INTENSIF, Vol.10 No.2 2026
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

Framework for Scenario-Driven Black-Box Testing of NPC Behavior in Virtual Reality Games

Received:
10 May 2026
Accepted:
15 July 2026
Published:
22 August 2026

[1*]Matahari Bhakti Nendya, [2]Antonius Rachmat Chrismanto, [3]Dan Daniel Pandapotan, [4]Vito Gautama, [5]Lunchakorn Wuttisittikulkij, [6]Gabriel Indra Widi Tamtama
[1,2,4]Department of Informatics, [3]Departement of Product Design, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia
[5]Department of Electrical Engineering, Chulalongkorn University, Bangkok, Thailand
[6]Institute of Service Science, National Tsing Hua University, Hsinchu City, Taiwan
E-mail: [1]didanendya@staff.ukdw.ac.id, [2]anton@ti.ukdw.ac.id, [3]danpandapotan@staff.ukdw.ac.id, [4]vito.gautama@ti.ukdw.ac.id, [5]lunchakorn.w@chula.ac.th, [6]gabrielindra.giwt@iss.nthu.edu.tw

*Corresponding Author

Abstract— Background: Non-Playable Characters (NPCs) play an important role in the quality of interaction in Virtual Reality (VR) games. As NPC systems become more complex and multi-role, their behavior must be evaluated not only from a design perspective but also in terms of technical stability and interaction consistency. However, systematic methods for validating NPC behavior in VR games remain limited. **Objective:** This study proposes and implements a scenario-driven black-box testing framework for evaluating NPC behavior in VR games based on observable interaction outcomes. **Methods:** NPC behavior was evaluated using normal, boundary, and stress-based interaction scenarios. Four behavioral quality criteria were assessed: determinism, state integrity, role consistency, and recovery stability. Selected dimensions of the Game Experience Questionnaire (GEQ), namely Social Behaviour Activity and Post-game Module, were also evaluated. **Results:** Six test scenarios were evaluated. Four scenarios (66.7%) met all criteria, while two (33.3%) were categorized as partial due to minor delays in state transitions and temporary idle-state resets. No crashes, system failures, or fatal interaction breakdowns were observed. GEQ results indicated moderate-to-good social behavior involvement (mean = 3.3) and a comfortable post-game experience (mean = 3.5). **Conclusion:** The proposed framework provides a structured and reproducible approach for assessing NPC behavior in VR games without requiring access to internal implementation details. The findings highlight the importance of technical behavioral testing, as positive player experience does not necessarily indicate fully stable and consistent NPC behavior.

Keywords— Virtual Reality Games; Non-Playable Characters; Scenario-Driven Testing; Black-Box Testing; NPC Behavior Evaluation

Corresponding Author:

Matahari Bhakti Nendya,
Department of Informatics,
Universitas Kristen Duta Wacana,
Email: didanendya@staff.ukdw.ac.id
Orchid ID: https://orcid.org/0000-0002-4708-9087

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

I. INTRODUCTION

Virtual Reality (VR) games have evolved into complex interactive software systems that integrate immersive environments, continuous user input, and autonomous agents capable of dynamic interaction. In these systems, the primary role of Non-Playable Characters (NPCs) is to assess the quality of interaction by mediating guidance, social presence and contextual feedback. Studies of immersive VR serious games [1] have shown that the quality of agent interaction has a strong influence on immersion and user engagement in learning and training applications. Large-scale reviews of VR use in educational and simulation environments have also suggested that the consistency of virtual character behavior is a major contributor to perceived realism and usability [2]. Meta-analyses reveal that immersion and presence, which are closely related to the quality of interaction[3] have significant effects on learning outcomes. Recent work on immersive environments from an artificial intelligence point of view, emphasizes the fact that autonomous agents have to exhibit stable and role-consistent behavior in order to maintain social presence and system credibility [4], [5]. Furthermore, empirical research on virtual agents and social interaction in VR shows that behavioral inconsistency can break down presence engagement and interaction quality even when visual fidelity is still high [6] [7]. In general, these findings show the relevance of NPC behavior as a quality software attribute for VR-based interactive systems.

Many recent works have been devoted to the modeling of NPC behavior through techniques such as Finite State Machines (FSM), Hierarchical FSM (HFSM), Behavior Trees (BT), and hybrid AI architectures, to enhance modularity, scalability, and behavioral realism. Hierarchical state-based models have been shown to be better structured to handle complex NPC behaviors than flat state machines[8], [9]. Previous works have studied FSM, HFSM, Behavior Trees, affective or personality-driven agents, and learning-based NPC models to enhance modularity, adaptivity, and behavioral realism [10], [11] [12]. More recent works have proposed learning-based models to improve the adaptivity and decision-making capabilities of NPCs [13], [14]. A common challenge of game AI, however, is not only to select an adequate decision model, but to create a coherent, human-like and context-relevant behavior in a variety of game situations [15]. At the same time, research in Game Software Engineering indicates that quality assurance of games faces particular difficulties such as fast iteration cycles and intricate interaction paths, which leads to testing practices that tend to focus on functional correctness rather than systematic behavioral validation [16], [17]. Empirical surveys of industry practices reveal that game testing remains dominated by manual playtesting and informal verification [18], [19].

In this study, functional correctness and behavioral stability are treated as different but related evaluation concerns. Functional correctness and behavioral stability are treated as related but

INTENSIF, Vol.10 No.2 2026
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

distinct evaluation concerns. An NPC may produce correct dialogue or animation in ordinary playtesting but still show unstable behavior during repeated, interrupted, or rapidly switched interaction sequences; therefore, NPC evaluation must examine behavioral coherence across time and interaction contexts[16], [18], [19].

