Assessing Procedural Content Generation in Reinforcement Learning Prompt-Based Feedback via Adaptive Virtual Reality Industrial Environments for Workforce Training

Joshua Hatfield*, Husnu S. Narman*, Sudipta Chowdhury†, Ammar Alzarrad‡
* Department of Computer Science {hatfield308, narman}@marshall.edu
† Department of Mechanical & Industrial Engineering, chowdhurys@marshall.edu
‡Department of Civil Engineering, alzarrad@marshall.edu

***Abstract*—Industrial training for hydroelectric maintenance requires learners to interpret the environment around them in real-time to make decisions in order to apply safety, regulations, and effective intervention under pressure. Traditional training approaches in a virtual reality environment fail to capitalize on the user's own performance by not adapting to their training skills. This work presents an adaptive virtual reality environment that utilizes a reinforcement learning (RL) feedback agent to deliver context-specific prompts during maintenance scenarios. Of the studies that choose to evolve their static environment to an adaptive one, there is little work that systematically isolates design choices such as user skill tiers. It works by extracting state information from user interactions within a detailed hydroelectric plant simulation and uses a lightweight policy model to determine when and how feedback should be presented to the user. Prompt quality is then assessed using programmatic metrics for clarity, specificity, safety, actionability, and domain accuracy, as well as an overall scalar representative of both task performance and text quality. We compare three feedback prompts under the following circumstances: (1) adaptive RL prompts, (2) random prompts, (3) static, hand-crafted prompts, and run a set of ablation studies that disable axis adaption, tier adaption, and individual reward components. Across simulated trials, the fully adaptive Proximal Policy Optimization (PPO) configuration achieves higher overall reward and domain targeting.**

***Index Terms*—virtual reality training, reinforcement learning, ablation study, adaptive feedback, hydroelectric maintenance**

I. Introduction

Hydroelectric power plants are stable sources of energy in the United States. Specifically, they have produced more than a quarter of the total renewable energy in 2022 for the United States in total [1]. In an industrious environment, it is crucial that workers are equipped with knowledge so that they can perform maintenance tasks while adhering to safety regulations to ensure an efficient, productive, and secure decision making process when faced with scenarios that require worker intervention. A study published by the National Library of Medicine focused on hydroelectric workers showed one of the most common risks to worker safety is lack of training [2]. This furthers the need for training in a hydroelectric plant environment specifically. This study introduces an adaptive VR training system to improve workplace safety. Traditional training methods lack adaptability to user performance. However, we find that adaptive environments can prove to also be ineffective when managed poorly.

In particular, the *objective* of this paper is to evaluate the quality of prompt-based feedback with a simulated human-in-the-loop process via metrics such as clarity, specificity, safety, actionability, and domain accuracy. The *key contributions* of this paper are:

- (i) By systematically comparing adaptive RL prompts against static and random baselines, this work highlights the limitations of traditional training approaches.
- (ii) This work also establishes that adaptive, policy-driven feedback provides measurable improvements in clarity, specificity, safety, and domain accuracy.
- (iii) The study also shows that a lightweight Proximal Policy Optimization (PPO) model can outperform both static (predefined) and randomly generated prompts. This validates PPO as a strong candidate for reinforcement learning–driven adaptive feedback in training environments.

The rest of the paper is organized as follows: In Section II, we present a literature review that evaluates prior work in the field. Section III describes the simulation environment and outlines the human interaction metrics. In Section V, we detail the feedback model and its associated reward criteria. Section VI reports the outcomes of an ablation study on the reinforcement learning model's weights, verifying its effectiveness and highlighting the potential of adaptive learning environments. Section VII provides a brief discussion of the results and their limitations. Finally, Section VIII concludes the paper with closing remarks and directions for future work.

To improve accessibility for a broader audience, we briefly define key concepts used throughout this paper. Reinforcement learning (RL) is a machine learning approach where an agent learns by trial and error, receiving rewards for good decisions. Proximal Policy Optimization (PPO) is a widely used RL algorithm that updates its decision-making strategy gradually to ensure stable learning. An ablation study refers to

systematically removing parts of a system to understand how each component contributes to overall performance. A tiered proficiency model categorizes users based on skill level (e.g., novice, intermediate, expert) and adjusts feedback accordingly. Finally, axis adaptation refers to dynamically selecting which part of a task (e.g., safety, execution, or cleanup) should receive feedback based on user performance.

