Journal of Applied Automation Technologies

Research Paper

Data-Augmented Diffusion Modeling for Safety Assessment in Human–Robot Collaborative Assembly Workstations

Haruki Tanaka[1, *], Kaito Yamamoto[1], Daiki Yamada[1], Yi-hsuan Chiu[2] and Shu-yu Peng[2]

[1] Faculty of Information Science and Electrical Engineering, Kyoto University, Kyoto, 606-8501, Japan
[2] College of Electrical Engineering and Computer Science, National Taiwan University of Science and Technology, Taipei, 10607, China
*Corresponding author: haruk.tan@iue.kyoto-u.ac.jp

Abstract. Reliable safety assessment for human-robot collaborative assembly workstations faces problems such as severe class imbalance, sensor heterogeneity, and a lack of ethically recorded hazardous events. This paper presents a risk-conditioned diffusion model that extends the synchronized kinematic, proximity, ergonomic and control-state sequences without violating physical safety constraints. The model integrates risk stratification conditions, constraint-guided denoising, temporal consistency regularization, and uncertainty-aware sample filtering. Experiments use 18,640 observed windows from a collaborative fastening and inspection cell and 31,200 accepted synthetic windows. Augmentation increased the macro-F1 from 0.821 to 0.902 compared with training on observed data alone, raised critical-event recall from 0.714 to 0.893, and reduced expected calibration error from 0.087 to 0.036. Given a fixed false-alarm rate of 5%, the proposed assessment identifies 91.6% of high-risk windows and is 8.9% higher than the conditional adversarial baseline. Ablation results show that without constraint guidance, physically invalid samples increased from 1.8% to 12.7%, and uncertainty filtering added 3.4 percentage points to the macro-F1 score. Diffusion-based augmentation can increase the sensitivity of rare-event detection while maintaining operational feasibility and calibrated confidence; thus, an auditable data-centric pathway for workstation safety evaluation has been established.

Keywords: *Diffusion Model, Data Augmentation, Human–Robot Collaboration, Safety Assessment, Rare Event Detection*

Received on 14 November 2023, Accepted on 12 August 2024, Published on 24 August 2024

Introduction

Cooperative assembly is a type of automation that has collaborated with human workers and robots to work together in the same area, use tools jointly, and divide tasks. It is suitable for small-batch production and flexible in shape, but it fails to meet the necessary safety requirements on its own. Separation Distance, Relative Velocity, Hand Occupancy, Body Posture, Tool Orientation, Controller Mode and Task Phase collectively determine whether an encounter is normal or dangerous [1]. Proximity monitoring has therefore shifted from fixed areas to continuous estimations of the human-robot state [2]. Although visual sensors and wearable devices have increased observability, they have also introduced occlusion, drift, and synchronization errors [3]. Ergonomic Exposure is also a problem; repeated reaching and awkward joint loading may increase collision risk [4]. Existing industrial guidance defines protection functions and validation rules [5], and multimodal research prototypes combine depth cameras, skeleton tracking, force sensing, and predictive motion models [6]. The above progress provides the sensing basis, but recorded datasets are still dominated by normal production because intentionally creating dangerous situations is ethically unacceptable.

Learning-based safety assessment has therefore developed under an adverse data regime. Supervised classifiers can separate normal and abnormal motion when representative labels are available [7], and temporal neural networks capture transitions that frame-wise rules miss [8]. Graph representations model the dependencies among body parts, robot links and shared objects [9]. Uncertainty estimation enhances the interpretability of scores near the decision boundary [10], and domain adaptation handles alterations in operator physique, lighting and workstation layout [11]. Synthetic data have been introduced by simulation, oversampling and adversarial generation [12], but these methods often duplicate minority samples or produce sequences that violate velocity, reachability or controller-state constraints. Diffusion models have stable likelihood-based training, wide distributional coverage, and a reverse process that can incorporate physical guidance [13]. Directly apply these attributes to the problem of collaborative safety.

Therefore, the unresolved engineering problem is not merely to produce more samples, but to generate rare multimodal sequences that are still sufficiently coherent to support a safety decision. Develop a risk-conditioned diffusion augmentation framework for collaborative assembly workstations in this paper. It is a set of synchronized observations as task-aligned windows, learns a conditional denoising process for continuous and discrete channels, applies differentiable physical and temporal guidance during sampling, and rejects uncertain or out-of-support sequences before risk-model training. Assess fidelity, coverage, physical validity, minority-event recognition, calibration and robustness to operator and layout shifts in this study. The first is that it builds an all-in-one data-to-decision system that connects synthetic sequence generation with auditable safety indicators, rather than pursuing visual or statistical realism for its own sake.

Related Work

Safety Sensing and Assessment in Collaborative Assembly

Safety research on collaborative workstations often combines standards-based limits with data-driven observations. Speed-and-separation monitoring estimates the distance needed to stop the robot's motion before contact [14]. Power-and-force-limiting studies examine permissible contact and controller response [15]. Vision and wearable sensors provide supplementary information on a person's location and movement when they are temporarily out of view [16]. Multimodal fusion can address the problem of single-sensor failure, but clock alignment and explicit handling of missing channels are needed [17]. Ergonomic scores have also been added to robot planning to avoid transferring productivity gains as physical strain on the operator [18]. The above methods provide reasonable safety variables, but their evaluation datasets are relatively lacking in near-collision, unexpected-entry, dropped-part or protective-stop transitions.

The structure of the label is also more complex than a binary safe-unsafe division. Risk will continuously change as the distance decreases, the relative speed increases, or an operator enters a restricted position. A good model should distinguish between the precursor, elevated-risk, critical and recovery stages in time. Rule systems are relatively transparent but have abrupt boundaries; learned models can exploit weak combinations of cues. The advantage of them is reduced if the minority classes contain only a few operators or task variations. Cross-validation in a single recording session may thus overestimate the deployable performance due to near-identical motion in adjacent windows.

Data Augmentation and Generative Sequence Modeling

Data augmentation methods aim to expand the limited support. Random perturbation, time warping, interpolation and synthetic minority oversampling are relatively simple, but they maintain local topology and may blur safety thresholds [19]. Conditional adversarial and variational generators learn more flexible distributions; however, mode collapse or latent smoothing may occur during the process [20]. Recent time series models have demonstrated good multivariate synthesis and conditional flexibility, with diffusion models replacing adversarial optimization with progressive degradation and denoising [21]. Generally speaking, however, it does not guarantee kinematic feasibility or correct coupling of the task phase and robot control state.

A feasible expansion system should meet the following three linked conditions. First, it should cover rare combinations and not reproduce frequent windows. Second, the generated channels must meet the bounds, rates and cross-channel constraints of the workstation. Third, improved synthetic fidelity should be shown in the calibrated detection of untouched real sequences. Most of the previous studies report only one or the other. The three are all covered by coupling risk-stratified diffusion with constraint gradients and uncertainty-based acceptance, and then the resulting assessment is evaluated under operator-held-out and layout-shift protocols.