However, despite advances in NPC behavior modeling and automated game testing, a significant gap remains in the systematic technical evaluation of NPC behavior from a black-box perspective. Recent regression testing frameworks for video games have proved effective in detecting unintended changes in behavior between software versions by modeling execution traces and state transitions, yet their main concern is functional regressions, rather than behavioral stability across sequences of interaction [20]. Likewise, reinforcement learning-assisted testing approaches have been demonstrated to improve test coverage and explore vast state spaces in complex games, but they are mainly developed to optimize exploration efficiency and stress system limits, rather than to assess determinism or role-consistent behavior of NPCs from players' perspective[21], [22]. Research in the meantime has pointed to the significance of keeping consistent role-dependent behavior to maintain immersion, social presence and system credibility in interactive environments by investigating personalized and role-based non-player characters [23], [24]. These results indicate that NPC correctness should be assessed not only in terms of functional outcomes, but also in terms of behavioral coherence across repeated and contextual interactions. Scenario-based testing has been shown to be effective for validating context-dependent behavior in complex systems beyond the game domain. Scenario-based testing has been widely used in complex autonomous and interactive systems to expose behavioral edge cases under normal, boundary, and stress conditions, particularly when system behavior emerges through context-dependent interaction sequences [25], [26], [27], [28]. These studies indicate that scenario-based testing is suitable for evaluating systems whose behavior emerges through sequences of context-dependent interactions. However, its application to NPC behavior evaluation in VR games remains limited, particularly when the evaluation is conducted from a black-box and player-centric perspective.

The difference between this work and prior work is that this work does not propose a new NPC behavior model, AI architecture, or dialogue generation technique, but a scenario-driven black-box testing framework to evaluate whether NPC behavior is stable, deterministic, role-consistent, and recoverable across different interaction conditions. Unlike prior work that mainly focuses on designing NPC behavior models or improving automated test coverage, this work proposes an implementation-independent evaluation framework based on observable player-facing behavior such as dialogue output, animation response, interaction timing, and state progression [8], [9], [11], [13], [14], [15]. [18], [19], [20], [21], [22]. In contrast, this study

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

contributes [8], [9], [11], [13], [14], [15]. In contrast, this study contributes an implementation-independent evaluation framework that can be applied to different NPC architectures because it relies only on observable behavior, including dialogue output, animation response, interaction timing, and state progression. This focus makes the proposed framework useful for VR game developers who need to verify NPC behavioral quality without accessing or modifying internal NPC control logic.

The goal of this research is to develop and apply a scenario driven black-box testing framework for evaluating the behavior of NPCs in Virtual Reality games. The proposed framework classifies NPC evaluation into normal, boundary and stress interaction scenarios and analyzes behavioral quality on observable criteria such as determinism, state integrity, role consistency and recovery stability. This work provides (1) a systematic and reproducible method for measuring NPC behavior from a player-centric point of view, (2) an operationalization of behavioral quality criteria for black-box testing of VR-based interactive systems, and (3) empirical evidence that the framework is capable of finding behavioral edge cases that are difficult to find with traditional playtesting. This work addresses the problem of validating NPC behavior as a software quality assurance issue to provide a practical methodological foundation for enhancing the reliability and robustness of VR games and other immersive interactive systems.

II. RESEARCH METHOD

In this paper we propose a scenario-driven black-box testing approach for the evaluation of Non-Playable Character (NPC) behavior in Virtual Reality (VR) games based on a framework-driven research methodology . The main methodological contribution of this research is the layered evaluation framework depicted in Figure 1, which is the basis of the proposed methodology. The framework explicitly models the transformation from interaction scenarios to behavioral observations, which are abstracted to quality criteria and aggregated to evaluation outcomes, unlike linear experimental workflows. The methodology is implementation independent, as it considers only observable NPC behaviour, rather than internal AI representations such as FSMs, HFSMs or behaviour trees. This design decision is consistent with the well-established principles of black-box testing in software engineering and game quality assurance research [17], [29], [30]. The framework maps player-NPC interaction scenarios to evidence of observable behavior . The results are classified into Pass, Partial or Fail as shown in Figure 1.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