II. Related Work

There are various studies that exist to inspect the intersection between RL and VR. In particular, they are shown in various high stakes scenarios with contextually rich simulations which stem from intricate surgical procedures to emergency response protocols. However, literature shows that there is conflicting results when applying adaptive training to virtual reality environments and requires further investigation [3], [4].

A. VR Training & Adaptive Feedback

Virtual reality has become an effective medium for training complex skills, particularly when combined with adaptive feedback mechanisms [3], [4]. Prior work emphasizes the importance of learner modeling, task adaptation, and real-time feedback to improve skill acquisition and efficiency [5], [6]. More recent systems integrate AI-driven agents, digital twins, and conversational interfaces to further enhance personalization and realism in training environments [7]–[9].

B. Reinforcement Learning for Instruction, Feedback, and Procedural Content Generation

Reinforcement learning (RL) has emerged as a powerful mechanism for adapting both instructional content and task difficulty in interactive environments. In the context of procedural content generation (PCG), Khalifa et al. formalize PCGRL, casting level generation itself as an RL problem where an agent learns to construct environments that satisfy design constraints and gameplay criteria [10]. Lopez et al. extend this line of work to 3D virtual environments: their early study uses RL to generate VR content tailored to design objectives [11], while subsequent work applies deep RL to PCG for 3D virtual environments, demonstrating how learned generators can produce diverse yet solvable spaces [12]. Gisslén et al. propose adversarial RL for PCG, pairing a generator agent with a solver agent so that training yields environments that are challenging but solvable, with auxiliary control inputs allowing designers to steer difficulty and style [13]. Mahmoudi-Nejad et al. further show that experience-driven PCG via RL can personalize therapeutic content for arachnophobia exposure therapy, adapting virtual spiders to patient-specific physiological responses [14]. Agarwal and Shridevi apply RL-based PCG to disaster evacuation training, using RL to generate evacuation scenarios in a 3D VR environment that better exercise trainees' decision-making under uncertainty [15]. More recently, Joshi introduces an RL-enhanced Wave Function Collapse framework for dynamic, narrative-driven AR experiences, where tile weights are adjusted via RL to reflect gameplay context and narrative needs [16].

C. Evaluation and Ablation in ML/RL Systems

Robust evaluation of machine learning and reinforcement learning systems increasingly relies on systematic ablation studies that isolate the contribution of individual components. Prior work shows that RL performance is sensitive to reward design, model structure, and training configuration, making component-level evaluation essential for understanding system behavior [17]–[20].

III. Task and Simulation Setup

Figure 1 presents the system as a sequential pipeline, from trainee interaction to feedback generation and logging.

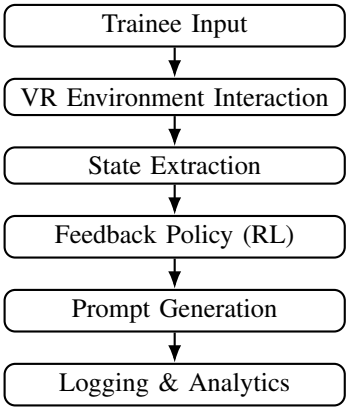

Fig. 1. Pipeline view of the proposed VR training system. User actions in the virtual environment are converted into state information, processed by the RL feedback policy, transformed into prompts, and logged for analysis. Arrows indicate sequential data flow.

A. Hydroelectric Abstraction

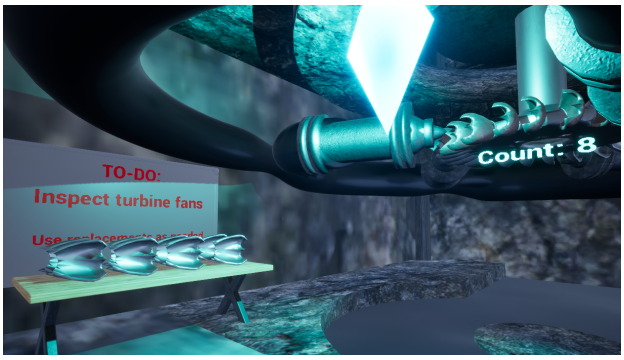

Fig. 2. The fan blade replacement task environment within the virtually realized hydroelectric plant

For this model, the user is given an interactable environment (see Fig. 2) in which they are given the option to perform safety features before and after a given repair scenario. Specifically, they are not required, but highly encouraged, to act out these safety precautions in order to receive a maximal return on their performance metrics. Failing to perform these tasks in a safe order will reduce their performance metrics and thus will lead to poor performance at the simulation termination.

