Stage 1: Current intent recognition. Given a skeleton-overlaid image, the model identifies ongoing construction activities through multi-step reasoning. The prompt is designed to analyze: (1) worker pose features, (2) tool identification, (3) environmental context, and (4) intent inference. This enables open-vocabulary recognition, generating descriptive labels such as “rebar tying at column base” without predefined categories.

Stage 2: Next action prediction. Based on the recognized intent, the model predicts the top-3 most probable next actions with priority rankings. The prompt instructs the model to consider standard construction workflows. For example, after “rebar tying”, predicting “continue tying adjacent joint”, “quality inspection”, or “move to next area”. Each prediction includes a priority level and rationale, enabling the robot to assist workers.

3.3 Data Annotation with Human-in-the-Loop

Training VLMs requires high-quality CoT annotations. To balance efficiency and domain accuracy, a two-step annotation pipeline is developed: (1) **Teacher model generation.** A large-scale VLM (Gemini 3 Pro) generates initial structured drafts using skeleton-overlaid images and designed CoT templates. These drafts include reasoning

processes, intent labels, and top-3 action predictions, serving as high-quality annotation candidates that capture the reasoning structure. (2) **Expert review and refinement.** Construction domain experts review teacher-generated annotations through a custom graphical user interface. Experts verify activity classifications against standard construction practices and experience, validate action predictions based on real workflows, and correct errors in reasoning chains. This pipeline minimizes annotation effort and accelerates the development of reasoning chains by allowing experts to focus on refining existing outputs rather than creating them from scratch.

3.4 Teacher-assisted Instruction Tuning and Full-Parameter Fine-tuning

While the teacher model (Gemini 3 Pro) performs well in the intent recognition task, its high computational cost and inference latency limit practical deployment in real-time robotic systems. To address these limitations, Qwen3-VL-8B-Instruct is selected for instruction tuning because its 8-billion parameter scale balances robust vision-language understanding with the efficiency required for edge deployment. To transfer advanced reasoning capabilities, the model is fine-tuned via full-parameter supervised fine-tuning (SFT) on expert-verified, CoT reasoning chains generated by the teacher model. This approach ensures higher inference efficiency and lower computational costs while preserving the high-quality reasoning inherited from the teacher. Detailed training configurations and hyperparameters are provided in Section 4.2.

4 Experiments

This section presents comprehensive experiments to validate the effectiveness of the proposed skeleton-enhanced vision-language framework for construction intent recognition. Section 4.1 describes the dataset construction process, including data sources, skeleton extraction, and the teacher-student annotation pipeline with expert verification. Section 4.2 details the implementation configurations for pose estimation, teacher model annotation generation, and student model fine-tuning. Section 4.3 introduces the LLM-as-judge evaluation approach for open-vocabulary intent recognition.

4.1 Dataset

The on-site construction activity dataset is constructed from multiple public construction safety datasets and online sources, totaling 1,819 images. The dataset aggregates images from four sources: CPPE [22], MOCS [23], pictor-ppe [24], and online construction images.

The dataset contains 1,074 single-person images (59%) and 745 multi-person images (41%). To align with robot-

assisted collaboration settings, we select a primary worker for skeleton annotation in multi-person scenarios. Selection prioritizes the lead actor based on visibility, joint completeness, and central positioning. This approach ensures high-quality annotations that reflect the realistic robotic focus. The resulting dataset comprises skeleton-overlaid images paired with expert-verified CoT reasoning chains (intent labels and top-3 action predictions). The data is split into training (1,637 images, 90%) and validation (182 images, 10%) sets, stratified by source to maintain diversity in multi-person scenarios.

