This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2026.3679121

Look2React: Making VR NPCs Come Alive with Dynamic Vision-Guided Reactions

Ritik Vatsal, Xincheng Huang, and Robert Xiao

(a) Waving (b) Cheering (c) Playing an Instrument

Fig. 1: *Look2React* NPC (right) reacting to different player actions (left). (a) NPC waves back to the player; (b) NPC starts cheering along with the player; (c) NPC starts dancing when the player plays an instrument

Abstract—A central promise of virtual reality (VR) games is the increased control players have over their character through pose and body language. However, many non-player character (NPC) systems fail to respond convincingly to these poses, user intent, and situational context, limiting immersion. We present *Look2React*, an interaction system that captures what NPCs see, using a vision-based reasoning model to select pose and text responses. *Look2React* endows NPCs with the ability to react dynamically and appropriately to player interactions. Through a gaze and proximity-based detection system inspired by stealth games, we trigger our system intuitively and only when intended, while also reducing resource costs. We invited 20 participants to play two versions of an RPG game: one with NPCs based on contemporary games and the other with *Look2React* NPCs. Our results demonstrate that *Look2React* increases engagement, leading to more frequent and repeated interactions with NPCs. Participants reported more satisfactory play sessions, significantly increased feelings of social presence, and felt that the dynamic reactions gave the NPCs more depth and personality — ultimately making them feel more human.

Index Terms—Games, non-player character AI, dynamic reactions, vision-based interactions

1 Introduction

A core promise of virtual reality (VR) is the control a player has over their experience. VR systems allow players to choose how they move, what they touch, and where they look, essentially giving them full rein over their character — a level of control rarely seen in traditional flatscreen games played with a gamepad or keyboard. Research shows that increasing levels of control in VR, particularly upper-body and full-body tracking, enhance users' sense of embodiment in terms of self-location and agency [22]. When combined with natural body positions and interaction modalities, this embodiment further strengthens immersion and enjoyment [27]. Even outside controlled lab environments, this capacity for agency tends to boost users' sense of presence when their actions directly affect the virtual world [32]. These benefits are well known, but current games still struggle to fully respond to this heightened embodiment, limiting the extent to which players' actions and agency are reflected in the virtual world [66].

Non-Player Characters (NPCs) are autonomous agents added to virtual worlds to make the environment more realistic, interactive, and socially engaging, providing players with responsive characters that populate the space, drive narratives, and create meaningful interactions [2]. These NPCs are points of interest for players exploring a new environment. Given the freedom afforded by most games, players will often try to interact with these NPCs in diverse and unique ways, from friendly waves to attacking with weapons, and this free-form interaction provides a great opportunity for dynamic reactions that will increase player immersion and pique their curiosity. However, current games offer limited reactions, often repeating the same preset dialogue and ignoring context, leaving players dissatisfied [8, 66]. With the rise of ML and AI technologies, researchers have used large-language models (LLMs) to simulate open-ended conversations with NPCs and have achieved notable success [8, 35, 51]. To increase believability, these LLMs need environmental context, and researchers have provided this by tokenizing in-game assets like inventory and interactable objects, and identifying simple gestures like pointing to highlight objects, allowing them to make more situationally-appropriate comments [9, 18, 29]. However, such methods have a high setup cost to tokenize each object and, more importantly, offer limited support for recognizing free-form player gestures.

To make NPCs truly human, there is a need for a system that is capable of identifying anything that the NPCs "see" the player do, without relying on predefined mappings of individual objects or a limited set of input gestures. Vision-Language Models (VLMs) have shown promise for understanding human interactions from visual inputs [33, 46, 68]. This capability makes them ideal for enabling contextual NPC reactions in virtual environments, allowing NPCs to respond contextually to player actions, gestures, and environmental cues, creating more natural interactions. Our study focuses on how we can use VLMs to create a system capable of generating real-time, believable, open-ended,

• *Ritik Vatsal, Xincheng Huang, and Robert Xiao are with the University of British Columbia. E-mails: ritikv@cs.ubc.ca, xincheng.huang@ubc.ca, brx@cs.ubc.ca*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

dynamic NPC responses for VR games. We start with the following research questions (RQ):

- **RQ1**: How can relevant pose data and user and environmental context be conveyed to a multimodal LLM for better player-NPC interactions?
- **RQ2**: What is the viability and performance of such a system in generating dynamic NPC reactions in role-play scenarios, based on provided environmental and user context?
- **RQ3**: What impact do dynamic NPC reactions have on players' immersiveness and engagement in VR games?