This specific study focuses on the repair needed for the fan blades on the runner at the base of the hydroelectric generator. Here, they can turn the generator off, drain the water from the turbine, open the turbine casing, remove old fan blades, replace with new fan blades, dispose of the old fan blades, close the casing, refill the generator, and turn the power back on to verify the system works accordingly.

The sub-tasks; prep, loto, drain, blade remove, blade install, disposal, casing close, refill, and restart, are measured within a ninth dimensional state vector and normalized to values between 0 and 1 for each dimension where each index from 0 to 8 represents the corresponding sub-task.

B. User Profile and Tiers

We simulate users at three skill levels: novice, adept, and expert, based on their performance in the environment. Specifically, we categorized these levels into three sections: pessimistic, neutral, and optimistic. For the pessimistic tier, we assume that the user performs between 0.00 and 0.33 for the training environment. The neutral assumption places the user's performance metrics between 0.34 and 0.66. Lastly, the optimistic assumption places the user's performance above 0.67 and up to maximum performance represented by 1.0. Each of these tiers falls into an instruction tier that is representative of their ability. The lowest tier is novice where the user is given a step-by-step approach that focuses heavily on safety. The moderate tier (adept) gives more substantial detail and assumes some familiarity with the subject matter. The most skilled user is given expert instruction, where the instructions are concise but technical, emphasizing constraints and checks.

C. State Dynamics

The state is sampled after each training iteration and then fed into a python script running the reinforcement learning model. The sampled state vector represents the partial success/failures of each run and returns feedback based on the user's performance. Given that this is a pre-deployment simulation to characterize feedback behavior, not a substitute for real learners, we have opted to simulate agents within the environment giving them a randomized value for performance based upon their perceived skill tier.

IV. Feedback Models and Metrics

In simple terms, the model observes the user's performance (state), decides how feedback should be delivered (action), and improves this decision process over time using rewards. The following equations formalize how the model represents user performance and generates feedback decisions.

A. Reward Design for PPO Feedback Policy

We model the hydropower fan-blade maintenance task with a 9-dimensional normalized state vector

$$s_t = \left[s_t^{(0)}, s_t^{(1)}, \ldots, s_t^{(8)}\right] \in [0,1]^9 \tag{1}$$

where each component corresponds to one subtask in the workflow (e.g., lockout/tagout, draining, casing open/close, blade removal and installation, disposal, refill, and restart).[1]

We summarize overall performance by averaging how well each subtask is completed. A scalar task-performance score is computed as the mean of the normalized dimensions:

$$\text{perf}(s_t) = \frac{1}{9}\sum_{i=0}^{8} s_t^{(i)} \tag{2}$$

For each time step, the feedback policy $\pi_\theta(a_t \mid s_t)$ produces a 5-dimensional action vector

$$a_t = [a_t^{(C)}, a_t^{(L)}, a_t^{(S)}, a_t^{(F)}, a_t^{(D)}] \tag{3}$$

which we interpret as style weights for clarity (C), concision/length (L), specificity (S), safety (F), and directness (D). This action vector is used by a template-based generator to synthesize a prompt, which is then scored along five text-quality metrics:

$$C_t, S_t, F_t, A_t, D_t \in [0,1] \tag{4}$$

corresponding to clarity, specificity, safety, actionability, and domain accuracy, respectively (cf. Table II).

The overall text-quality score is a weighted sum:

$$J_{\text{text}}(s_t, a_t) = w_C\, C_t + w_S\, S_t + w_F\, F_t + w_A\, A_t + w_D\, D_t, \tag{5}$$

where, in our experiments,

$$\begin{aligned} &w_C = 0.30, \quad w_S = 0.35, \\ &w_F = 0.20, \quad w_A = 0.10, \quad w_D = 0.05 \end{aligned} \tag{6}$$

The scalar reward used for policy optimization is a convex combination of task performance and text quality:

$$r_t = \lambda_{\text{task}}\, \text{perf}(s_t) + \lambda_{\text{text}}\, J_{\text{text}}(s_t, a_t) \tag{7}$$

with

$$\lambda_{\text{task}} = \lambda_{\text{text}} = 0.5 \tag{8}$$

Thus, the feedback policy is encouraged to simultaneously improve the state of the underlying simulated task and the quality of the generated prompt. In practice, this means the model learns to give clearer, safer, and more relevant feedback over time.