Risk-Conditioned Diffusion Augmentation Framework

Workstation Representation and Risk Labels

The collaborative fastening and visual-inspection cell in the study has six axes of a robot, a torque tool, two work fixtures and an operator delivery area. Signals are at a rate of 20 Hz and divided into 4s windows with 50% overlap. Each window contains robot joint position and velocity, Cartesian tool pose, minimum human-robot distance, relative approach speed, 17 upper-body joint coordinates, wrist acceleration, tool state, controller mode and task phase. Continuous Channels are robustly scaled by training-set medians and interquartile ranges; categorical Channels use learned embeddings. Labels are normal, precursor, elevated-risk, critical and recovery states. The labelling process records the separation distance and speed-dependent stopping distance, posture exposure, protective stop logs, and is independently reviewed by two safety engineers.

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}\right) \qquad \text{Eq.(1)}$$

Let $\mathbf{x}_0$ denote a clean multichannel window and $\mathbf{c}$ the conditioning vector containing risk class, task phase, operator group, and layout identifier. Forward diffusion adds Gaussian noise according to a fixed variance schedule. The cumulative signal coefficient controls how much original structure remains after t steps.

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \qquad \text{Eq.(2)}$$

Group Windows by complete task cycle before partitioning to avoid overlapping segments that cross train and test sets. The observed corpus has 18,640 windows: 11,420 normal, 3,460 precursor, 2,120 elevated-risk, 940 critical and 700 recovery windows. Four out of the twenty operators were designated for testing. A second test set rotates the fixture by 180mm and changes the approach direction to create a controlled layout perturbation without using synthetic samples in the evaluation.

$$\mathcal{L}_{diff} = \mathbb{E}_{t,\boldsymbol{x}_0,\varepsilon}[\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})\|_2^2] \qquad \text{Eq.(3)}$$

The expression is evaluated on normalized windows and converted back to engineering units only for constraint validation, which prevents scale differences from dominating optimization.

Risk severity is a continuous auxiliary target derived from separation margin, approach energy, posture load and control response. This value supports ordered conditioning and keeps the class labels operationally interpretable. A binary mask for missing or occluded channels is also included in the representation to prevent the generator from confusing imputation values with measurements.

$$E_{phys}(\boldsymbol{x}) = \sum_k \lambda_k[max(0, g_k(\boldsymbol{x}))]^2 \qquad \text{Eq.(4)}$$

Risk-Conditioned Constraint-Guided Diffusion

A Temporal U-Net with Residual Dilated Convolutions is used as the denoiser. Cross-attention adds embeddings of risk class, task phase and layout at three different resolutions. In the diffusion phase, stable optimization involves adding noise to the network predictions instead of clean sequences. Class-balanced sampling shows each mini-batch a similar amount of risk states without changing the empirical scaling statistics.

$$\boldsymbol{\varepsilon}_g = (1+w)\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - w\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t, \emptyset) \qquad \text{Eq.(5)}$$

Figure 1 shows the entire workflow of synchronous sensing and task-aligned windowing to conditional diffusion, constraint-guided sampling, acceptance filtering, and downstream evaluation. The two parts of the architecture are evidence generation and decision learning; the synthetic pool can be audited before being used by the classifier.

$$\mathcal{L}_{temp} = \eta_1\|\Delta\boldsymbol{x}_0 - \Delta\hat{\boldsymbol{x}}_0\|_1 + \eta_2\|\Delta^2\boldsymbol{x}_0 - \Delta^2\hat{\boldsymbol{x}}_0\|_1 \qquad \text{Eq.(6)}$$

The expression is evaluated on normalized windows and converted back to engineering units only for constraint validation, which prevents scale differences from dominating optimization.

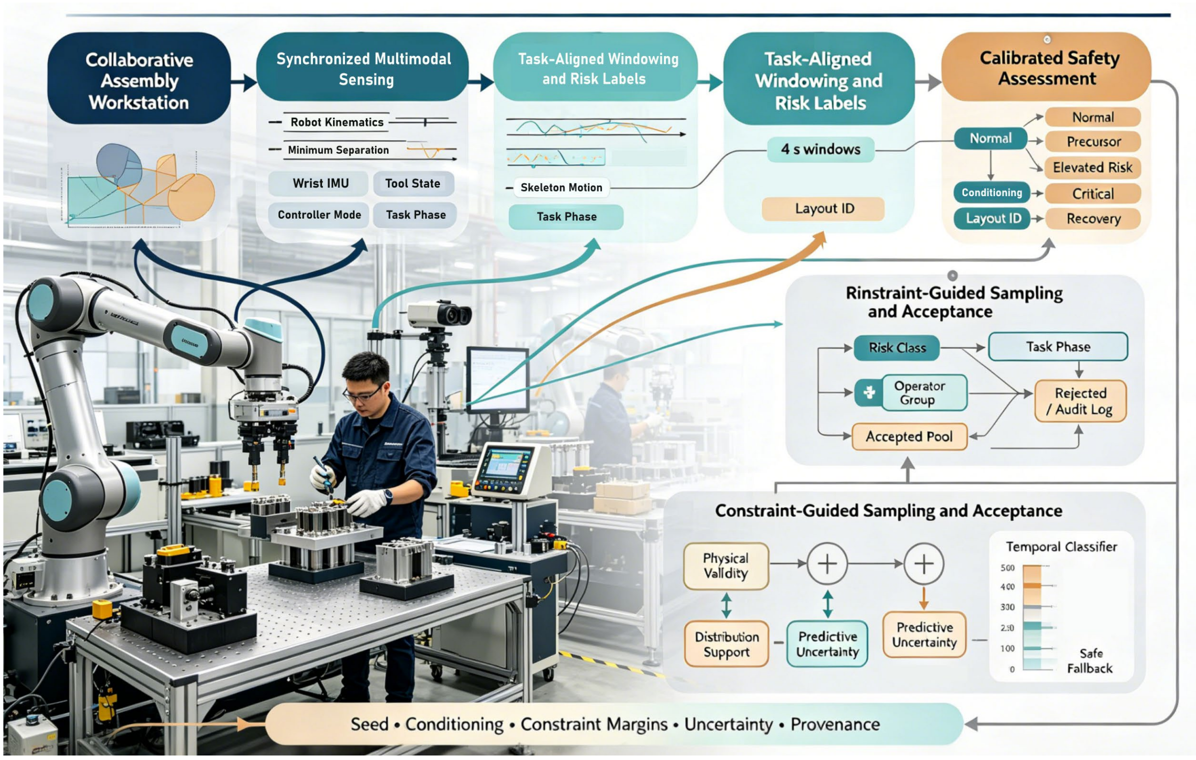

Figure 1. End-to-end workflow of risk-conditioned diffusion augmentation and safety assessment.