To answer these research questions, we present *Look2React*, a system that generates situational and dynamic NPC reactions just by "looking" at the player. *Look2React* takes rendered frames from within the VR world, capturing player actions and environment details, and returns a gesture selected from a set of available character animations, along with an appropriate textual reply. To test the effectiveness of this system, we created an RPG game and conducted a user study with 20 participants, which showed that *Look2React* greatly increases satisfaction, presence, and immersiveness. *Look2React* NPCs are perceived as more human, with their own personalities and traits.

The contributions of this work are -

- *Look2React*, a novel system to generate real-time, dynamic, personalised, and context-aware NPC reactions using vision in VR environments.
- A user study with 20 participants who played and compared two versions of a custom VR role-playing game, one modeled after NPC behaviours in recent RPG games and one with *Look2React*, showcasing dynamic NPC reactions.

2 Related Works

2.1 Non-Player Characters (NPCs)

NPCs are in-game characters with distinct personalities and roles. NPCs often play a major role in the progression of the game (e.g., quest-giving NPCs), provide in-game services (e.g., a shopkeeper NPC), and can function as allies or foes [15, 58]. Seasoned players have certain expectations of NPCs, and thoughtful design to meet these expectations leads to greater immersion and engagement with the environment [56, 57]. In most cases, the first encounter with NPCs is visual, which means that visual design alone can strongly cue narrative role and traits, affecting perceived immersion. To make encounters realistic, precise character design and research-driven workflows map trait expectations to silhouettes, costuming, and static body language so players can estimate if the NPC is "friend or foe" at a glance before any animation or dialogue is added [26, 43, 44]. The way NPCs are scripted and the personality they have also affect immersion and long-term emotional investment [4]. Designers craft primary NPC arcs to intertwine with the main plot, promoting repeat contextual interactions while using secondary NPCs for side-quests, exposition, and contextual hints to balance a sense of realism and belonging to the world, rather than making the world revolve around the protagonist [24, 37]. Well-implemented NPCs make the players return to relive the experience, increasing replayability and connection with the game world [53].

The research in NPCs is situated in the intersection of social science, game studies, and Human-Computer Interaction. Prior social science and HCI research suggests that people naturally apply social rules to interactive media and computer agents [34, 41]. A potential outcome of more lively game NPCs is social presence, which is influenced by contingent behaviour, perceived agency, and anthropomorphic cues [6, 36]. Game studies also show that players may cultivate parasocial friendships and emotional attachment to carefully designed NPCs, leading to enhanced engagement and motivation in gaming [6, 64]. For example, Yin states that by reflecting on how NPCs guide, challenge, assist, and contextualize player experiences, practitioners in adjacent domains (e.g., sports coaching) can derive actionable insights to enhance engagement, motivation, and immersion in real-world settings [64]. As these systems become more human-like, Yin et al. examine how players form and adjust trust in NPCs that can lie, showing that initial default trust is calibrated through dialogue, context, and verification behaviors, and that deception meaningfully reshapes players' interaction strategies and takeaways about NPC reliability [65]. These interactions also give valuable insights into Human–AI interaction by surfacing design patterns, such as responsiveness, transparent communication cues, calibrated emotional signaling, role-consistent behaviors, and supportive team structures, that help humans coordinate, build calibrated trust, and remain engaged with artificial partners in complex tasks [59].

2.2 Dynamic NPCs and VR

At the core of all NPC interactions is the NPCs' ability to react to player input. Early NPCs relied on hand-authored decision trees and finite-state machines (FSMs) to select between attack and idle states [14, 49]. However, as games became more complex, these methods quickly became unscalable and infeasible. Yahyavi et al. analyzed how such static schemes diverge from human behavior in first-person shooter (FPS) contexts and proposed "pheromone maps": influence fields tied to items, events, and locations which drive more adaptive, human-like decisions in tasks such as goal selection, ambush setup, and area defense [60].

