Uncertainty-aware collaboration request prediction for proactive human-robot collaboration in construction

Zaolin Pan and Yantao Yu

Abstract— **Proactive Human-Robot Collaboration (HRC) in construction holds transformative potential for enhancing operational fluency by enabling robots to anticipate human intentions—a capability termed collaboration request prediction. However, current HRC systems predominantly rely on deterministic, single-human action prediction models. These approaches are unable to account for the inherent uncertainties in dynamic, multi-human interactions and long-term predictions, risking inappropriate robotic actions when predictions are overtrusted. Despite the critical need for robust Uncertainty Quantification (UQ) in such high-stakes, dynamic environments, discussions on UQ for HRC in construction remain limited. To bridge this gap, we propose a holistic uncertainty-aware framework for long-horizon multi-human action anticipation. We address the central question: What uncertainty—epistemic (stemming from model ignorance or atypical data), aleatoric (intrinsic to human behavior and environmental noise), or both—are most critical for robust long-term action anticipation in multi-human HRC? Our approach explicitly disentangles epistemic uncertainty and aleatoric uncertainty through a combination of Monte Carlo dropout for epistemic uncertainty estimation and heteroscedastic modeling for input-dependent aleatoric uncertainty quantification. Additionally, we explore evidential deep learning to model both uncertainties simultaneously via learned distributional parameters. Empirical validation on real-world scaffolding datasets reveals that explicit modeling of both uncertainties improves prediction reliability and decision-making resilience. These findings underscore the necessity of holistic UQ in deploying adaptive, trustworthy HRC systems within complex, dynamic construction environments.**

I. Introduction

Inherently hazardous and rapidly changing environments like construction sites present unique challenges—and opportunities—for Human-Robot Collaboration (HRC) [1]. By partnering with robots, construction teams can potentially achieve greater productivity, safety, and efficiency [2]. Yet, realizing this potential requires moving beyond robots as mere reactive assistants that are limited to responding to immediate human actions. Instead, robots must become proactive collaborators, capable of anticipating future requests and needs, and participating in dynamic multi-agent interaction processes [3].

Proactive HRC demands robotic systems that not only interpret real-time human actions but also forecast upcoming tasks, collaborative needs, and group dynamics over extended time horizons. For instance, robots may require up to 15 seconds to complete preparatory actions, such as Atlas retrieving and positioning a plank [4], underscoring the necessity for foresight to align assistance with evolving workflows. This anticipatory capability enables preemptive adaptation, minimizes disruptions, and facilitates seamless integration into human-centric teams.

Yet, long-term action anticipation in multi-worker construction settings is inherently uncertain. Sources of ambiguity include variable human behaviors, robot sensor limitations, and task ambiguities (e.g., overlapping or ill-defined workflows) [5]. These challenges manifest as two distinct types of uncertainty in robot decision-making:

- **Epistemic uncertainty,** arising from incomplete knowledge (e.g., sparse training data for rare activities or unconventional task sequences).
- **Aleatoric uncertainty,** stemming from intrinsic randomness in sensory inputs or human pacing variability.

To manage the inherent ambiguity of long-horizon reasoning, models must explicitly quantify the uncertainty, as it helps achieve transparent and trustworthy decision-making in HRC. However, existing approaches to long-term action anticipation for HRC typically rely on deterministic models that produce single-point predictions without associated confidence measures [6]. While other studies on single-agent action anticipation have explored uncertainty quantification, they often focus on either epistemic or aleatoric sources in isolation [7], [8], [9]. This gap make it unclear on the distinct and combined impact of both uncertainties on the long-term, multi-human action anticipation context. To fill these gaps, we introduce a unified, uncertainty-aware framework for long-term, multi-worker action anticipation. Our contributions are:

(1) A baseline model that predicts collaboration requests among multiple workers over extended time horizons.

(2) A systematic comparison of epistemic and aleatoric uncertainty estimation strategies within the same framework.

(3) Empirical validation on real-world scaffolding construction videos, demonstrating the strength of uncertainty-aware decision-making.