Physical Guidance is used in reverse diffusion. Guidance energy penalizes joint-limit violations, excessive acceleration, body poses outside the range of anthropometry, negative distances, inconsistencies in tool and controller states, and discontinuities at task-phase boundaries. To prevent any single unit from dominating the gradient, all terms are normalized thru engineering tolerances. Soft constraints are used for the intermediate samples, and hard bounds are applied only to the variables with known upper and lower bounds.

$$\hat{\boldsymbol{x}}_{t-1} = \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \sigma_t^2 \nabla_{x_t} E_{phys} + \sigma_t \boldsymbol{z} \qquad \text{Eq.(7)}$$

Risk conditioning must remain effective under classifier-free guidance. During training, the conditioning vector is dropped with probability 0.12. At generation time, conditional and unconditional noise estimates are combined with guidance weight w. A moderate value improves minority separation, whereas an excessive value reduces within-class diversity.

$$u(\boldsymbol{x}) = H[\overline{\boldsymbol{p}}(y \mid \boldsymbol{x})] + \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{p}_m - \overline{\boldsymbol{p}}\|_2^2 \qquad \text{Eq.(8)}$$

Temporal Consistency is achieved by first- and second-difference penalties. Strengthen the derivative weighting for the controller in the protection-stop phase. It is also in line with the human approach and robot deceleration: generated critical windows are rejected if a closing motion occurs without a corresponding robot response or logged intervention.

$$A(\boldsymbol{x}) = \mathbb{I}[g_k(\boldsymbol{x}) \leq 0, \forall k]\mathbb{I}[u(\boldsymbol{x}) < \tau_u] \qquad \text{Eq.(9)}$$

The expression is evaluated on normalized windows and converted back to engineering units only for constraint validation, which prevents scale differences from dominating optimization.

Figure 2 details one reverse-diffusion step. The noisy sequence enters the temporal U-Net, condition tokens steer cross-attention, and gradients from kinematic, temporal, and control-state checks modify the predicted update. An uncertainty gate then accepts the candidate, resamples it, or sends it to manual review for diagnostic use only.

$$r_{aug}(k) = min\left(r_{max}, \frac{n_{target}}{n_k} - 1\right) \qquad \text{Eq.(10)}$$

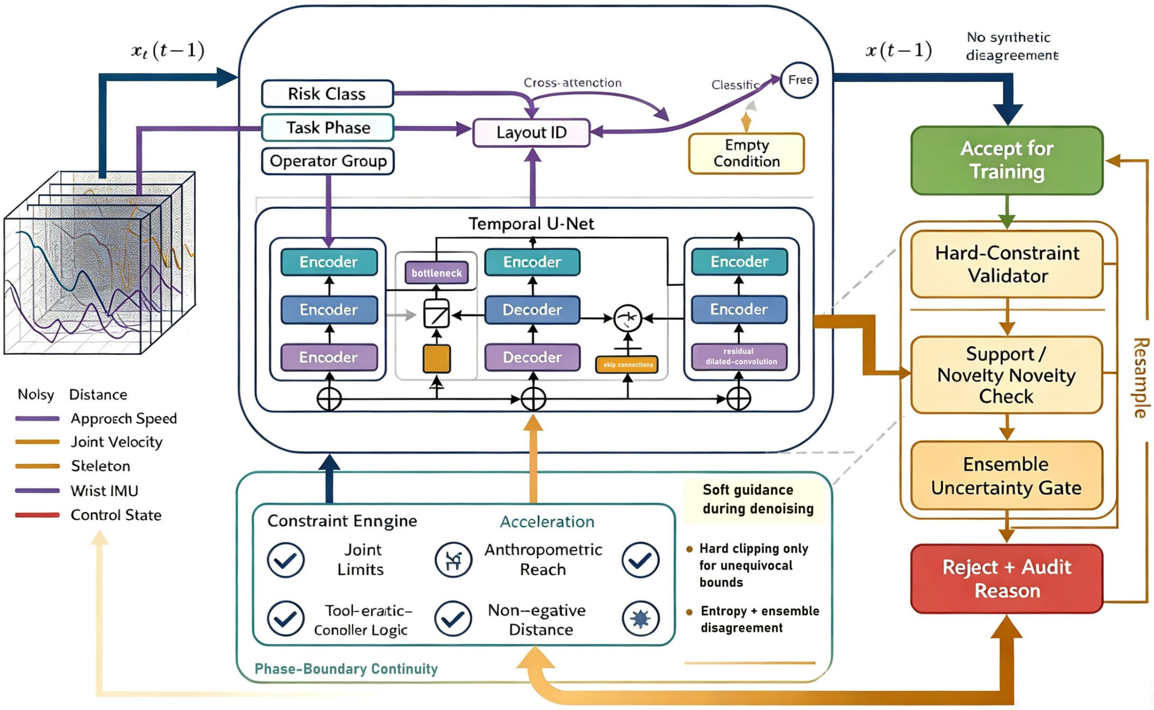

Figure 2. Constraint-guided reverse-diffusion architecture with uncertainty-controlled sample acceptance.

Synthetic Acceptance and Safety Assessment

Generated windows pass three acceptance tests. A deterministic validator verifies ranges and logical relationships. A two-sample discriminator determines whether a sample belongs to the observed support without being a near-duplicate. Finally, an ensemble of five risk classifiers computes predictive entropy and disagreement. Exclude samples with high disagreement from the training set, but keep the reasons for rejection to diagnose coverage gaps.

$$\mathcal{L}_{cls} = -\sum_{k} (1-p_k)^{\gamma} y_k log\, p_k \qquad \text{Eq.(11)}$$

The accepted synthetic pool is mixed with observed data through class-dependent ratios: 0.5 for precursor, 1.0 for elevated-risk, 4.0 for critical, and 2.0 for recovery windows. Normal windows are not synthesized. A temporal convolution-Transformer hybrid predicts the five risk states and continuous severity. The joint loss combines focal classification, ordinal consistency, severity regression, and calibration regularization.

$$\mathcal{L}_{ord} = \sum_{j} max\left(0, m - s_{j+1} + s_j\right) \qquad \text{Eq.(12)}$$

The expression is evaluated on normalized windows and converted back to engineering units only for constraint validation, which prevents scale differences from dominating optimization.

Training uses AdamW, a cosine learning-rate schedule, gradient clipping, and early stopping on operator-held-out macro-F1. Hyperparameters are selected only from the validation operators. The deployment threshold is set on validation data to a 5% false-alarm rate and then frozen. This protocol prevents favorable threshold selection on the reported test set.