B. PPO Objective

We train the feedback policy using a clipped PPO surrogate objective. Let π_θ be the current policy and $\pi_{\theta_{\text{old}}}$ the behavior policy. The probability ratio is

$$\rho_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \tag{9}$$

Given advantage estimates $\hat{A}_t$ constructed from the rewards in (7), the clipped surrogate is

$$\begin{aligned} L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[&\min\left(\rho_t(\theta)\,\hat{A}_t,\right. \\ &\left.\text{clip}(\rho_t(\theta),\, 1-\epsilon,\, 1+\epsilon)\,\hat{A}_t\right)\Big] \end{aligned} \tag{10}$$

[1] In code, raw scores are sampled as $\text{raw}_i \in [0, \max_i]$ with $\max = [1,1,1,8,8,8,1,1,1]$, then normalized as $s_i = \text{raw}_i / \max_i$.

where ϵ is the clipping parameter (set to $\epsilon = 0.2$ in our implementation). In our lightweight implementation, we use the scalar reward r_t (optionally normalized) as a proxy for the advantage when updating the policy parameters.

1) Random Baseline: The random condition samples (i) the feedback axis (preparation/safety, blade work, or restore/cleanup) and (ii) the instruction tier (novice, adept, expert) uniformly at random, independent of the current state. Prompts are still generated from the same constrained template and hydropower lexicon as the RL agent, which ensures that text remains syntactically coherent and domain-appropriate. However, RAND has no notion of which subtask is currently weak and cannot systematically target the lowest-performing dimension. This kind of stochastic baseline is standard practice in RL and interactive learning to distinguish genuine policy learning from performance that could be achieved by an uninformed but structurally similar policy.

2) Static Baseline: The static condition approximates the "traditional training" approach: prompts are drawn from a fixed pool of hand-authored messages that do not depend on the instantaneous state. In our implementation, STAT shares the same template family and vocabulary as the RL agent but typically defaults to an "adept-style" tier and does not adapt the targeted axis based on measured performance. Static baselines mirror prior work in intelligent tutoring and adaptive training systems, where rule-based or fixed feedback messages are compared against adaptive or data-driven policies to quantify the marginal benefit of personalization.

As seen in Table I we layout each condition of the prompt generation, how each axis is chosen, the templates utilized for each, and whether or not it is adaptable.

TABLE I
PROMPT GENERATION CONDITIONS

Cond.	Axis / Tier Selection	Templates & Lexicon	Adaptive
RL (adaptive)	Axis chosen as weakest state dimension; tier (novice / adept / expert) chosen from simulated user profile.	Uses shared hydropower templates and safety/domain lexicon; policy decides which combination to emit.	Yes
RAND (random)	Axis and tier sampled uniformly at random, independent of state.	Same template and lexicon pool; no learning or targeting.	No
STAT (static)	Axis typically fixed or sampled independent of state; tier often fixed to adept-like phrasing.	Same template family; prompts are hand-crafted variants, state-agnostic.	No

Positioning our adaptive PPO agent between these two baselines parallels how other VR+AI systems have been evaluated. For example, adaptive or AI-assisted agents in immersive medical and industrial scenarios are typically compared against non-adaptive scripted content and, in some cases, minimally structured control conditions [7]–[9], [21]. Similarly, reinforcement learning and procedural content generation work in XR often contrasts adaptive generators with static or randomly parameterized environments to demonstrate that performance gains arise from learning rather than from incidental variability [10]–[16].

C. Text Metrics

Table II summarizes the metrics used in the reward function. Clarity is scored to aim for shorter sentences, imperative verbs, low redundancy. Specificity is scored based on domain vocabulary, numeric constraints, sequencing words. Safety is presence and early placement of safety phrases (lockout/tagout, PPE, de-energize) Actionability checks if the prompt includes a concrete action-object-check pattern (do Y, then X, then verify Z). DomainAcc checks whether the prompt targets the same axis as the weakest part of the state.