II. Problem Formulation

Given a short video clip $V \in \mathbb{R}^{K\times C\times H\times W}$, where K denotes the number of frames, C the channel dimension (e.g., RGB), and $H\times W$ the spatial resolution, we propose a framework to jointly infer human-centric attributes and enable proactive robotic assistance. The goal is to learn a model f that, for each detected worker i, predicts: (1) the current action label

Corresponding author: Yantao Yu. The authors are with the Hong Kong University of Science and Technology, Hong Kong SAR, China (email: zpanaq@connect.ust.hk; ceyantao@ust.hk).

$\text{cur}_i \in \mathbb{R}^{C_{cls}}$, (2) a binary role $r \in \{0, 1\}$ that can be generalized to bystander (0) or scaffolder (1), (3) a sequence of T future actions $\text{fut}_i^t \in \mathbb{R}^{T \times (C_{cls}+1)}$, and (4) temporally grounded action start times s_i^t and uncertainty estimates u_i^t (aleatoric, epistemic, or both) for each future action t. To ensure robustness, predictions are filtered using confidence (τ_{conf}) and uncertainty (τ_{unc}) thresholds, discarding low-confidence or high-uncertainty outputs. The remaining predictions are prioritized into a future action request queue Q by combining temporal urgency (earlier s_i^t) and initiative support (prioritizing scaffolders). Finally, the robot executes assistance actions in the order of Q, dynamically updating the queue as new video data arrives.

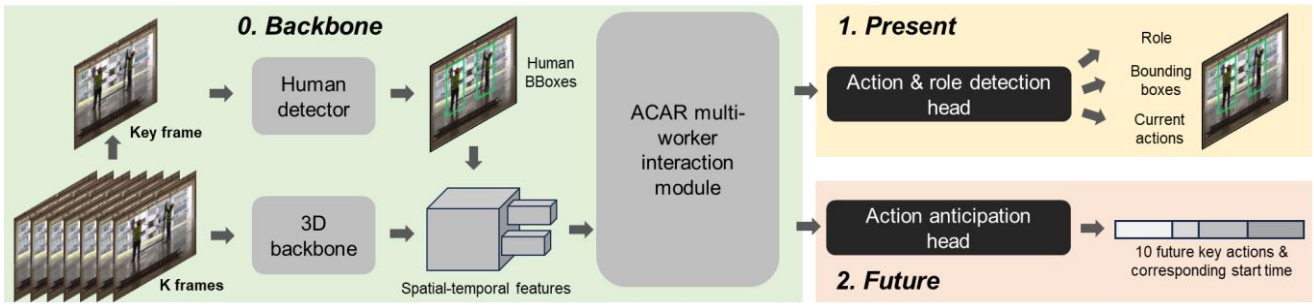

Figure 1. Framework of the proposed multi-human long-term action anticipation baseline.

III. METHODOLOGY

A. Multi-human Long-term Action Anticipation

Building on existing action detection models capable of advanced multi-human action localization and classification, we propose a framework for multi-human long-term action anticipation. Unlike auto-regressive approaches—which perform stepwise, conditional predictions prone to error propagation and entangled uncertainty estimation—we formulate intention anticipation as a fixed-horizon, parallel action-duration prediction task. This parallel approach directly models the global joint distribution of actions and durations, explicitly decoupling aleatoric and epistemic uncertainties while avoiding the computational inefficiency of iterative predictions. Fixed-horizon prediction further enhances practicality in multi-human construction scenarios, where autoregressive methods incur prohibitive computational costs due to repeated forward passes.

The network structure is shown in Fig. 1. Our framework adapts the ACAR model [10], chosen for its explicit modeling of multi-human interactions critical to collaborative construction tasks. The backbone integrates YOLOv12 [11] for human detection and SlowFast R-50 [12] for spatiotemporal feature extraction. The original ACAR action detection head is extended to classify worker roles (e.g., scaffolder vs. bystander), reducing noise from irrelevant personnel in intention anticipation. The action anticipation head employs two parallel MLPs: one predicts ten future actions, while the other estimates their corresponding start times. These ten horizons, spanning the main scaffold assembly workflow, balance granularity and computational feasibility.

B. Uncertainty Estimation

In this section, we present the uncertainty modeling strategies categorized into three types: epistemic, aleatoric, and a combination of both. It's important to note that these strategies are applied solely to the action prediction branch, excluding the duration branch. This approach is designed to prevent interference from the other branches and to ensure training stability.