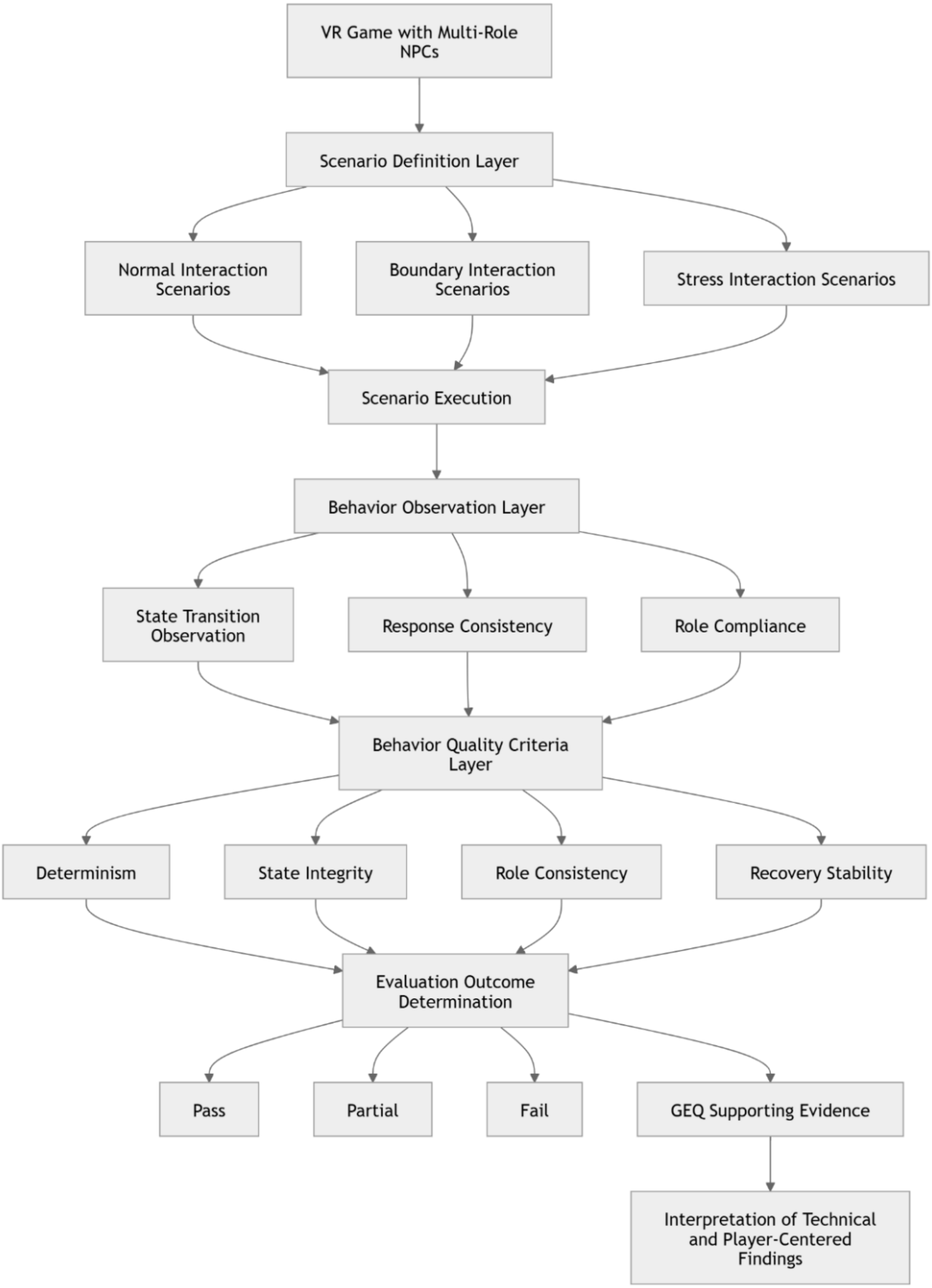

Fig 1. Research flow of the scenario driven black-box NPC test of the proposed framework. The novelty of the proposed framework is embodied in Scenario Definition Layer, Behavior Observation Layer, Behavior Quality Criteria Layer and Evaluation Outcome Determination.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

A. Scenario Definition Layer

The methodology starts with the Scenario Definition Layer (Fig. 1) where interaction scenarios are defined as structured test cases. Each scenario generates a set of player actions and environmental conditions intended to trigger NPC reactions over time. In this context, the main units to organize and control the behavioral evaluation are scenarios, following the traditional practices in scenario-based software testing and validation of interactive systems [31].

Interaction scenarios are concrete exchanges between players and NPCs (Fig. 2) where the player chooses dialogue options and the NPC generates observable actions and feedback. The example shown is one of a normal interaction scenario and can be used as a black-box test case. Here, the behavior of the NPC is only tested by its externally visible responses, irrespective of the internal control logic.