TABLE II
TEXT QUALITY METRICS USED IN THE REWARD FUNCTION

Metric	Summary	Meaning of High Score
Clarity (C)	Penalizes long sentences and redundancy; rewards imperative, simple phrasing.	Instructions are easy to read and follow.
Specificity (S)	Counts domain terms, numeric constraints (e.g., torque values), and sequencing words ("before", "then").	Guidance is detailed, technical, and unambiguous.
Safety (F)	Detects presence and early placement of safety tokens (LOTO, PPE, de-energize).	Safety-first behavior is emphasized.
Actionability (A)	Checks action–object patterns ("drain the penstock", "close the casing").	Prompt can be executed as a concrete set of steps.
DomainAcc (D)	Compares the subtask referenced in the text to the weakest dimension of the state.	Feedback targets the correct step in the workflow.

V. EXPERIMENTAL DESIGN AND ABLATION STUDY

A. Why Ablation?

Ablation studies are critical for identifying the specific contributions of each component in a reinforcement learning (RL)–driven adaptive feedback system. Prior work has demonstrated that deep RL performance is highly sensitive to reward shaping, model structure, and implementation details, making careful component-level evaluation essential for interpreting results [17]. In VR training contexts, where instructional quality and user safety are directly tied to the policy's behavior, distinguishing the effects of adaptive difficulty, feedback logic, and state representation is necessary to ensure reliability and generalizability. By systematically disabling or modifying individual modules of our framework such as axis-weighting mechanisms, tier-based adaptation, and sub-reward signals, our ablation study clarifies which architectural elements meaningfully enhance training guidance. This approach follows best practices seen in adaptive training research [18], [19], ensuring

that improvements in user performance can be attributed to specific algorithmic contributions rather than incidental interactions between system components.

B. Ablation Conditions

To evaluate the contribution of each subsystem, we designed four targeted ablation conditions in addition to the complete adaptive RL model. Each variant was trained using identical environment configurations and scenario seeds for fair comparison, following reproducibility standards recommended by Henderson et al. [17].

1) Fixed-Axis Baseline: The adaptive axis-selection mechanism was removed and replaced with uniform, static weights across all axes. This condition assesses whether personalized weakness identification contributes to improved feedback quality and user progression.

2) No Tier-Based Adaptation: The tiered proficiency model was disabled so that all users received the same class of feedback regardless of demonstrated skill. This ablation isolates the effect of progressive feedback difficulty on learning stability and performance improvement.

3) Reward Component Removal: Individual reward terms—corresponding to preparation and safety, procedural correctness, diagnostic reasoning, and decision-making—were removed one at a time. This condition shows how each part of the reward affects the model's behavior, reflecting common reward-shaping analyses in RL literature [19], [22].

4) Static Scripted Feedback: The RL policy was replaced by a rule-based system with predefined heuristics and non-adaptive prompts. This serves as a strict baseline, enabling comparison between dynamic RL-generated guidance and traditional handcrafted instructional logic.

Across all ablated variants, we evaluate performance using metrics derived from feedback relevance, rate of user improvement, and final task completion accuracy. These comparisons illuminate the roles of adaptive difficulty, hierarchical feedback modeling, and reward shaping in guiding effective training behavior.

VI. Results

As shown in Table III we performed an ablation study using the RL condition where we configured various iterations of the model to inspect the effectiveness of the axis adaptation, tier adaptation, clarity reward and specificity reward. Since we have simulated users in this study, we can expect that the average reward for the user-based interactions, not the prompt quality metrics, may cause the overall reward to not be fully realized.

TABLE III
Ablation Study for RL Condition

Config.	Reward	Clarity	Safety	Dom. Acc
Full (axis + tier + full reward)	0.636	0.380	0.897	1.000
No axis adapt	0.621	0.386	0.924	0.337
No tier adapt (fixed tier)	0.602	0.281	0.701	1.000
Clarity-only reward	0.439	0.376	0.887	1.000
Specificity-only reward	0.719	0.376	0.867	1.000

A. Main Comparison

The full adaptive framework outperformed all baseline conditions across measures of training efficiency, task correctness, and feedback relevance. Compared to the static scripted feedback, the RL-based model produced feedback that aligned more closely to user performance trajectories resulting in faster improvement and higher domain accuracy. Across all evaluation runs, the adaptive RL model achieved the highest reward score, reflecting its ability to tailor instructional guidance to evolving user behaviors. The fixed-axis and non-adaptive baselines frequently consistently scored lower on overall reward metrics which resulted in feedback that was mismatched to user needs. In contrast, the RL policy dynamically modified feedback depth and specificity, enabling a more personalized learning pathway that could be translated to measurable skill gains.