4.2 Implementation Details

Human pose estimation is performed using RTMPose-L [21] via MMPoseInferencer with a keypoint confidence threshold of 0.3. During the data annotation phase, Gemini 3 Pro is accessed via Google AI Studio API to generate initial annotations, with temperature set to 0.7 and maximum output tokens to 2,048. The student model (Qwen3-VL-8B-Instruct [25]) is fine-tuned using LLaMA-Factory. Instead of parameter-efficient methods, a full-parameter SFT strategy is employed to maximize the reasoning capacity transfer. Training is conducted on four NVIDIA H800 GPUs (80GB each) with mixed-precision (bfloat16) and DeepSpeed ZeRO-3 optimization. Per-device batch size is 2 with gradient accumulation over 8 steps, achieving an effective global batch size of 64. The model is trained for 3 epochs with a learning rate of $5e-5$, a cosine learning rate schedule, and 10% linear warmup. In the context of instruction fine-tuning for VLMs, limiting the training to a small number of epochs is an empirical best practice [26]. This limitation prevents the generative model from overfitting its sequence generation style to the training distribution and mitigates catastrophic forgetting of its broad pre-trained capabilities, thus allowing the model to optimally master the required CoT reasoning format while preserving its inherent open-vocabulary potential.

To validate the effectiveness of skeleton overlay and CoT reasoning, four model configurations are evaluated. The baseline configuration uses the pretrained Qwen3-VL-8B model without fine-tuning on skeleton-overlaid images, testing baseline model’s capability. Three fine-tuned variants are trained with identical SFT configurations and procedures as detailed above, differing only along two dimensions: input representation (RGB images versus skeleton-overlaid images) and annotation format (standard question-answering versus CoT reasoning). Each variant is trained independently from the same pretrained checkpoint to ensure controlled comparison.

4.3 Evaluation Metrics

Due to the open-vocabulary nature of intent generation, traditional classification metrics based on exact label matching are inappropriate. Instead, semantic consistency is evaluated using an LLM-as-judge protocol [27]. Specifically, for each validation sample i , three state-of-the-art language models $\mathcal{J} = \{j_1, j_2, j_3\}$ (GPT-4o, Claude 3.5 Sonnet, and Gemini 3 Pro) independently assess whether the predicted intent $\hat{y}_i$ is semantically consistent with the ground truth y_i , producing binary judgments $s_{i,j} \in \{0, 1\}$ where 1 indicates consistency. The final decision for sample i is determined by majority voting v_i :

$$v_i = \begin{cases} 1 & \text{if } \sum_{j \in \mathcal{J}} s_{i,j} \geq 2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

For N validation samples, Intent Recognition Accuracy is then computed as:

$$\text{Acc}_{\text{intent}} = \frac{1}{N} \sum_{i=1}^N v_i \quad (2)$$

To quantify inter-judge agreement and ensure evaluation reliability, Fleiss' Kappa coefficient [28] is computed. For N samples evaluated by $|\mathcal{J}| = 3$ judges with binary decisions, Fleiss' Kappa κ is defined as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (3)$$

where $\bar{P}$ is the observed agreement, and $\bar{P}_e$ is the expected agreement by chance. Specifically, for each sample i , the proportion of agreement P_i is:

$$P_i = \frac{1}{|\mathcal{J}|(|\mathcal{J}| - 1)} \sum_{k \in \{0,1\}} n_{ik}(n_{ik} - 1) \quad (4)$$

where n_{ik} is the number of judges assigning category k to sample i . The observed agreement is the average across all samples:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i \quad (5)$$

The expected agreement is computed from the overall proportion of judgments in each category:

$$\bar{P}_e = \sum_{k \in \{0,1\}} p_k^2, \quad \text{where } p_k = \frac{1}{N \cdot |\mathcal{J}|} \sum_{i=1}^N n_{ik} \quad (6)$$

Fleiss' Kappa ranges from -1 to 1 , where $\kappa < 0$ indicates agreement worse than chance, $0 \leq \kappa < 0.20$ indicates slight agreement, $0.20 \leq \kappa < 0.40$ indicates fair agreement, $0.40 \leq \kappa < 0.60$ indicates moderate agreement, $0.60 \leq \kappa < 0.80$ indicates substantial agreement, and $0.80 \leq \kappa \leq 1.00$ indicates perfect agreement. This metric ensures that high accuracy is not achieved through inconsistent or arbitrary judgments across LLM evaluators.