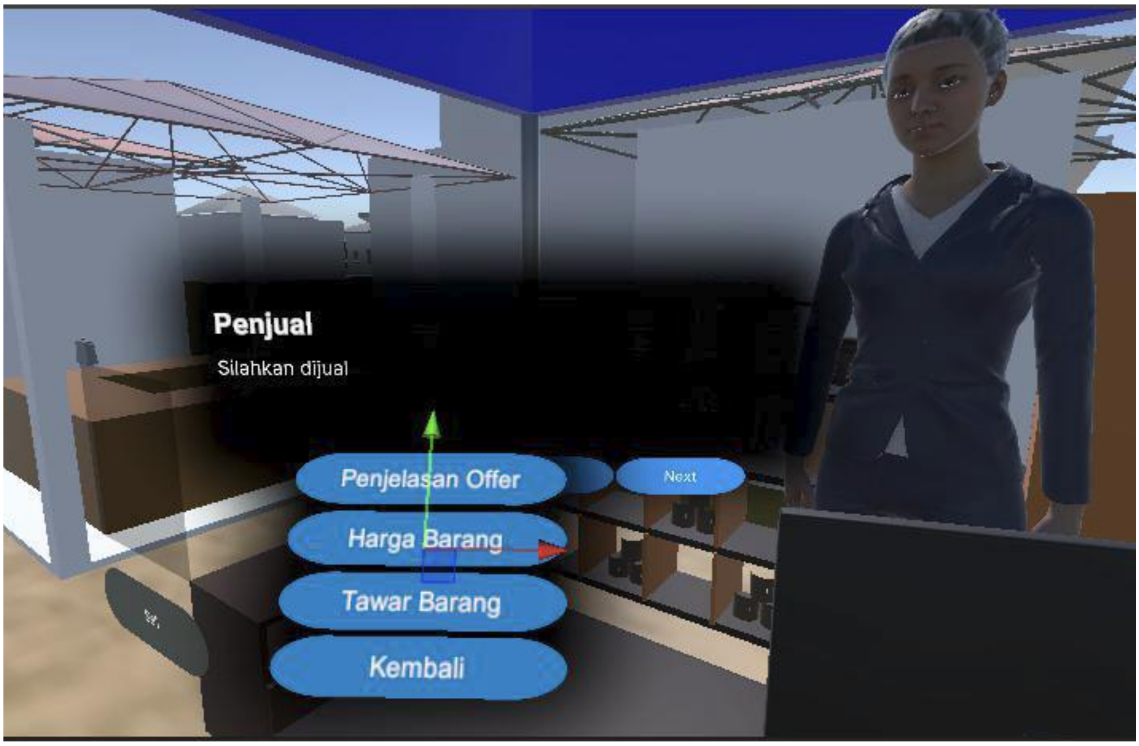

Fig 2. A Typical Interaction Scenario illustrating observable NPC behavior as input for scenario-driven black-box testing

Scenario-based testing has been widely used in recent research on complex and autonomous systems to expose behavioral edge cases that are not detectable through isolated functional tests [25], [27], [28], [32]. Scenarios are divided into three types to cover systematically:

1. Normal interaction scenarios, showing what interactions to expect in the gameplay.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

2. Boundary interaction scenarios, i.e. interaction at the limits of behavioural conditions like repeated triggering or fast context switching.
3. Stress interaction cases. Uncover long or abnormal interaction patterns to diagnose instability and recovery issues.

This classification is in line with recent practice in scenario based testing of autonomous and interactive systems [25], [27]. Table 1 presents operational definitions, example test actions, and the primary evaluation focus for each scenario type.

Table 1. Operational Definition of Scenario Types

Scenario Type	Operational Defintion	Example Test Action	Main Evaluation Focus
Normal	Expected player-NPC interaction under ordinary gameplay conditions	The player approaches the Guidance NPC, selects a dialogue option, receives a response, and ends the interaction normally	Correct response, role consistency, and normal state transition
Boundary	Interaction at the limit of expected behavior	The player rapidly switches between NPCs or repeatedly opens and closes the same dialogue within a short interval	Determinism and state integrity under rapid interaction changes
Stress	Prolonged or abnormal interaction designed to expose instability	The player repeatedly triggers the same NPC interaction, interrupts the interaction by moving away, and reactivates the NPC immediately	Recovery stability, idle reset handling, and delayed transition detection

B. Behavior Observation Layer

Once defined, the execution of the scenarios generates observable interaction traces that can serve as the basis for behavioral analysis. All scenarios are contained within the Behavior Observation Layer. Here the NPC's behavior is assessed from a black-box perspective. At this stage we do not assume any internal control structures for the NPCs, but evaluate solely on the basis of external observable behavior, such as dialogue output, animation response, interaction timing, and state progression. Behavioral observation focuses on three dimensions illustrated in Fig 1:

1. State transition observation, examining whether NPCs move coherently between interaction states.
2. Response consistency, analysis on how similar interactions lead to response consistency across repeated executions.
3. Role compliance, whether the behavior observed is consistent with the intended functional role of the NPC.

This observation strategy is consistent with recent studies that stress external behavior as the main indicator of correctness for interactive agents and game systems [17], [19] , [32].

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

At each scenario run a structured observation sheet was filled in with the following information: Type of scenario, actions of the player, expected behavior, observed behavior, time of transition, role consistency, recovery behavior and final outcome.

C. Behavior Quality Criteria Layer

The Behavior Quality Criteria Layer then summarizes the observed behaviors into four complementary criterion-based assessments:

1. Determinism: The responses of the NPCs to multiple executions of the same scenario are consistent.
2. Integrity of the state Maintaining valid behavioral states across sequences of interactions.
3. Role consistency. Matching between observed behaviour and pre-defined roles of NPCs.
4. Recovery stability: The ability of NPCs to return to valid behavioral states after an interruption or a deviant interaction.

Together, these criteria capture different but related aspects of behavioral correctness from a black-box point of view, allowing for a systematic evaluation of NPC behavior as perceived by players rather than as implemented internally. The selection of these criteria is motivated by recent work in game quality assurance, automated game testing, agent behaviour assessment and reliability of interactive systems [16], [33], [34] [16]. These criteria have been chosen because they measure behavioral stability in repeated, interrupted, and context-sensitive interactions, not just functional correctness.

D. Evaluation Outcome Determination

The evaluation result of a scenario execution is obtained by aggregating the behavioral quality criteria according to the aggregated assessment of the defined behavioral quality . The results are grouped as follows:

1. Pass, when all behavioral criteria are satisfied.
2. Partial, when minor behavioral issues are observed without disrupting overall interaction coherence.
3. Fail, in case of serious breaches of the behavioural code.

This classification scheme enables reproducible and transparent reporting of NPC behavioural quality across scenarios. This is in line with evaluation practices in software testing research [20], [29].

Pass means all requirements were met. Partial means minor problems, such as state transition delays, temporary idle resets, or short response inconsistencies, without breaking the interaction. Fail is for invalid state, wrong role behavior, failed recovery or breakdown in interaction.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

E. Participants and Supporting Data Collection

The proposed framework was applied to a VR game environment with multi-role NPCs. We analyzed behavioral determinism and recovery stability by repeating six predefined interaction scenarios. In addition to the technical evaluation, player-centered data were collected using selected dimensions of the Game Experience Questionnaire (GEQ) [35], a validated tool widely used in VR and game research [36]. To provide supporting evidence of the perceived quality of NPC interaction and post-interaction comfort, the dimensions of Social Behaviour Involvement and Post-game Module were selected.