1) Epistemic Uncertainty Only

Epistemic uncertainty reflects our lack of knowledge about the true model parameters and can be reduced with more data. Common techniques for capturing this form of uncertainty are Monte Carlo (MC) dropout and deep ensembles. Balancing training complexity and inference-time efficiency, we opt for MC dropout due to its simplicity, minimal computational overhead, and seamless integration into existing architectures.

2) Aleatoric Uncertainty Only

Aleatoric uncertainty captures the irreducible noise in the data—the variability inherent to the observations themselves. It splits into two types [13]: homoscedastic uncertainty, which is constant across all inputs (e.g. sensor noise of fixed magnitude), and heteroscedastic uncertainty, which varies with each input (e.g. some scenes or actions are intrinsically more ambiguous than others).

In an action anticipation setting, heteroscedastic uncertainty is especially important: for instance, when a worker is installing a cross-brace, the next action is fairly predictable (low uncertainty), whereas a worker merely wandering yields much greater ambiguity (high uncertainty). To model this, we equip our network—parameterized by $\mathbf{W}$—to predict, at each future horizon i, both a vector of logits $\mathbf{f}_i^{\mathbf{W}}$ and a variance parameter $\boldsymbol{\sigma}_i^{\mathbf{W}}$ scaling input-dependent noise. The logits are perturbed with Gaussian noise, as shown below:

$$\hat{\mathbf{x}}_i \mid \mathbf{W} \sim \mathcal{N}\left(\mathbf{f}_i^{\mathbf{W}}, (\boldsymbol{\sigma}_i^{\mathbf{W}})^2\right). \quad (1)$$

The perturbed logits $\hat{\mathbf{x}}_i$ are normalized via a softmax function to generate a probability distribution over actions:

$$\hat{\mathbf{p}}_i = \text{Softmax}(\hat{\mathbf{x}}_i). \quad (2)$$

Training the model involves a Monte Carlo sampling approach to approximate the expected log-likelihood of

actions. For each input i, T noise samples $\epsilon_t \sim \mathcal{N}(0, I)$ are drawn to perturb the logits:

$$\hat{\mathbf{x}}_{i,t} = \mathbf{f}_i^{\mathbf{W}} + \boldsymbol{\sigma}_i^{\mathbf{W}} \epsilon_t . \quad (3)$$

The loss function marginalizes over these samples to encourage learning of input-specific variances:

$$\mathcal{L}_x = \sum_i \log \frac{1}{T} \sum_t \exp\left(\hat{x}_{i,t,c} - \log \sum_{c'} \exp \hat{x}_{i,t,c'} \right), \quad (4)$$

where c indexes the ground-truth action class. This strategy enables the model to distinguish high-confidence predictions (e.g., tool retrieval) from uncertain ones (e.g., ambiguous movements).

3) Both Epistemic and Aleatoric Uncertainty

Combining epistemic and aleatoric uncertainty quantification is critical for robust decision-making in HRC. While MC dropout offers a straightforward approach by sampling stochastic forward passes to estimate epistemic uncertainty—and can be combined with aleatoric uncertainty modeling—this method incurs significant computational overhead due to repeated inference steps. Alternative frameworks, Evidential Deep Learning (EDL) [14], provide a more unified solution by directly modeling both uncertainty types through probability distributions, inspired by Bayesian statistics and Dempster-Shafer theory. EDL replaces conventional softmax outputs with parameters of a Dirichlet distribution, which represents the model's belief over class probabilities. For a classification task with K classes, the model predicts concentration parameters $\alpha = (\alpha_1, ..., \alpha_K)$, where $\alpha_K > 0$ reflects the evidence for class k. Key metrics derived from these parameters include:

- Total evidence: $S = \sum_{k=1}^{K} \alpha_k$, quantifying the overall confidence in predictions.
- Expected probability: $p_k = \frac{\alpha_k}{S}$, the normalized class probability.
- Epistemic uncertainty: $u = \frac{K}{S}$, inversely proportional to total evidence, capturing model ignorance.
- Aleatoric uncertainty: $a = \text{Entropy}(\boldsymbol{p}) - u$, measuring inherent data noise via the entropy of expected probabilities.