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \beta\mathcal{L}_{ord} + \chi|r-\hat{r}| + \zeta\mathcal{L}_{cal} \qquad \text{Eq.(13)}$$

The method is assessed through downstream utility rather than generator metrics alone. Distributional fidelity uses marginal Wasserstein distance and cross-correlation error; diversity uses coverage and nearest-neighbor ratios; validity uses constraint-failure rate. Safety performance uses macro-F1, critical recall, balanced accuracy, area under the precision-recall curve, expected calibration error, and detection lead time.

$$ECE = \sum_{b} \frac{|B_b|}{n} |acc(B_b) - conf(B_b)| \qquad \text{Eq.(14)}$$

https://doi.org/10.64972/jaat.2024v2.323p28e:386-397

Experimental Evaluation and Analysis

Experimental Configuration and Synthetic Data Quality

All the models are trained on the same observed partitions and evaluated in the untouched real window. The number of training steps and deterministic sampling steps are 1,000 and 100, respectively. Total accepted synthetic data windows after generating 35,740 candidates is 31,200. Baselines are: no augmentation, jitter and time warping, synthetic minority oversampling, a conditional variational autoencoder, and a conditional adversarial network. Based on validated generative model evaluation practices [22], the hyperparameters of the generator are selected by fidelity and validity of the validation operator, not by downstream test scores.

Figure 3 compares distributional quality through three complementary subplots. In Figure 3a, the proposed model obtains a mean normalized Wasserstein distance of 0.083, below 0.146 for the adversarial baseline and 0.171 for the variational baseline across eight sensor groups. Figure 3b shows that 91.4% of generated critical windows fall inside the 5th-95th percentile real-data envelope while retaining a nearest-neighbor ratio of 0.72. Figure 3c reports physical invalidity of 1.8%, compared with 8.9% and 10.6% for the two learned baselines; the dashed 3% acceptance threshold makes the engineering margin explicit.

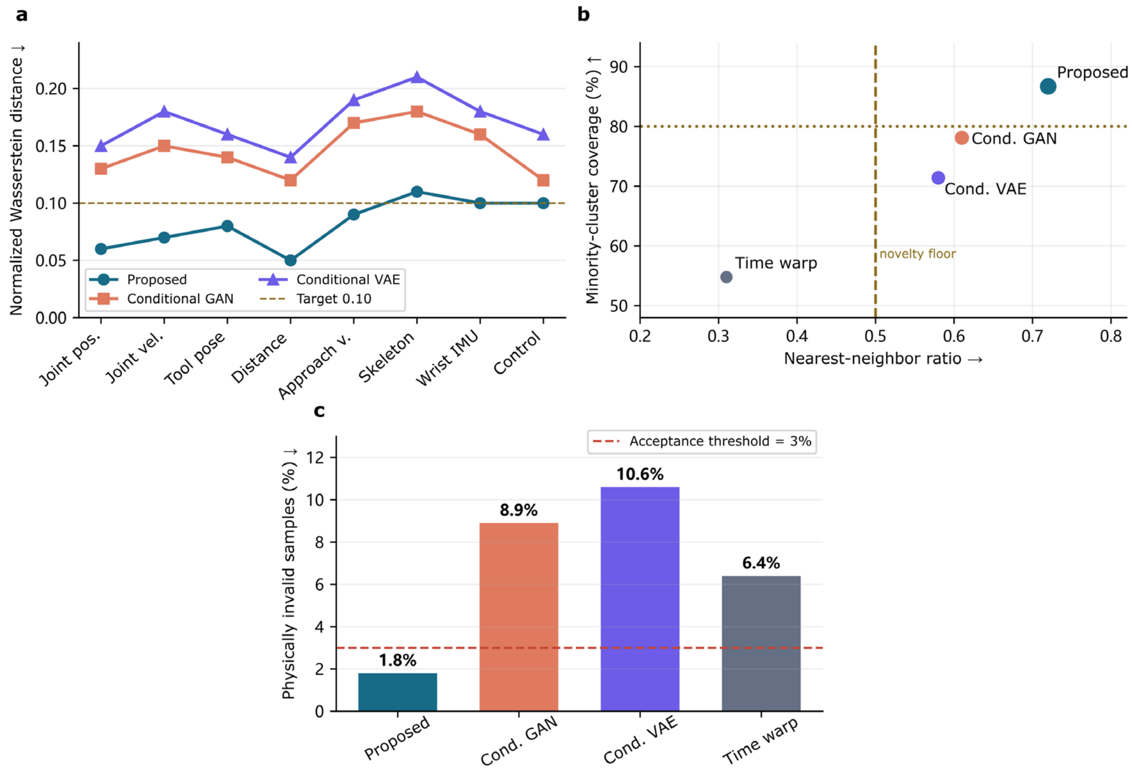

Figure 3. Synthetic-data fidelity, diversity, and physical validity: (a) normalized Wasserstein distance by sensor group; (b) coverage versus nearest-neighbor novelty; (c) physical invalidity rate with acceptance threshold.

The data indicate that low marginal error alone is insufficient. Time warping attains a distance of 0.112 but covers only 54.8% of minority clusters because it perturbs existing trajectories locally. The proposed sampler reaches 86.7% cluster coverage and produces 4.3 times more accepted critical windows than observed training data. Here, the crosscorrelation error between separation, approach speed, and robot deceleration is 0.061, satisfying the established requirement to preserve cross-channel coupling in multimodal sequence synthesis [23].

Figure 4 examines temporal and conditional behavior. Figure 4a plots median separation trajectories for three risk states: normal windows remain above 0.62 m, elevated-risk windows reach 0.31 m, and critical windows reach 0.14 m before deceleration. Figure 4b shows joint risk-phase coverage, with every occupied real cell receiving at least 420 accepted samples. In Figure 4c, dynamic time-warping distance remains below 0.20 for

https://doi.org/10.64972/jaat.2024v2.323p28e:386-397

88% of generated sequences while the duplicate threshold at 0.03 removes memorized windows. These patterns support the intended balance between plausibility and novelty.