5 Results and Analysis

Table 1 presents intent recognition accuracy evaluated through LLM-as-judge consensus voting. The proposed method achieves 71.43% accuracy with strong inter-judge agreement ($\kappa = 0.936$), improving over the baseline ($\text{Acc}_{\text{intent}} = 43.40\%$, $\kappa = 0.915$) by 28.03% in accuracy.

Fine-tuning with RGB images and SFT improves performance to 61.54% ($\kappa = 0.941$), demonstrating the value of domain adaptation through supervised learning. Adding skeleton overlay without CoT ($\text{Acc}_{\text{intent}} = 65.38\%$, $\kappa = 0.919$) shows an increase compared to RGB images with SFT fine-tuning (+3.84%), suggesting that explicit pose representation alone provides useful guidance. The full method incorporating both skeleton overlay and CoT reasoning achieves an additional +9.89% gain over RGB images with SFT, indicating that structured reasoning enables the model to systematically leverage pose information for robust intent recognition. Moreover, it is crucial to benchmark this 71.43% accuracy against the intrinsic difficulty of the task. Unlike traditional closed-set action recognition which often achieves high accuracy by classifying bounded labels, open-vocabulary intentional understanding requires the model to not only recognize geometric postures but also causally reason about the worker's underlying intention in plain text. A recent comprehensive benchmark by Liu et al. [29] on fine-grained action intelligence reveals that state-of-the-art MLLMs exhibit profound limitations in the higher-order reasoning of intentional understanding, achieving only around 60% accuracy in general scenarios. In light of this intrinsic difficulty, achieving 71.43% accuracy with a compact and edge-deployable model, enhanced by skeleton overlays and structured CoT reasoning, demonstrates practical value.

Analysis of experimental results reveals how skeleton enhancement and CoT reasoning address challenges of low-light conditions and occlusions through three representative scenarios and case studies. Skeleton overlay provides critical advantages in challenging conditions where RGB features alone are insufficient (see Figure 2). Figure 2(a) demonstrates that skeletal cues remain detectable even under low-light, shadowed conditions. While RGB features degrade in the dark foreground due to low contrast, the skeleton structure remains highly interpretable regardless of the challenging illumination. Figure 2(b) illustrates that the skeleton structure remains recognizable under partial occlusion. When construction panels block portions of the worker's body, the skeleton extraction algorithm successfully captures visible limbs and infers occluded joint positions through anatomical constraints. The resulting overlay maintains structural coherence even across obstructed regions. Figure 2(c) presents a complex scenario combining both low light and physical occlusion from formwork and support structures. In this case, the

Table 1. Stage 1 intent recognition performance on validation set (182 images)

Configuration	Intent Recognition Accuracy Acc_{intent} (%)				Δ vs Baseline	κ
	GPT-4o	Claude 3.5 Sonnet	Gemini 3 Pro	Majority Vote		
Baseline	37.36	39.56	52.74	43.40	–	0.915
RGB + SFT	47.25	63.19	65.38	61.54	+18.14	0.941
Overlay + SFT	49.45	62.09	70.33	65.38	+21.98	0.919
Overlay + CoT + SFT (Ours)	52.75	67.03	75.27	71.43	+28.03	0.936