Participants were 12 students (age range 19-23 years) from a university with previous VR experience. Confounding effects of first time VR users were reduced by using purposive sampling. Participants experienced the predefined interaction scenarios in the VR environment and then filled out the selected dimensions of the GEQ. All participants provided informed consent before data collection. Participants were informed of the aims of the study, the interaction procedures and their right to withdraw at any time without penalty.

The sample size was considered suitable for validation of the methodological framework as small samples are generally used in structured usability and exploratory testing studies to identify interaction and behavioral problems [37], [36], [19], [17].

III. RESULT AND DISCUSSION

The NPC behaviour was tested using the proposed scenario-driven black-box testing framework with six predefined interaction scenarios. Scenarios included normal, boundary and stress conditions. Each scenario was run multiple times to check for behavioral stability and recovery. The evaluation results are based on four behavioral quality criteria: determinism, state integrity, role consistency and recovery stability.

The results of this research are that the proposed scenario-driven black-box testing framework can identify subtle NPC behavioral instability that is not necessarily observable in conventional playtesting or player experience scores. The system baseline was robust in terms of not crashing or failing fatally during testing, but the framework detected behavioral edge cases under boundary and stress interaction conditions such as small state transition delays and temporary idle resets. They did not induce immediate interruptions of the flow of gameplay but suggested possible shortcomings of the NPC behavioral recovery mechanism. Table 2 presents the summary of the evaluation results of the six predefined scenarios.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

Table 2. NPC Behavior Evaluation Results Summary

Scenario Type	Number of Scenarios	Pass	Partial	Fail	Main Observation
Normal	Normal	Normal	Normal	Normal	NPCs reacted consistently in dialogue, animation, and role appropriate responses during expected game play interactions.
Boundary	Boundary	Boundary	Boundary	Boundary	In one case, a small delay was seen in the state transition for fast interaction switching, but the NPC eventually returned to the correct behavior.
Stress	Stress	Stress	Stress	Stress	One scenario showed a temporary idle reset during long or abnormal interaction, but the NPC recovered without system failure.
Total	6	4 (66.7%)	2 (33.3%)	0	No crash or fatal behavioral failures were observed.

Table 2 shows that four scenarios (66.7%) were classified as Pass, two scenarios (33.3%) as Partial due to transition delays and temporary idle resets and zero as Fail.

A. Behavioral Result by Evaluation Criteria

In standard interaction situations NPC responses were mostly deterministic when the same situation was repeated several times. If scenarios were repeated, NPCs always gave the same sequences of dialogue and animation. This result shows that the NPC system could give predictable replies during normal player-NPC interaction. However, under edge conditions of fast switching of interaction small deviations in the time of state transition were noticed. While these deviations did not result in wrong answers, they did impact temporal consistency and resulted in Partial classification.

The tested scenarios have generally preserved the integrity of the state. NPCs have not entered an invalid or unintended state during normal interaction. However, the boundary scenario showed that the fast switching of interactions can have a transient effect on the timing of state changes. This implies that the NPC system is functionally correct, but its behavioral timing is not completely stable under edge-case interaction patterns. This is important because black-box evaluation can expose problems that are not apparent when you only test whether a feature gives the expected output.

Role consistency was largely maintained in all scenarios. NPCs behaved according to their pre-defined functional roles in routine interactions, thus supporting the effectiveness of role-based NPC design. We did not observe any role violations across the tested scenarios. For example, we found that the NPCs we tested did not fall back on behavior that contradicted their assigned roles, even when the interaction sequences were repeated, interrupted, or otherwise abnormal. This

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

means that, at the level of functional identity, role-based behavior remained the same and only timing and recovery still needed to be improved.

In stress scenarios, stability of recovery was the most sensitive criterion. Occasionally, after long or abnormal interactions, NPC's would go into idle states for a short period of time before returning to normal behaviour. Despite the fact that the recovery was successful, the observed delay signals possible flaws in the way state recovery is managed. The delay justified the classification as Partial. This result suggests that recovery stability is an important criterion to evaluate the behavior of an NPC, because an NPC might be correct during normal interaction but still has delayed recovery when the player repeatedly interrupts or reactivates the interaction sequence.

Prior research in game software engineering and scenario-based testing supports the results of this research. Interaction problems in complex systems are often revealed under boundary or stress conditions, not during normal operation [16], [18], [19], [25], [27], [28]. The results are also in line with previous research on NPC believability and social interaction in VR, which shows that consistent and role-aligned agent behavior is important to maintain immersion, social presence and user trust in interactive environments [4], [5], [6], [23], [24]. Thus, the Partial results presented in this work should be interpreted not as a major system failure, but as a diagnostic evidence that the proposed framework is sensitive enough to detect subtle behavioral vulnerabilities.

B. Player-Centered Supporting Evidence

The technical evaluation results were supplemented by the analysis of selected dimensions of the Game Experience Questionnaire (GEQ). Results showed moderate to good social behavior involvement (mean = 3.3) and enjoyable post-game experience (mean = 3.5). These scores indicate a satisfactory quality of perceived interaction. The Social Behaviour involvement score shows that the interaction with the NPC was acceptable and socially engaging for the participants, despite the small behavioural problems found from technical testing. The Post-game Module score suggests that the interaction session in VR was generally comfortable for the participants and that they did not experience severe post-interaction discomfort. This means that from user perspective, the NPC system still provided an acceptable interaction experience. This divergence shows that acceptable GEQ scores do not necessarily indicate full behavioral robustness, because black-box testing still revealed subtle technical issues.