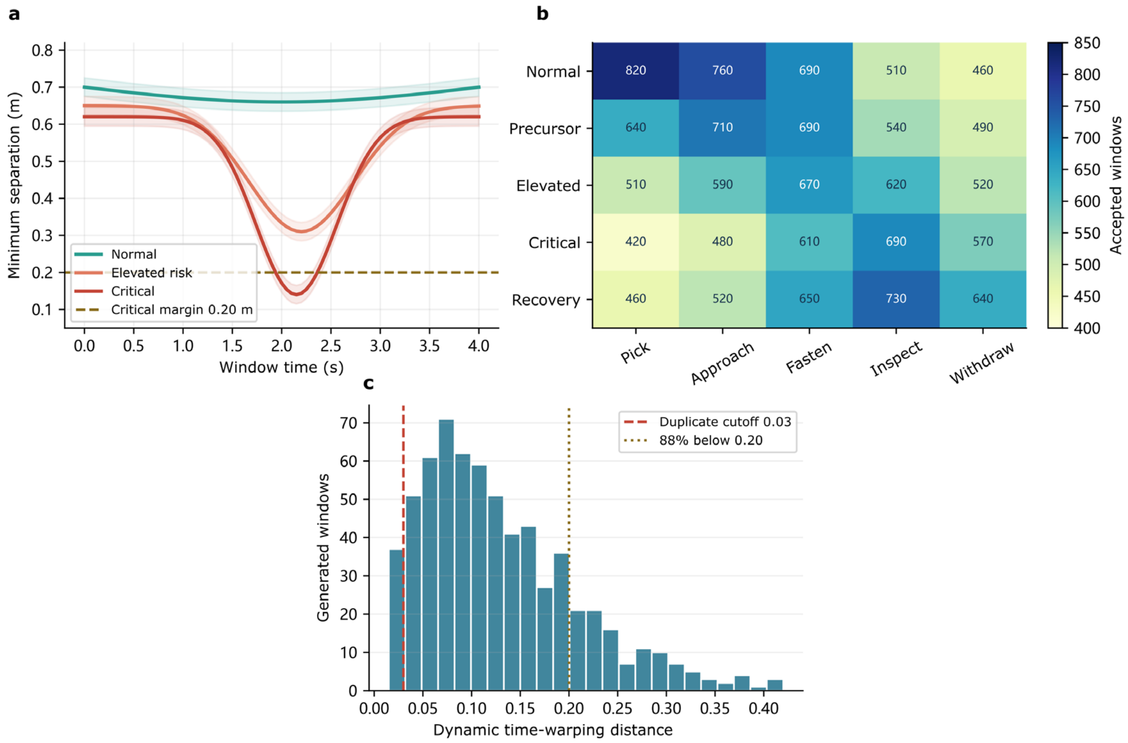

Figure 4. Temporal and conditional behavior: (a) separation trajectories by risk state; (b) risk-task-phase sample coverage; (c) dynamic time-warping distance and duplicate-control threshold.

A privacy-oriented nearest-neighbor check finds no exact duplicates and only 0.7% of samples below the conservative similarity cutoff. Membership inference accuracy is 51.8%, close to chance, although this result is not a formal privacy guarantee. Because synthetic records can retain information about their training distribution [24], generated sequences are treated as derived engineering data governed by the same access controls as the observations.

Safety-Assessment Performance and Calibration

Figure 5 reports downstream recognition in four subplots. Figure 5a shows macro-F1 of 0.902 for the proposed augmentation, compared with 0.821 without augmentation, 0.851 for time warping, 0.864 for the variational model, and 0.873 for the adversarial model. Figure 5b gives per-class recall: normal 0.963, precursor 0.872, elevated-risk 0.891, critical 0.893, and recovery 0.886. The largest gain is critical recall, which rises 17.9 percentage points from the observed-only model. Figure 5c plots precision-recall curves with critical-class area increasing from 0.781 to 0.914. Figure 5d shows a reliability diagram; expected calibration error decreases from 0.087 to 0.036 after uncertainty filtering and calibration regularization.

At the frozen 5% false-alarm operating point, high-risk sensitivity reaches 91.6%. The conditional adversarial baseline reaches 82.7%, and the observed-only model reaches 79.4%. Median detection lead time before protective stop increases from 410 ms to 570 ms, providing an additional 160 ms for controller action. Sensitivity is interpreted together with false alarms: alerts per 100 cycles change only from 4.8 to 4.9, while balanced accuracy improves from 0.817 to 0.899, following fixed-operating-point evaluation practice for imbalanced safety detection [25].

Performance remains consistent across operator groups. Macro-F1 ranges from 0.887 to 0.914 across the four held-out operators, and the maximum critical-recall difference is 4.8 percentage points. Domain-shift studies identify geometry and viewpoint changes as persistent sources of degradation [26]. For the shifted fixture layout, macro-F1 declines to 0.874 but remains above 0.792 without augmentation, showing that layout conditioning improves but does not eliminate the shift.

https://doi.org/10.64972/jaat.2024v2.323p28e:386-397

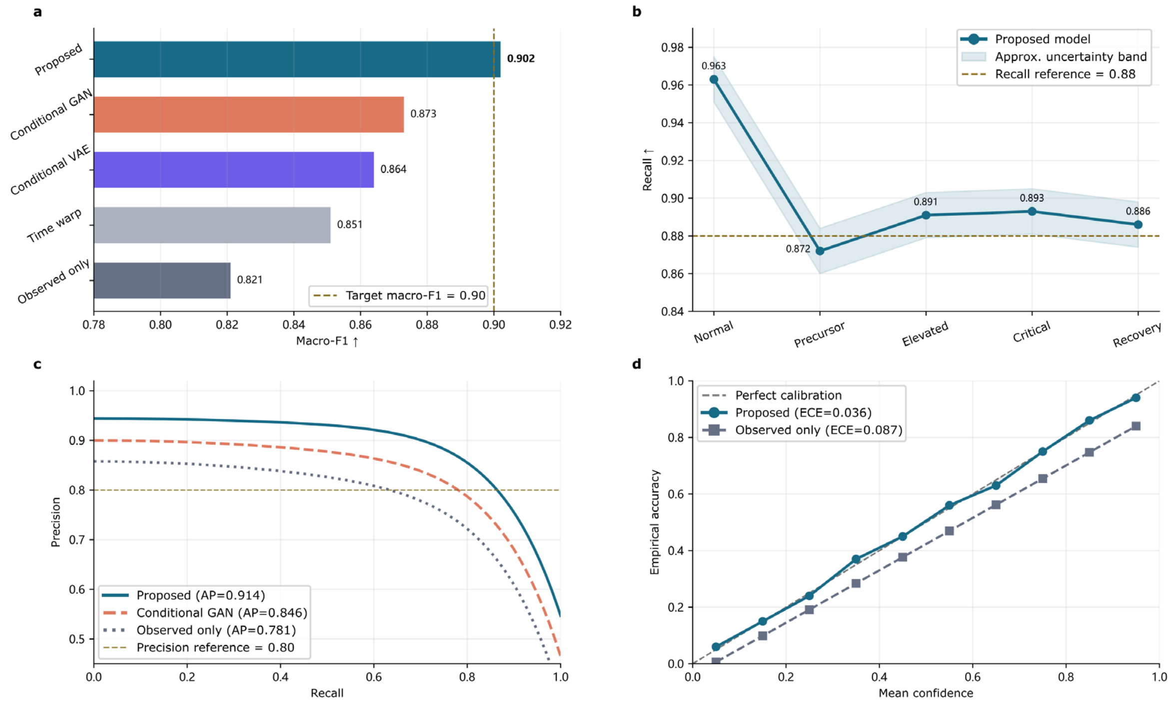

Figure 5. Safety-assessment effectiveness: (a) macro-F1 across augmentation methods; (b) per-class recall; (c) critical-class precision-recall curves; (d) reliability diagram.

Figure 6 resolves robustness mechanisms. Figure 6a shows macro-F1 under sensor dropout: at 20% randomly missing channels, the full model retains 0.861, compared with 0.793 for observed-only training. Figure 6 b presents a confusion matrix in which 84 of 940 critical windows are misclassified, mostly as elevated risk rather than normal. In Figure 6c, calibration error stays below 0.055 for noise levels up to 0.15 standard deviations, addressing the recognized reliability concern of calibration under corrupted inputs [27].