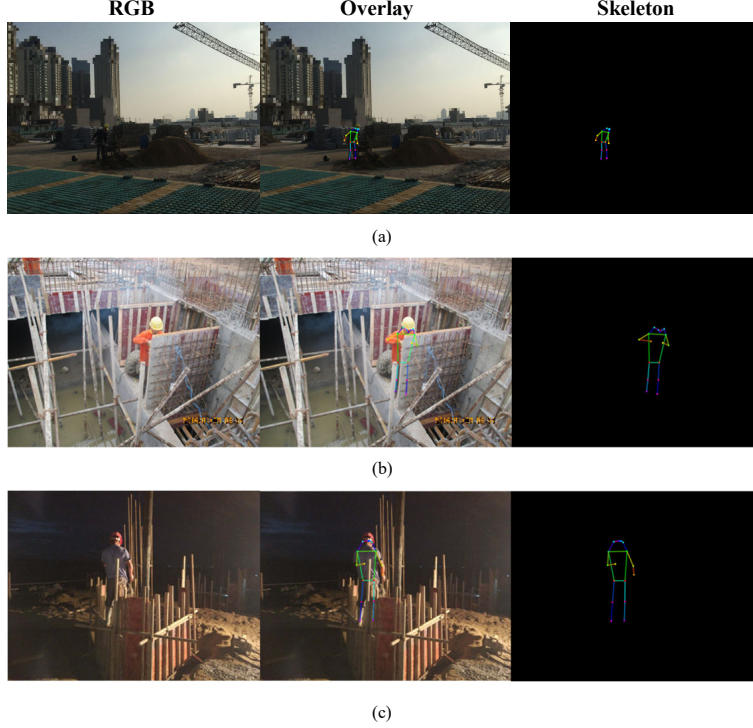


Figure 2. Skeleton enhancement provides robustness in challenging scenarios.

framework reliably reconstructs the human pose despite the simultaneous impact of poor lighting and dense structural blockage.

Figure 3 demonstrates successful two-stage reasoning for construction intent recognition and action prediction. The model’s CoT integrates pose, tool, and environmental information to correctly identify asphalt pavement leveling. Based on this intent, the predicted top-3 next actions match practical road-construction workflow needs: continuing leveling with machine progress, inspecting pavement flatness, and clearing residual materials. This case shows that skeleton enhancement improves pose recognition, while CoT reasoning supports accurate activity understanding and workflow prediction. In contrast, Figure 4 reveals tool-level confusion under limited observability. The dense rebar-and-steel-column background and ambiguous hand-tool interaction obscure discriminative grinding cues, making fine-grained activity distinction unreliable. Specifically, the model predicts “rebar node binding operation”, whereas the ground truth is “steel structure grinding operation”. This highlights a key limitation:

when explicit tool evidence is weak or occluded, the current reasoning may over-rely on posture-context priors and produce overconfident intent classification.

6 Discussion and Conclusion

This paper presents a skeleton-enhanced vision-language framework for construction worker intent recognition and action prediction. The framework integrates skeleton overlay for visual robustness, CoT reasoning for structured analysis, and teacher-student-instructed fine-tune for practical deployment. Experimental results demonstrate significant performance improvements from both skeleton enhancement and CoT reasoning, with real-time inference suitable for robot applications. While inferring a worker’s current work intent from visual observations remains an extremely challenging frontier for even the most massive VLMs, this framework provides a paradigm shift by coupling domain-specific geometric guidance with reasoning logic. This allows edge-friendly models to produce reliable rationales of *why* a worker is moving, going beyond bounded classification labels. The