C. Discussion and Implications

These results indicate that scenario-driven black-box testing can reveal NPC behavioral issues that remain hidden during ordinary playtesting. These results indicate that scenario-driven black-

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

box testing can reveal NPC behavioral issues related to determinism and recovery stability that might be hidden during normal playtesting. The absence of Fail outcomes suggests baseline behavioral robustness while the Partial outcomes indicate remaining vulnerabilities in transition and recovery handling. Methodologically, the findings validate the employment of scenario-driven testing to evaluate NPC behavior in VR games with different NPC architectures. For example, the same evaluation logic can be applied to evaluate NPCs implemented in FSM, HFSM, Behavior Tree or learning-based control models, as the framework only evaluates observable behavior instead of internal implementation. From a practical viewpoint, the identified temporary idle reset under stress conditions suggests that the NPC recovery logic should be considered a dedicated architectural concern.

Developers should not default to a generic idle state as the fallback after interruption. NPCs should instead have explicit recovery states that store the previous valid state, interrupted interaction context and expected return state after recovery. This would allow the NPC to return to a more coherent behavior after abnormal or repeated player interaction. The slight delay in state transitions under boundary conditions also hints at the need for transition guards, interaction cooldowns, or input debounce mechanisms to prevent an NPC from receiving conflicting triggers before the previous transition has completed. And each role of an NPC should have a role-specific recovery rule. A Guide NPC should be reset to a guide ready state, a Culprit NPC should be reset to its violation or reaction cycle, and a Bystander NPC should be reset to its observation or ambient behavior cycle. These improvements can help decrease idle resets, and help maintain role consistency and behavioral stability for repeated or abnormal player interactions. The results also imply that VR game developers should combine player-centered evaluation and technical behavior testing. GEQ can tell if the users feel comfortable and socially meaningful during the interaction. Scenario-driven black-box testing can tell if the NPC system is stable under repeated, interrupted or abnormal interaction patterns. Combining both approaches provides a more comprehensive assessment of NPC quality than either approach alone.

IV. CONCLUSION

In this paper we have shown that a scenario-driven black-box testing framework can be a systematic and reproducible approach to validate the behavior of non-player character (NPC) in Virtual Reality games. The framework can expose behavioral edge cases of determinism and recovery stability that are hard to detect using traditional playtesting, by using predefined interaction scenarios and explicit criteria for behavioral quality. Evaluation of six scenarios resulted in four scenarios (66.7%) being classified as Pass. Two scenarios (33.3%) were classified as Partial due to mild state transition delays and temporary idle resets. with no Fail outcomes,

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

system crashes or fatal interaction breakdowns observed. The corroborating results from the GEQ of Social Behaviour Involvement, with a mean score of 3.3, and Post-game experience, with a mean score of 3.5, suggest that participants perceived the NPC interaction as acceptable and comfortable. Subtle behavioral vulnerabilities were still detected via technical testing. These results confirm that the positive user experience scores should not be seen as enough evidence of behavioral robustness, as NPC systems may still have issues with stability for repeated, interrupted or abnormal interaction patterns. However, limitations of this study included the use of six predefined scenarios, 12 participants with previous VR experience and manual observation of behavior through structured observation sheets . Future work should focus on more diverse scenarios, include participants with varying levels of VR experience, implement automated logging and scripted scenario execution, and compare the framework across different NPC architectures. In general, the proposed framework contributes to game software engineering and VR interaction design by positioning NPC behavioral correctness as a software quality assurance problem and by providing an implementation-independent approach to improve the reliability and stability of immersive interactive systems.

AuthorContributions. M*atahari Bhakti Nendya*: Conceptualization, Methodology, Software (base code development), Writing – Original Draft, Writing – Review & Editing, and Supervision. *Antonius Rachmat Chrismanto*: Conceptualization, Investigation, and Supervision. *Dan Daniel Pandapotan*: Conceptualization, Investigation, Data Curation, and Software (3D environment development). *Vito Gautama*: Investigation, Data Curation, and Software (application development). *Lunchakorn Wuttisittikulkij*: Writing – Review and Supervision. *Gabriel Indra Widi Tamtama*: Writing – Review and Supervision.

Funding: The research was funded by Institute of Research and Community Service (LPPM), Universitas Kristen Duta Wacana under Skema Riset Produk Inovasi grant number 122/D.02/LPPM/2025.

Acknowledgments: The authors would like to thank the Faculty of Information Technology and the Faculty of Architecture and Design, Universitas Kristen Duta Wacana, for their support and collaboration in the implementation of this research. The authors would also like to thank SMA BOPKRI 1 Yogyakarta for their valuable cooperation as a partner in the user testing phase of the PasukanKuningVR application. Furthermore, the authors thank the volunteers, game and virtual reality experts and academic colleagues for their feedback and participation which enriched the depth and quality of the study.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data supporting the findings of this study are not publicly available due to the inclusion of human participant interaction data and questionnaire responses, which are subject to privacy and ethical restrictions. Aggregated evaluation results and framework specifications are fully described within the paper. Further information may be made available from the corresponding author upon reasonable request.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

Informed Consent: Informed consent was obtained from all participants prior to data collection. Participants were informed about the purpose of the study, the interaction procedures, and their right to withdraw at any time without penalty. Further details regarding participant recruitment and study procedures are described in the Methods section.