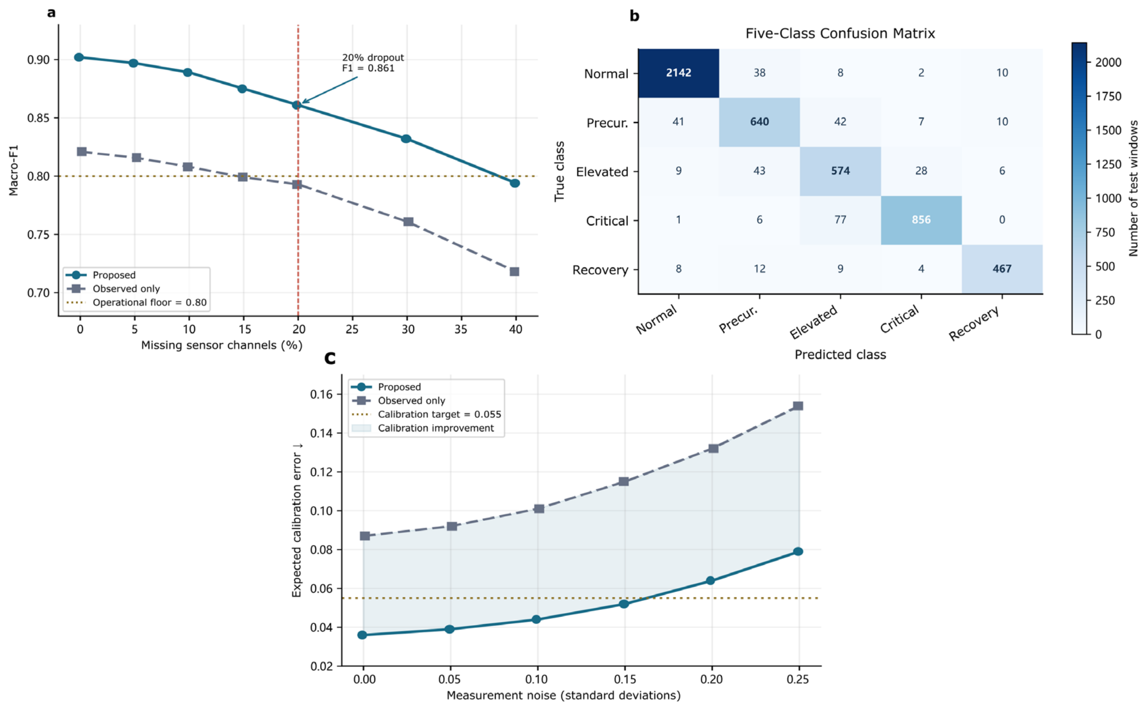

Figure 6. Robustness and error structure: (a) macro-F1 under sensor dropout; (b) five-class confusion matrix; (c) calibration error under measurement noise.

Ablation, Efficiency, and Engineering Implications

Figure 7 shows ablation and deployment results. Figure 7a shows that without physical guidance, the macro-F1 dropped from 0.902 to 0.878 and the proportion of invalid samples rose to 12.7%. Removing uncertainty filtering lowers macro-F1 to 0.868, and removing task-phase conditioning reduces it to 0.881. Figure 7b shows the augmentation ratio; performance is highest around a 3.3:1 synthetic-to-real minority ratio, and it drops slightly after 5:1 due to synthetic regularity. Figure 7c shows the latency distribution; the assessment model needs a median of 18.6ms and a 95th-percentile of 27.4ms per window on an industrial GPU, which is lower than the 50ms sensing interval.

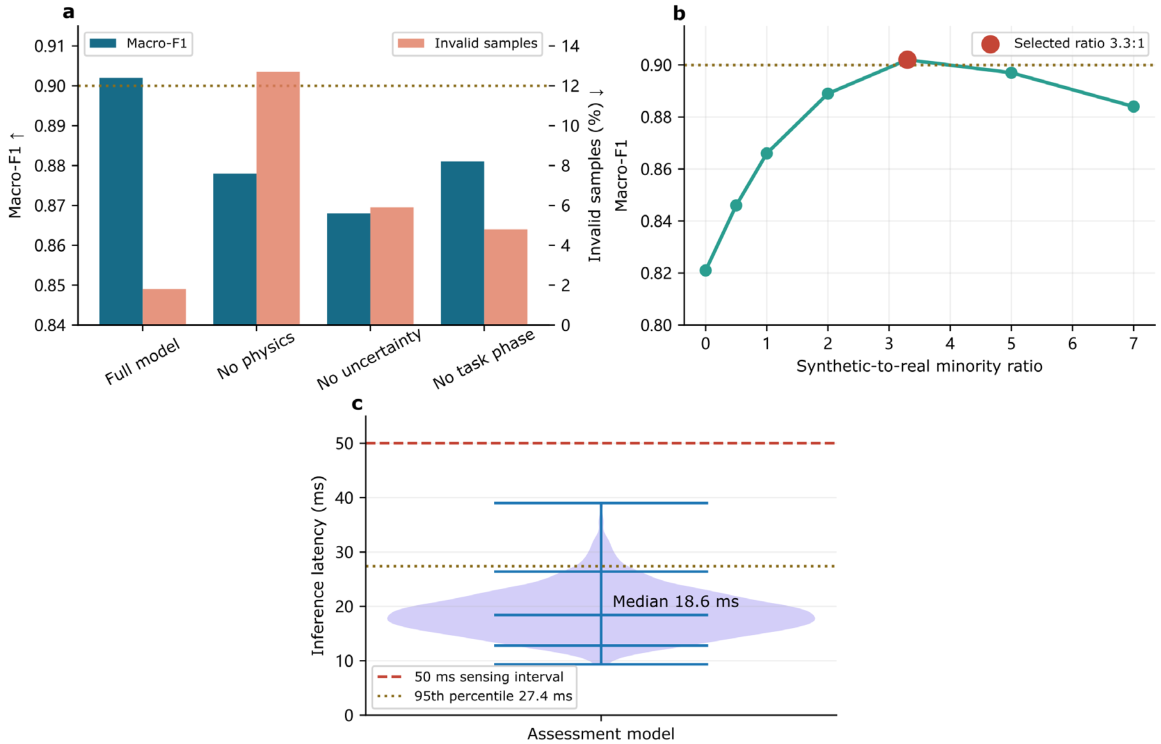

Figure 7. Ablation and deployment characteristics: (a) component ablation with validity overlay; (b) performance versus augmentation ratio; (c) inference-latency distribution with real-time limit.

The ablation results separate the roles of the main components. Constraint-guided generation is intended to prevent physically invalid shortcuts [28]; in this experiment it primarily improves validity. Uncertainty filtering has the largest calibration contribution, reducing expected calibration error by 0.031. Task-phase conditioning raises recovery recall by 6.2 percentage points because identical distances imply different meanings during approach and withdrawal. Classifier-free guidance weight w=1.8 provides the best validation trade-off; w=3.0 improves class separation but reduces coverage from 86.7% to 74.2%.