The EDL loss combines two components to balance data fitting and uncertainty calibration, Bayes risk loss for data fitting and Kullback-Leibler (KL) divergence to penalize incorrect evidence:

$$\begin{aligned} \mathcal{L}_{\text{BR}} &= \sum_{k=1}^{K} y_k \left(\psi(S) - \psi(\alpha_k) \right), \\ \mathcal{L}_{\text{KL}} &= \log\left(\frac{\Gamma(S)}{\Gamma(K)} \right) + \sum_{k=1}^{K} (\alpha_k - 1) \left(\psi(\alpha_k) - \psi(S) \right), \\ \mathcal{L} &= \mathcal{L}_{\text{BR}} + \lambda_t \mathcal{L}_{\text{KL}}, \end{aligned} \quad (5)$$

where y_k denotes ground-truth one-hot label, ψ and Γ represent digamma and Gamma functions, respectively, and λ_t is annealing weight.

IV. Experiment

A. Experimental Setup

Experiments used real-world videos of duo-worker scaffolding construction, featuring flexible task flows and occasionally unrelated workers. The dataset consists of 33 videos (~70 minutes total), annotated with 11 action labels and split 4:1 for training and testing. We assume a robot assistant capable of simple tasks like transporting jack plates or cross braces and stabilizing vertical frames. Based on the Atlas robot's speed (15s to transport a baseplate), we allocate 20s of preparation time for the robot to initiate supportive actions.

B. Evaluation Metric

We evaluate performance using frame mean Average Precision (frame-mAP) for action detection, top-1 accuracy for action prediction, and Mean over Class (MoC) accuracy for action anticipation. Uncertainty is quantified via Expected (ECE) and Maximum Calibration Error (MCE). For collaboration request prediction—determining when and what assistance to provide based on anticipated worker intent and confidence thresholds—we use mAP@0.5:0.95tIoU.

C. Qualitative Results

The visualization of the inspection results is shown in Fig. 2. The worker's locations, role, current and future actions, and corresponding start times and uncertainties are predicted.

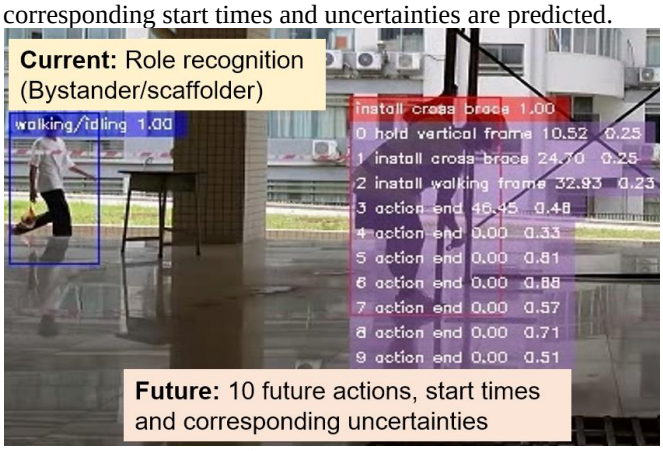

Figure 2. Visualization of multi-worker long-term action anticipation.

D. Quantitative Results

Table I summarizes the baseline model's performance. It achieves strong action detection (frame-mAP: 0.936), with moderate yet acceptable performance in action prediction and anticipation, likely due to worker behavior uncertainty and long prediction horizons. These results collectively confirm the model's reliability.

TABLE I. Multi-human Long-term Action Anticipation.

Action detection		**Action prediction**		**Action anticipation**	
Class label	*f-mAP*	*Prediction*	*Accuracy*	*Temporal*	*MoC*

		horizon		*horizon*	
gather cross brace	0.998	fut-0 (current)	0.924	3s	0.735
gather jack base	0.811	fut-1	0.814	5s	0.692
gather vertical frame	0.927	fut-2	0.550	10s	0.639
gather walking frame	0.834	fut-3	0.469	15s	0.628
hold vertical frame	0.880	fut-4	0.527	20s	0.623
install cross brace	0.987	fut-5	0.485	30s	0.601
install jack base	0.959	fut-6	0.512	50s	0.591
install vertical frame	0.970	fut-7	0.340	100s	0.572
install walking frame	0.963	fut-8	0.379	200s	0.566
walking/idling	0.998	fut-9	0.359	-	-
Overall	0.936	Overall	0.558	Overall	0.566