Animal Subjects: There were no animal subjects.

ORCID:
Matahari Bhakti Nendya: https://orcid.org/0000-0002-4708-9087
Antonius Rachmat Chrismanto: https://orcid.org/0000-0003-3247-2419
Dan Daniel Pandapotan: https://orcid.org/0009-0001-2600-6609
Vito Gautama: https://orcid.org/0009-0003-7175-3413
Lunchakorn Wuttisittikulkij: https://orcid.org/0000-0002-3033-3020
Gabriel Indra Widi Tamtama: https://orcid.org/0000-0002-4779-5563

REFERENCES

[1] D. Checa and A. Bustillo, “A review of immersive virtual reality serious games to enhance learning and training,” Multimed. Tools Appl., vol. 79, no. 9–10, pp. 5501–5527, Mar. 2020, doi: 10.1007/s11042-019-08348-9.
[2] G. J. Hwang and S. Y. Chien, “Definition, roles, and potential research issues of the metaverse in education: An artificial intelligence perspective,” Computers and Education: Artificial Intelligence, vol. 3, Jan. 2022, doi: 10.1016/j.caeai.2022.100082.
[3] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, “A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda,” Comput. Educ., vol. 147, p. 103778, Apr. 2020, doi: 10.1016/J.COMPEDU.2019.103778.
[4] M. Coban, Y. I. Bolat, and I. Goksu, “The potential of immersive virtual reality to enhance learning: A meta-analysis,” Educ. Res. Rev., vol. 36, p. 100452, 2022, doi: https://doi.org/10.1016/j.edurev.2022.100452.
[5] V. Souza, A. Maciel, L. Nedel, and R. Kopper, “Measuring Presence in Virtual Environments: A Survey,” ACM Comput. Surv., vol. 54, no. 8, Oct. 2021, doi: 10.1145/3466817.
[6] A. Fraser, R. Hollett, C. Speelman, and S. L. Rogers, “Behavioural Realism and Its Impact on Virtual Reality Social Interactions Involving Self-Disclosure,” Applied Sciences (Switzerland), vol. 15, no. 6, Mar. 2025, doi: 10.3390/app15062896.
[7] C. Kyrlitsias and D. Michael-Grigoriou, “Social Interaction With Agents and Avatars in Immersive Virtual Environments: A Survey,” Jan. 11, 2022, Frontiers Media S.A. doi: 10.3389/frvir.2021.786665.
[8] R. Fauzi, M. Hariadi, S. M. S. Nugroho, and M. Lubis, “Defense behavior of real time strategy games: Comparison between HFSM and FSM,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 13, no. 2, pp. 634–642, Feb. 2019, doi: 10.11591/ijeecs.v13.i2.pp634-642.
[9] E. N. F. Dewi, A. N. Rachman, and R. H. Z. Hartadji, “Implementation of Hierarchical Finite State Machine for Controlling 2D Character Animation in Action Video Game,” in 2023 Eighth International Conference on Informatics and Computing (ICIC), 2023, pp. 1–6. doi: 10.1109/ICIC60109.2023.10381978.
[10] X. Liu, Z. Xie, and S. Jiang, “Personalized Non-Player Characters: A Framework for Character-Consistent Dialogue Generation,” AI (Switzerland), vol. 6, no. 5, May 2025, doi: 10.3390/ai6050093.
[11] M. Iovino, E. Scukins, J. Styrud, P. Ögren, and C. Smith, “A survey of Behavior Trees in robotics and AI,” Aug. 01, 2022, Elsevier B.V. doi: 10.1016/j.robot.2022.104096.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i2.28379

[12] R. C. Jain, "Multimodal Agents: From Vision to Reality," IEEE MultiMedia, vol. 31, no. 4, pp. 4–16, 2024, doi: 10.1109/MMUL.2024.3485253.

[13] G. Alves Da Silva, M. Wagner, and S. Ribeiro, "Development of Non-Player Character with Believable Behavior: a systematic literature review," in Anais Estendidos do XX Simpósio Brasileiro de Jogos e Entretenimento Digital, Porto Alegre, 2021, pp. 391–323. doi: doi:10.5753/sbgames_estendido.2021.19660.

[14] Q. Y. Yin et al., "AI in Human-computer Gaming: Techniques, Challenges and Opportunities," Jun. 01, 2023, Chinese Academy of Sciences. doi: 10.1007/s11633-022-1384-6.

[15] M. Ç. Uludağlı and K. Oğuz, "Non-player character decision-making in computer games," Artif. Intell. Rev., vol. 56, no. 12, pp. 14159–14191, 2023, doi: 10.1007/s10462-023-10491-7.

[16] M. Billah, A. Seher, A. Bahar, M. Ashfaq, and A. Hamid, "Game Brains: NPCs Intelligence Using Neural Network Brains," International Journal of Innovations in Science & Technology, vol. 6, no. 5, pp. 74–82, 2024, Accessed: Jan. 31, 2026. [Online]. Available: https://journal.50sea.com/index.php/IJIST/article/view/779

[17] S. K. Shukla, P. Jain, S. Rana, P. Pandey, V. K. Gupta, and R. Mohana, "A Comprehensive Overview of Software Quality Assurance for Computer Gaming," in 2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2024, pp. 749–754. doi: 10.1109/PDGC64653.2024.10984084.