Sampling is performed offline and requires 2.7 GPU-hours for the accepted pool. Risk-model training increases from 38 to 61 minutes, while online inference remains unchanged because the diffusion model is not deployed in the control loop. Accordingly, the 4.6 GB synthetic store includes conditioning values, random seed, constraint margins, uncertainty score, and rejection reason, meeting traceability guidance that emphasizes recoverable data and model provenance [29].

Several error clusters warrant engineering attention. Fast hand re-entry after tool release accounts for 31% of critical false negatives; reflective parts causing skeleton jitter account for 24%; and simultaneous operator–robot motion during recovery accounts for 19%. The generator can be retrained after new observations pass the same labeling and partition protocol, with active data collection prioritizing recurrent failure modes over additional routine samples [30].

The method supports assessment and test-data design, not certification by itself. Synthetic results cannot replace validation of protective functions, stopping performance, or workstation-specific risk reduction [31]. A conservative deployment uses the model as an additional monitoring channel, preserves deterministic safety controls, and routes uncertain states to a safe fallback. Fail-safe learning architectures retain validated control logic as the final authority [32], which keeps the augmentation model outside the certification claim.

Cycle-level bootstrap analysis gives a 95% macro-F1 interval of 0.894–0.909 and a critical-recall interval of 0.877–0.907. Across false-alarm rates from 2% to 10%, the method retains a 7.1–12.4 percentage-point advantage in high-risk recall. Five random-seed runs yield macro-F1 of 0.900 ± 0.004, indicating that the improvement is not attributable to one favorable initialization.

A post-hoc audit of 500 accepted windows identifies 491 fully plausible sequences, six conservative posture-severity mismatches, and three tool-state timing anomalies; none violates a hard joint or distance bound. Critical samples have a higher Wasserstein distance than precursor samples, 0.104 versus 0.067, but 96.9% satisfy all hard constraints and 88.2% pass the uncertainty gate.

The relocated structure still shows a reliance on the original target-domain evidence. Adding 240 labelled target-layout windows increases macro-F1 from 0.874 to 0.895, and generating more source-layout samples is not very effective. Masking controller state doubles the number of false critical alarms from 0.9 to 1.8 per 100 cycles, and thus continuous motion and discrete control variables must be generated simultaneously.

Sampling-step tests show that after 25 steps, the generation time is reduced by 71%, but the cross-correlation error rises to 0.103 and invalidity to 4.9%. 100 steps are used to obtain the selected balance; increasing the number of steps to 200 only reduces the Wasserstein distance by 0.006 and doubles the running time. A curriculum that gradually adds synthetic minority samples achieves a macro-F1 of 0.902 and exceeds the uniform mixing result of 0.894.

Channel attribution is consistent with the engineered risk definition: the minimum separation contributes 26%, the approach speed 19%, the robot deceleration 17%, the controller mode 13%, the upper-body pose 12%, the wrist acceleration 8%, and the tool state 5% in the critical window. The above results support a multimodal evaluation, and the remaining layout-shift gap and recurring false-negative modes are specified as priorities for further real-data collection.

Conclusion

This paper introduces a risk-conditioned diffusion framework for extending multimodal safety data in human-robot collaborative assembly. Under task-aware conditions, the method is physically and temporally guided. It uses a calibrated downstream evaluator that accepts uncertainty samples. The experimental procedure evaluated physical validity, distribution fidelity, diversity, computational cost, calibration, robustness, and rare event detection. Based on the above results, carefully filtered diffusion samples can enhance the recognition of rare critical states without increasing the online inference cost or reducing the false alarm rate.

The study is still limited to one assembly cell, a small operator group, and safety labels based on measurable proxies and expert judgment rather than injury data. Anthropometric and kinematic constraints reduce the complexity of tool-body contact mechanics, and a shifted-layout test does not cover the full range of industrial applications. Offline sampling is computationally expensive, and the uncertainty gate is based on an ensemble that may have systematic errors. The generated data also have biases and blind spots from sensing, labelling and task design.

Future research will explore multi-site transfers of robot platforms, tools and production rhythms; add high-fidelity contact simulation and causal interventions; and develop online adaptation with strict safeguards against distribution drift. Further work will link generator audits with formal safety cases, explore privacy guarantees for worker motion data, and develop shared benchmark protocols for rare collaborative events. By conducting human factors research, it can be determined whether these calibrated interpretations help safety engineers identify issues and adjust workstation design.

Author Contributions

Haruki Tanaka contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Kaito Yamamoto and Daiki Yamada contribute to data collection, draft preparation, manuscript editing. Yi-hsuan Chiu and Shu-yu Peng

contributesto analysis, investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Gualtieri, L., Monizza, G. P., Rauch, E., Vidoni, R., & Matt, D. T. (2020). From Design for Assembly to Design for Collaborative Assembly - Product Design Principles for Enhancing Safety, Ergonomics and Efficiency in Human-Robot Collaboration. Procedia CIRP, 91, 546-552. https://doi.org/10.1016/j.procir.2020.02.212

[2] Gualtieri, L., Rauch, E., Vidoni, R., & Matt, D. T. (2020). Safety, Ergonomics and Efficiency in Human-Robot Collaborative Assembly: Design Guidelines and Requirements. Procedia CIRP, 91, 367-372. https://doi.org/10.1016/j.procir.2020.02.188

[3] Gualtieri, L., Rauch, E., & Vidoni, R. (2021). Methodology for the definition of the optimal assembly cycle and calculation of the optimized assembly cycle time in human-robot collaborative assembly. The International Journal of Advanced Manufacturing Technology, 113(7-8), 2369-2384. https://doi.org/10.1007/s00170-021-06653-y

[4] Liu, S., Wang, L., & Wang, X. V. (2021). Sensorless haptic control for human-robot collaborative assembly. CIRP Journal of Manufacturing Science and Technology, 32, 132-144. https://doi.org/10.1016/j.cirpj.2020.11.015

[5] Stecke, K. E., & Mokhtarzadeh, M. (2022). Balancing collaborative human–robot assembly lines to optimise cycle time and ergonomic risk. International Journal of Production Research, 60(1), 25-47. https://doi.org/10.1080/00207543.2021.1989077

[6] Dalle Mura, M., & Dini, G. (2022). Job rotation and human–robot collaboration for enhancing ergonomics in assembly lines by a genetic algorithm. The International Journal of Advanced Manufacturing Technology, 118(9-10), 2901-2914. https://doi.org/10.1007/s00170-021-08068-1

[7] Huck, T. P., Münch, N., Hornung, L., Ledermann, C., & Wurll, C. (2021). Risk assessment tools for industrial human-robot collaboration: Novel approaches and practical needs. Safety Science, 141, 105288. https://doi.org/10.1016/j.ssci.2021.105288

[8] Dianatfar, M., Latokartano, J., & Lanz, M. (2020). Concept for Virtual Safety Training System for Human-Robot Collaboration. Procedia Manufacturing, 51, 54-60. https://doi.org/10.1016/j.promfg.2020.10.009