TABLE IV
Adaptive Feedback Examples by Skill Tier

Tier	Axis	Feedback Examples
Novice	prep / safety	**Good:** Start by donning PPE. Then verify lockout/tagout and check for debris before draining the penstock. *(Step-by-step, safety-focused)* **Bad:** "Verify LOTO and proceed." *(Too vague, assumes prior knowledge)*
Adept	blade_remove	**Good:** With the unit de-energized, remove blades in sequence and confirm alignment before replacement. *(Moderate detail, assumes familiarity)* **Bad:** "Remove blades." *(Lacks guidance and specificity)*
Expert	restart	**Good:** Clear LOTO, perform a slow-roll check, and verify vibration and pressure are within tolerance. *(Concise, technical, verification-focused)* **Bad:** "Turn the system back on carefully." *(Too simplistic, lacks technical depth)*

Table IV demonstrates how feedback adapts to user skill level, highlighting both appropriate and inappropriate examples to illustrate differences in feedback depth and specificity.

B. Ablation Effects

The ablation study reveals the relative importance of each architectural component in shaping effective adaptive feedback. Removing the adaptive axis-selection mechanism led to the largest performance decline among the ablated variants. Without dynamic axis weighting, the model was unable to prioritize user-specific weaknesses, resulting in more generic feedback and slower advancement through proficiency tiers.

Disabling the tier-based adaptation module produced a moderate degradation in results. The absence of progressively structured feedback limited the model's ability to scale difficulty appropriately, causing users to plateau on mid-tier competencies. This outcome reflects broader findings in adaptive training literature, where hierarchical progression cues are central to sustained skill development [5], [6].

Reward component ablations showed distinct patterns depending on which signal was removed. Eliminating safety and preparation-related rewards caused the policy to underemphasize foundational procedural steps, while removing diagnostic or decision-making rewards yielded policies that improved rapidly but lacked depth in critical reasoning tasks. These observations echo trends in RL reward design studies, where the omission of key components leads to biased behavioral strategies [18], [23].

Finally, replacing the RL policy with a static scripted system produced the weakest performance across all metrics, confirming that the adaptive mechanisms, which are not merely the presence of feedback, drive the observed learning gains. Together, the ablation results highlight the necessity of combining adaptive axis weighting, hierarchical proficiency modeling, and multi-component reward shaping to achieve robust, individualized VR training performance.

VII. Discussion and Limitations

The results show that adaptive feedback improves training effectiveness by matching instruction to user skill level. Beginners benefit from detailed, step-by-step guidance, while experienced users benefit from concise, technical feedback. When adaptation is removed, feedback becomes less relevant and less effective.

Limitations include the use of simulated users and automated evaluation metrics rather than real participants. Future work will validate these findings with human subjects and extend the system to additional maintenance tasks.

VIII. Conclusion and Future Works

Overall, this study reveals keen insights into the possibility of increasing adaptable training environments into workplace training norms by creating a prompt-quality evaluation loop for an RL feedback agent in a hydroelectric plant. However, with it only covering a small portion of maintenance expectations for hydroelectric plant workers, we plan to incorporate this method of reinforcement learning to more tasks such as reservoir cleaning, clearing blockages, turbine rotor repair, and patching leaks in the various piping systems of the penstock.

References

[1] R. Uria Martinez and M. Johnson, "Us hydropower market report 2023," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), Tech. Rep., 2023. [Online]. Available: https://www.energy.gov/eere/water/hydropower-market-reports

[2] A. Acakpovi and L. Dzamikumah, "An investigation of health and safety measures in a hydroelectric power plant," *Safety and health at work*, vol. 7, no. 4, pp. 331–339, 2016.

[3] M. Zahabi and A. M. Abdul Razak, "Adaptive virtual reality-based training: a systematic literature review and framework," *Virtual Reality*, vol. 24, no. 4, pp. 725–752, 2020.

[4] L. F. Lui, U. Radhakrishnan, F. Chinello, and K. Koumaditis, "The efficacy of adaptive training in immersive virtual reality for a fine motor skill task," *Virtual Reality*, vol. 29, no. 1, p. 20, 2025.

[5] K. M. Stanney, A. Skinner, and C. Hughes, "Exercisable learning-theory and evidence-based andragogy for training effectiveness using xr (elevate-xr): elevating the roi of immersive technologies," *International Journal of Human–Computer Interaction*, vol. 39, no. 11, pp. 2177–2198, 2023.