[18] J. Chueca, J. Verón, J. Font, F. Pérez, and C. Cetina, "The consolidation of game software engineering: A systematic literature review of software engineering for industry-scale computer games," Inf. Softw. Technol., vol. 165, p. 107330, 2024, doi: https://doi.org/10.1016/j.infsof.2023.107330.

[19] C. Politowski, F. Petrillo, and Y.-G. Guéhéneuc, "A Survey of Video Game Testing," in 2021 IEEE/ACM International Conference on Automation of Software Test (AST), Madrid, Spain, 2021, pp. 90–99. doi: 10.1109/AST52587.2021.00018.

[20] J. Yu et al., "GameRTS: A Regression Testing Framework for Video Games," in 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), 2023, pp. 1393–1404. doi: 10.1109/ICSE48619.2023.00122.

[21] J. Bergdahl, C. Gordillo, K. Tollmar, and L. Gisslén, "Augmenting Automated Game Testing with Deep Reinforcement Learning," in 2020 IEEE Conference on Games (CoG), Osaka, Japan: IEEE, 2020, pp. 600–603. doi: 10.1109/CoG47356.2020.9231552.

[22] R. Tufano, S. Scalabrino, L. Pascarella, E. Aghajani, R. Oliveto, and G. Bavota, "Using reinforcement learning for load testing of video games," in Proceedings of the 44th International Conference on Software Engineering, in ICSE '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2303–2314. doi: 10.1145/3510003.3510625.

[23] H. Lee, "A unified conceptual model of immersive experience in extended reality," May 01, 2025, Elsevier B.V. doi: 10.1016/j.chbr.2025.100663.

[24] K. Kreijns, K. Xu, and J. Weidlich, "Social Presence: Conceptualization and Measurement," Mar. 01, 2022, Springer. doi: 10.1007/s10648-021-09623-8.

[25] D. Nalic, H. Li, A. Eichberger, C. Wellershaus, A. Pandurevic, and B. Rogic, "Stress Testing Method for Scenario-Based Testing of Automated Driving Systems," IEEE Access, vol. 8, pp. 224974–224984, 2020, doi: 10.1109/ACCESS.2020.3044024.

[26] D. Nalic, T. Mihalj, M. Bäumler, M. Lehmann, A. Eichberger, and S. Bernsteiner, "Scenario based testing of automated driving systems: A literature survey," in FISITA web Congress, 2020. doi: 10.46720/f2020-acm-096.

[27] F. Finkeldei, C. Thees, J. N. Weghorn, and M. Althoff, "Scenario Factory 2.0: Scenario-Based Testing of Automated Vehicles with CommonRoad," Automotive Innovation, vol. 8, no. 2, pp. 207–220, May 2025, doi: 10.1007/s42154-025-00360-0.

[28] Y. Chang, C. Xi, and Z. Luo, "Specific Scenario Generation Method for Trustworthiness Testing of Autonomous Vehicles Based on Interaction Coding," Applied Sciences, vol. 15, no. 19, p. 10656, Oct. 2025, doi: 10.3390/app151910656.

[29] V. Garousi, A. Rainer, P. Lauvås, and A. Arcuri, "Software-testing education: A systematic literature mapping," Journal of Systems and Software, vol. 165, p. 110570, 2020, doi: https://doi.org/10.1016/j.jss.2020.110570.

[30] K.-J. Stol and B. Fitzgerald, "Guidelines for Conducting Software Engineering Research," in Contemporary Empirical Methods in Software Engineering, Springer International Publishing, 2020, pp. 27–62. doi: 10.1007/978-3-030-32489-6_2.

[31] S. Pargaonkar, "A Comprehensive Review of Performance Testing Methodologies and Best Practices: Software Quality Engineering," International Journal of Science and Research (IJSR), vol. 12, no. 8, pp. 2008–2014, Aug. 2023, doi: 10.21275/sr23822111402.

[32] C. Politowski, F. Petrillo, G. C. Ullmann, and Y.-G. Guéhéneuc, "Game industry problems: An extensive analysis of the gray literature," Inf. Softw. Technol., vol. 134, p. 106538, 2021, doi: https://doi.org/10.1016/j.infsof.2021.106538.

[33] S. Shirzadehhajimahmood, I. S. W. B. Prasetya, D. Prandi, F. Dignum, M. Dastani, and G. Keller, "Automated Game Testing With Online Search Agent and Model Construction, a Study," Software Testing Verification and Reliability, vol. 35, no. 2, Mar. 2025, doi: 10.1002/stvr.70002.

[34] U. C. Türker, R. M. Hierons, K. El-Fakih, M. R. Mousavi, and I. Y. Tyukin, "Accelerating Finite State Machine-Based Testing Using Reinforcement Learning," IEEE Trans. Softw. Eng., vol. 50, no. 3, pp. 574–597, Mar. 2024, doi: 10.1109/TSE.2024.3358416.

[35] W. A. Ijsselsteijn, D. Kort, and Y. A. W. & Poels, "The Game Experience Questionnaire," 2013. Accessed: Jan. 11, 2025. [Online]. Available: https://pure.tue.nl/ws/files/21666907/Game_Experience_Questionnaire_English.pdf

[36] A. Somrak, M. Pogačnik, and J. Guna, "Suitability and comparison of questionnaires assessing virtual reality-induced symptoms and effects and user experience in virtual environments," Sensors (Switzerland), vol. 21, no. 4, pp. 1–24, Feb. 2021, doi: 10.3390/s21041185.

[37] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, in CHI '93. New York, NY, USA: Association for Computing Machinery, 1993, pp. 206–213. doi: 10.1145/169059.169166.