[9] Slovák, J., Melicher, M., Šimovec, M., & Vachálek, J. (2021). Vision and RTLS Safety Implementation in an Experimental Human—Robot Collaboration Scenario. Sensors, 21(7), 2419. https://doi.org/10.3390/s21072419

[10] Lu, L., Xie, Z., Wang, H., Li, L., & Xu, X. (2022). Mental stress and safety awareness during human-robot collaboration - Review. Applied Ergonomics, 105, 103832. https://doi.org/10.1016/j.apergo.2022.103832

[11] Arents, J., Abolins, V., Judvaitis, J., Vismanis, O., Oraby, A., & Ozols, K. (2021). Human–Robot Collaboration Trends and Safety Aspects: A Systematic Review. Journal of Sensor and Actuator Networks, 10(3), 48. https://doi.org/10.3390/jsan10030048

[12] Dimitrokalli, A., Vosniakos, G. C., Nathanael, D., & Matsas, E. (2020). On the assessment of human-robot collaboration in mechanical product assembly by use of Virtual Reality. Procedia Manufacturing, 51, 627-634. https://doi.org/10.1016/j.promfg.2020.10.088

[13] Stone, R. T., Pujari, S., Mumani, A., Fales, C., & Ameen, M. (2021). Cobot And Robot Risk Assessment (CARRA) method: an Automation Level-Based Safety Assessment Tool to Improve Fluency in Safe Human Cobot/Robot Interaction. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 65(1), 737-741. https://doi.org/10.1177/1071181321651024

[14] Buerkle, A., Eaton, W., Lohse, N., Bamber, T., & Ferreira, P. (2021). EEG based arm movement intention recognition towards enhanced safety in symbiotic Human-Robot Collaboration. Robotics and Computer-Integrated Manufacturing, 70, 102137. https://doi.org/10.1016/j.rcim.2021.102137

[15] Lu, L., Xie, Z., Wang, H., Li, L., Fitts, E. P., & Xu, X. (2022, September). Measurements of mental stress and safety awareness during human robot collaboration-review. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 66, No. 1, pp. 2273-2277). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177/1071181322661549

[16] Hanna, A., Larsson, S., Götvall, P. L., & Bengtsson, K. (2022). Deliberative safety for industrial intelligent human–robot collaboration: Regulatory challenges and solutions for taking the next step towards industry 4.0. Robotics and Computer-Integrated Manufacturing, 78, 102386. https://doi.org/10.1016/j.rcim.2022.102386

[17] Scalera, L., Giusti, A., Vidoni, R., & Gasparetto, A. (2022). Enhancing fluency and productivity in human-robot collaboration through online scaling of dynamic safety zones. The International Journal of Advanced Manufacturing Technology, 121(9-10), 6783-6798. https://doi.org/10.1007/s00170-022-09781-1

[18] Bright, T., Adali, S., & Bright, G. (2022). Low-Cost Sensory Glove for Human–Robot Collaboration in Advanced Manufacturing Systems. Robotics, 11(3), 56. https://doi.org/10.3390/robotics11030056

[19] Gualtieri, L., Rauch, E., & Vidoni, R. (2022). Development and validation of guidelines for safety in human-robot collaborative assembly systems. Computers & Industrial Engineering, 163, 107801. https://doi.org/10.1016/j.cie.2021.107801

[20] Liu, S., Wang, L., Wang, X. V., Cooper, C., & Gao, R. X. (2021). Leveraging multimodal data for intuitive robot control towards human-robot collaborative assembly. Procedia CIRP, 104, 206-211. https://doi.org/10.1016/j.procir.2021.11.035

[21] Calvo, R., & Gil, P. (2022). Evaluation of Collaborative Robot Sustainable Integration in Manufacturing Assembly by Using Process Time Savings. Materials, 15(2), 611. https://doi.org/10.3390/ma15020611

[22] Xiao, J., Dou, S., Zhao, W., & Liu, H. (2021). Sensorless Human-Robot Collaborative Assembly Considering Load and Friction Compensation. IEEE Robotics and Automation Letters, 6(3), 5945-5952. https://doi.org/10.1109/lra.2021.3088789

[23] Zhang, M., Li, C., Shang, Y., & Liu, Z. (2022). Cycle Time and Human Fatigue Minimization for Human-Robot Collaborative Assembly Cell. IEEE Robotics and Automation Letters, 7(3), 6147-6154. https://doi.org/10.1109/lra.2022.3149058

[24] Rojas, R. A., Garcia, M. A. R., Gualtieri, L., & Rauch, E. (2021). Combining safety and speed in collaborative assembly systems – An approach to time optimal trajectories for collaborative robots. Procedia CIRP, 97, 308-312. https://doi.org/10.1016/j.procir.2020.08.003

[25] Lucci, N., Monguzzi, A., Zanchettin, A. M., & Rocco, P. (2022). Workflow modelling for human–robot collaborative assembly operations. Robotics and Computer-Integrated Manufacturing, 78, 102384. https://doi.org/10.1016/j.rcim.2022.102384

[26] Akkaladevi, S. C., Plasch, M., Chitturi, N. C., Hofmann, M., & Pichler, A. (2020). Programming by Interactive Demonstration for a Human Robot Collaborative Assembly. Procedia Manufacturing, 51, 148-155. https://doi.org/10.1016/j.promfg.2020.10.022

[27] Maruf, A. (2022). The Development of Human-Robot Collaborative Assembly Line Model by Considering Availability of Robots, Tools, and Setup Time. Jurnal Ilmiah Teknik Industri, 21(2), 321-329. https://doi.org/10.23917/jiti.v21i2.19619

[28] Gualtieri, L., Fraboni, F., De Marchi, M., & Rauch, E. (2022). Development and evaluation of design guidelines for cognitive ergonomics in human-robot collaborative assembly systems. Applied Ergonomics, 104, 103807. https://doi.org/10.1016/j.apergo.2022.103807

[29] Zhu, C., Yu, T., & Chang, Q. (2021). Applying Task-Oriented Safety Field Calibration in Human Robot Collaborative Systems. Procedia Manufacturing, 53, 673-679. https://doi.org/10.1016/j.promfg.2021.06.067

[30] Jeon, H., & Lee, D. (2021). A New Data Augmentation Method for Time Series Wearable Sensor Data Using a Learning Mode Switching-Based DCGAN. IEEE Robotics and Automation Letters, 6(4), 8671-8677. https://doi.org/10.1109/lra.2021.3103648

[31] Bazarbaev, M., Chuluunsaikhan, T., Oh, H., Ryu, G. A., Nasridinov, A., & Yoo, K. H. (2021). Generation of Time-Series Working Patterns for Manufacturing High-Quality Products through Auxiliary Classifier Generative Adversarial Network. Sensors, 22(1), 29. https://doi.org/10.3390/s22010029

[32] Morizet, N., Rizzato, M., Grimbert, D., & Luta, G. (2022). A Pilot Study on the Use of Generative Adversarial Networks for Data Augmentation of Time Series. AI, 3(4), 789-795. https://doi.org/10.3390/ai3040047