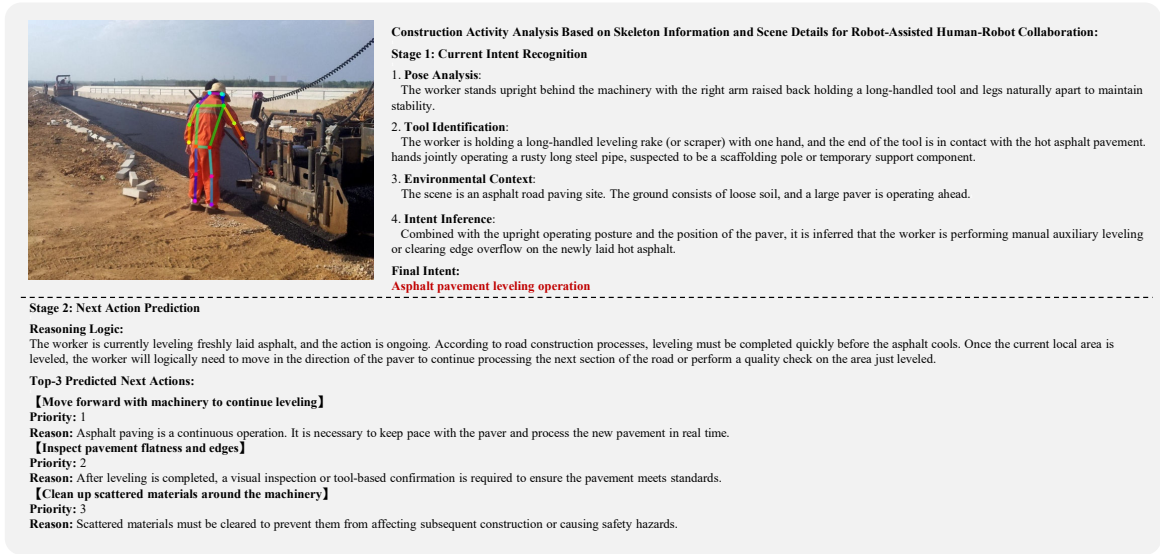


Figure 3. Illustration of a successful two-stage reasoning case.

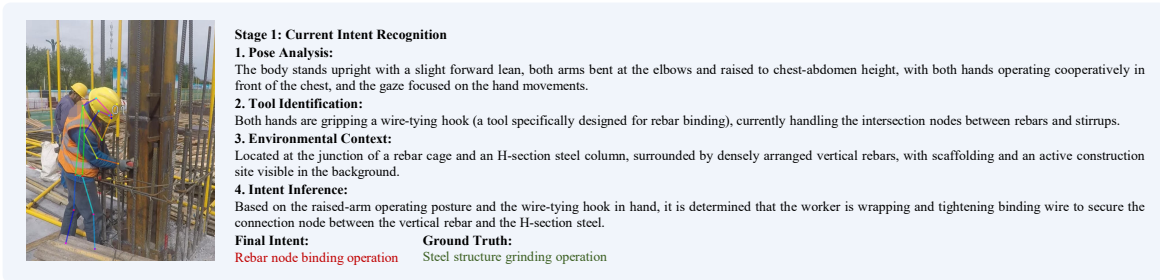


Figure 4. Illustration of a failure case.

open-vocabulary formulation handles diverse construction activities without predefined categories, establishing a foundation for proactive and safe HRC in construction.

Several limitations remain. The work relies on images selected from open-source datasets rather than collected from specific construction projects, resulting in inconsistent image quality. The limited dataset size constrains model accuracy improvements, and training examples for rare specialized activities remain insufficient. Single-frame inference lacks temporal context for activity progression, and skeleton extraction remains inaccurate in certain cases involving extreme occlusion or overhead camera perspectives. Future research should address these limitations through three primary directions: (1) developing more accurate skeleton extraction models with enhanced robustness to challenging viewpoints and severe occlusion conditions will improve pose representation quality; (2) extending the framework from single-frame images to video sequences as source data will enable temporal modeling of activity progression and action continuity; (3) incorporating temporal information into the action prediction evaluation protocol will better assess the model's

understanding of construction workflow dynamics. Additional research directions include multi-agent analysis for collaborative workflows, domain adaptation techniques for cross-site generalization, few-shot learning approaches for rare specialized activities, and real-world deployment studies with edge device optimization.

7 Acknowledgements

This work was funded by The Science and Technology Development Fund, Macau SAR (File no. 0101/2024/RIB2).

References

- [1] Y. Kim, H. Kim, R. Murphy, and et al. Delegation or collaboration: Understanding different construction stakeholders' perceptions of robotization. *Journal of Management in Engineering*, 38(1):04021084, 2022.
- [2] Z. Wu, Q. Liao, K. Chen, and et al. A systematic review of digital auxiliary technologies enabling human-robot collaboration in construction. *Engi-*