E. Uncertainty Estimation Results

Through systematic comparison of different uncertainty modeling strategies implemented on our baseline model, we draw three key conclusions from the results presented in Table 2. First, joint modeling of both epistemic and aleatoric uncertainties yields performance improvements in both action prediction and anticipation accuracy compared to the uncertainty-agnostic baseline. Second, uncertainty modeling enhances probability calibration, as evidenced by reduced ECE and MCE, indicating better alignment between predicted probabilities and empirical likelihoods. Third, the result reveals that aleatoric uncertainty constitutes the dominant component, as models incorporating only aleatoric uncertainty achieve comparable performance to those modeling both uncertainty types. These findings collectively demonstrate that while both uncertainties contribute to prediction quality, aleatoric uncertainty plays the primary role in long-term multi-human action anticipation scenarios.

TABLE II. UNCERTAINTY ESTIMATION.

Task (Metric)	Baseline	Epistemic	Aleatoric	Aleatoric+ Epistemic	EDL
Action prediction (*Accuracy*)	0.558	0.559	0.626	0.626	**0.635**
Action anticipation (*MoC@40s*)	0.591	0.588	0.596	0.594	**0.598**
ECE	0.385	0.384	**0.158**	0.161	**0.158**
MCE	0.456	0.745	0.219	0.224	**0.197**
Collaboration request prediction (mAP)	0.161	0.154	0.170	0.176	**0.178**

a. Under the optimal confidence threshold

V. CONCLUSION

This paper presents an uncertainty-aware framework for robust collaboration request prediction in multi-human construction scenarios. We develop a novel baseline model capable of long-term multi-human action anticipation, and systematic evaluation of epistemic and aleatoric uncertainty quantification for improving prediction reliability. Experimental results demonstrate that joint modeling of both uncertainties yields optimal performance, with aleatoric uncertainty playing the dominant role in dynamic construction environments. Future work will focus on developing more sophisticated baseline models, collecting and evaluating on more diverse, in-situ multi-worker interaction datasets. These directions will further bridge the gap between laboratory validation and practical deployment of uncertainty-aware HRC systems in complex construction environments.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China [grant number 72201226]; the Collaborative Research Fund (CRF) from the Research Grants Council (Hong Kong) [grant number C6044-23GF]; and the Research Grants Council (Hong Kong) [grant number 26208323].

REFERENCES

[1] C. Brosque, E. Galbally, O. Khatib, and M. Fischer, "Human-robot collaboration in construction: opportunities and challenges," in *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Jun. 2020, pp. 1–8.

[2] M. Zhang, R. Xu, H. Wu, J. Pan, and X. Luo, "Human–robot collaboration for on-site construction," *Automation in Construction*, vol. 150, p. 104812, Jun. 2023

[3] S. Li *et al.*, "Proactive human–robot collaboration: mutual-cognitive, predictable, and self-organising perspectives," *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102510, Jun. 2023.

[4] SITECH Solutions, *Video: Watch "Atlas" the Humanoid Fetch Tools for Worker on Scaffold*, (Feb. 01, 2023).

[5] P. Zheng, S. Li, J. Fan, C. Li, and L. Wang, "A collaborative intelligence-based approach for handling human-robot collaboration uncertainties," *CIRP Annals*, vol. 72, no. 1, pp. 1–4, Jan. 2023.

[6] S. Bhagat, S. Li, J. Campbell, Y. Xie, K. Sycara, and S. Stepputtis, "Let me help you! neuro-symbolic short-context action anticipation," *IEEE Robotics and Automation Letters*, pp. 1–8, 2024.

[7] C. Canuto, P. Moreno, J. Samatelo, R. Vassallo, and J. Santos-Victor, "Action anticipation for collaborative environments: The impact of contextual information and uncertainty-based prediction," *Neurocomputing*, vol. 444, pp. 301–318, Jul. 2021.

[8] Y. Abu Farha and J. Gall, "Uncertainty-aware anticipation of activities," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0. Accessed: Aug. 21, 2024.

[9] H. Guo, N. Agarwal, S.-Y. Lo, K. Lee, and Q. Ji, "Uncertainty-aware action decoupling transformer for action anticipation," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18644–18654.

[10] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li, "Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization," Apr. 20, 2021.

[11] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," Feb. 18, 2025.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," Oct. 29, 2019.

[13] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

[14] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty," Oct. 31, 2018.