[6] A. Verniani, E. Galvin, S. Tredinnick, E. Putman, E. A. Vance, T. K. Clark, and A. P. Anderson, "Features of adaptive training algorithms for improved complex skill acquisition," *Frontiers in Virtual Reality*, vol. 5, 2024.

[7] X. T. Zhu, H. Cheerman, M. Cheng, S. R. Kiami, L. Chukoskie, and E. McGivney, "Designing vr simulation system for clinical communication training with llms-based embodied conversational agents," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–9.

[8] T. Duricic, P. MÃžllner, N. Weidinger, N. ElSayed, D. Kowald, and E. Veas, "Ai-powered immersive assistance for interactive task execution in industrial environments," *arXiv preprint arXiv:2407.09147*, 2024. [Online]. Available: https://arxiv.org/abs/2407.09147

[9] Y.-Z. Lin, K. Petal, A. H. Alhamadah, S. Ghimire, M. W. Redondo, D. R. V. Corona, J. Pacheco, S. Salehi, and P. Satam, "Personalized education with generative ai and digital twins: Vr, rag, and zero-shot sentiment analysis for industry 4.0 workforce development," *arXiv preprint arXiv:2502.14080*, 2025. [Online]. Available: https://arxiv.org/abs/2502.14080

[10] A. Khalifa, P. Bontrager, S. Earle, and J. Togelius, "Pcgrl: Procedural content generation via reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 16, no. 1, 2020, pp. 95–101.

[11] C. E. Lopez, O. Ashour, and C. S. Tucker, "Reinforcement learning content generation for virtual reality applications," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 59179. American Society of Mechanical Engineers, 2019.

[12] C. E. López, J. Cunningham, O. Ashour, and C. S. Tucker, "Deep reinforcement learning for procedural content generation of 3d virtual environments," *Journal of Computing and Information Science in Engineering*, vol. 20, no. 5, 2020.

[13] L. Gisslén, A. Eakins, C. Gordillo, J. Bergdahl, and K. Tollmar, "Adversarial reinforcement learning for procedural content generation," in *IEEE Conference on Games (CoG)*, 2021, pp. 1–8.

[14] A. Mahmoudi-Nejad, M. Guzdial, and P. Boulanger, "Arachnophobia exposure therapy using experience-driven procedural content generation via reinforcement learning (edpcgrl)," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 17, no. 1, 2021, pp. 164–171.

[15] J. Agarwal and S. Shridevi, "Procedural content generation using reinforcement learning for disaster evacuation training in a virtual 3d environment," *IEEE Access*, vol. 11, pp. 98 607–98 617, 2023.

[16] A. S. Joshi, "Reinforcement learning-enhanced procedural generation for dynamic narrative-driven ar experiences," *arXiv preprint arXiv:2501.08552*, 2025. [Online]. Available: https://arxiv.org/abs/2501.08552

[17] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[18] R. Fuchs, R. Gieseke, and A. Dockhorn, "Personalized dynamic difficulty adjustment imitation learning meets reinforcement learning," in *IEEE Conference on Games (CoG)*, 2024, pp. 1–2.

[19] R. Stolz, H. Krasowski, J. Thumm, M. Eichelbeck, P. Gassert, and M. Althoff, "Excluding the irrelevant: Focusing reinforcement learning through continuous action masking," *Advances in Neural Information Processing Systems*, vol. 37, pp. 95 067–95 094, 2024.

[20] X. Tang, F. Li, Z. Cao, Q. Yu, and Y. Gong, "Optimising random forest machine learning algorithms for user vr experience prediction based on iterative local search-sparrow search algorithm," in *6th International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2024, pp. 1387–1391.

[21] P. Spiegler, A. Harirpoush, and Y. Xiao, "Towards user-centered interactive medical image segmentation in vr with an assistive ai agent," *arXiv preprint arXiv:2505.07214*, 2025. [Online]. Available: https://arxiv.org/abs/2505.07214

[22] K. Mitsis, K. Zarkogianni, E. Kalafatis, K. Dalakleidi, A. Jaafar, G. Mourkousis, and K. S. Nikita, "A multimodal approach for real time recognition of engagement towards adaptive serious games for health," *Sensors*, vol. 22, no. 7, p. 2472, 2022.

[23] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.