- neering, *Construction and Architectural Management*, 2025.
- [3] Y. Tian, J. Chen, J. I. Kim, and et al. Multiple-input streams attention (misa) network for skeleton-based construction workers' action recognition using body-segment representation strategies. *Automation in Construction*, 156:105104, 2023.
 - [4] F. Lin. Vision language models: A survey of 26k papers. *arXiv preprint arXiv:2510.09586*, 2025.
 - [5] M. Wang, J. Chen, and J. Ma. Monitoring and evaluating the status and behaviour of construction workers using wearable sensing technologies. *Automation in Construction*, 165:105555, 2024.
 - [6] M. Zhang, R. Xu, H. Wu, and et al. Human-robot collaboration for on-site construction. *Automation in Construction*, 150:104812, 2023.
 - [7] Y. Gong, J. Seo, K.-S. Kang, and et al. Automated recognition of construction worker activities using multimodal decision-level fusion. *Automation in Construction*, 172:106032, 2025.
 - [8] B. Zhong, L. Shen, X. Pan, and et al. Visual attention framework for identifying semantic information from construction monitoring video. *Safety Science*, 163:106122, 2023.
 - [9] R. Dwivedi, D. Dave, H. Naik, and et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9):1–33, 2023.
 - [10] M. Park, D. Q. Tran, J. Bak, and et al. Small and overlapping worker detection at construction sites. *Automation in Construction*, 151:104856, 2023.
 - [11] S. Yun, S. Hong, S. Hwang, and et al. Analysis of masonry work activity recognition accuracy using a spatiotemporal graph convolutional network across different camera angles. *Automation in Construction*, 175:106178, 2025.
 - [12] S. Hong, B. Lee, T. Kim, and et al. Hybrid continuous construction worker pose estimation under occlusion and missing data using st-gcn and kalman filtering. *Advanced Engineering Informatics*, 69:104072, 2026.
 - [13] A. Sinha, D. Reilly, F. Bremond, and et al. Ski models: Skeleton induced vision-language embeddings for understanding activities of daily living. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:6931–6939, Apr. 2025.
 - [14] S. Danish, A. Sadeghi-Niaraki, S. U. Khan, and et al. A comprehensive survey of vision-language models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Information Fusion*, 126:103623, 2026.
 - [15] H. Gauttam, G. Nain, K. K. Pattanaik, and et al. Edge-ai: A systematic review on architectures, applications, and challenges. *Journal of Network and Computer Applications*, 245:104375, 2026.
 - [16] J. Fan, Y. Yin, T. Wang, and et al. Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey. *Frontiers of Engineering Management*, 12(1):177–200, 2025.
 - [17] Y. Ma, Z. Yao, X. Liu, and et al. Efficient 3d object annotation via vision-derived pseudo-lidar and vision language model (vlm) validation. *Transportation Research Part C: Emerging Technologies*, 182:105429, 2026.
 - [18] Y. Ye, Z. Huang, Y. Xiao, and et al. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.
 - [19] N. Ho, L. Schmid, and S.-Y. Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 14852–14882, Toronto, Canada, 2023. Association for Computational Linguistics.
 - [20] E. J. Hu, Y. Shen, P. Wallis, and et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
 - [21] T. Jiang, P. Lu, L. Zhang, and et al. RtmPose: Real-time multi-person pose estimation based on mmpose. *CoRR*, abs/2303.07399, 2023.
 - [22] R. Xiong and P. Tang. Pose guided anchoring for detecting proper use of personal protective equipment. *Automation in Construction*, 130:103828, 2021.
 - [23] X. An, L. Zhou, Z. Liu, and et al. Dataset and benchmark for detecting moving objects in construction sites. *Automation in Construction*, 122:103482, 2021.
 - [24] N. D. Nath, A. H. Behzadan, and S. G. Paal. Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, 112:103085, 2020.
 - [25] S. Bai, Y. Cai, R. Chen, and et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
 - [26] H. Liu, C. Li, Q. Wu, and et al. Visual instruction tuning. *NeurIPS*, 2023.
 - [27] L. Zheng, W.-L. Chiang, Y. Sheng, and et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023.
 - [28] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
 - [29] Dayong Liu, Chao Xu, Weihong Chen, and et al. Beyond description: Cognitively benchmarking fine-grained action for embodied agents. *arXiv preprint arXiv:2511.18685*, 2025.