Large-language models (LLMs) have recently been used to augment the design and implementation of NPC characters. Stegeren and Myśliwiec fine-tuned GPT-2 on annotated RPG quests to generate NPC dialogue that aligns with quest structure and role, demonstrating early success of LLMs to shape voice, intent, and consistency in NPC interactions [54]. Park et. al. also show how LLM-driven action selection and memory recall yield NPC-like behaviors that feel intentional and consistent over time [38]. As LLMs improve, researchers have begun to use them as real-time decision makers for unconstrained and dynamic NPC reactions [12, 18].

VR is an ideal platform for deploying realistic NPCs due to increased immersiveness and enhanced controls when compared to flatscreen games [52, 55]. Korkiakoski et al. created GPT-4-powered NPCs in a VR speech-based interrogation simulator. They showed good usability and moderate believability, but noted an average interaction latency of around 7 seconds, which rises as conversation context increases [23]. Other researchers have used speech in AR/VR scenarios for sentiment analysis and to create more natural ways to interact with NPCs [8, 35, 51]. To create NPCs that are aware of the environment and other necessary context, not just limited to speech, Li et al. proposed a prompting schema that lets virtual agents "perceive" VR scenes (spots/locations, objects, characters, system context, and communication cues) by passing object identifiers to an LLM, which allows them to generate environment-aware interactions and dialogue [29]. However, these systems are unable to "see" the player in real-time and either rely on tokenised descriptions of the environment, leading to significant setup overhead, or drop this context entirely, focusing largely on speech as free-form input [8, 47].

2.3 Vision-Based Activity Recognition

Using vision to analyze human motion is not a new concept; Johansson (1973) recorded participants in a dark room, with LEDs stuck to the body joints, to create early representations of the 2D human body in motion [19]. Yamato (1992) used Hidden Markov Models on time-sequential images of tennis players to recognize human actions (e.g., tennis strokes), establishing a probabilistic temporal framework for gesture recognition from video [61]. Since then, several frameworks utilizing motion and space-time trajectories have been developed to analyze motion gestures from multi-temporal images [1, 21, 31, 50]. Ryoo and Aggarwal (2009) proposed a framework for automated recognition of 8 complex human actions (approach, depart, hug, punch, kick, push, and handshake) from image sequences [45].

Recent research uses deep learning with large activity datasets and convolutional networks for more robust activity recognition [10, 11, 39, 42, 62]. These methods achieve high accuracy in identifying up to 60 different actions, making them useful for inferring user action context in VR scenes. However, skeleton-based activity recognition means that items equipped in the environment are omitted, which can misclassify

contextually distinct behaviors [67]. For example, a gesture detected as a "wave" may actually be a player waving a weapon, significantly reducing situational accuracy.

Vision-Language Models extend skeleton/action pipelines by jointly grounding visual input and text, enabling open-ended scene understanding and action description, which is useful for a more complete perception of virtual environments [13, 68]. Researchers have used vision-language-action models to create Embodied AI that takes visual inputs and language goals and outputs actions, unifying perception and decision-making for language-conditioned tasks like navigation, manipulation, and household routines [30]. In the physical space, VLMs are being rapidly adapted for human–robot interaction to perceive human actions, interpret instructions, and choose socially appropriate actions in near-real time [33, 46, 48]. This shows that VLMs are well-suited for creating responsive and dynamic NPC reactions in virtual environments.

3 LOOK2REACT

In this section, we go over the creation of *Look2React* and how each component works. We start by discussing model selection, which determines several facets of the architectural design. We then illustrate the interaction design and player perspective within *Look2React*. Finally, we discuss the inputs to the system and how each component is implemented.

3.1 System Architecture

Model selection plays a crucial role in an AI system, as the specific capabilities and strengths of the model chosen can affect the design and architecture of the overall system. Although these capabilities are converging over time, current models expose diverse capabilities that can prevent them from being freely interchangeable.

Initially, we envisioned a two-stage system, to be run fully locally to minimize latency: first, our system captures user actions in the virtual environment, sending them to an "observer" model to produce a description of these actions along with any notable contextual details. Next, the observer's output would be fed into a reasoning model, which evaluates the possible NPC poses, selects the most appropriate one, and generates a corresponding NPC comment in response to the first model's analysis.

We initially experimented with pose-based action-inference models for the observer model, such as MM-Skeleton [63], but quickly ran into limitations. First, the use of inverse kinematics (IK) to reconstruct poses in VR from sparse data (head and hand pose) meant that the observed poses were not exactly accurate, leading to reduced activity recognition accuracy. Secondly, the inability to capture virtual objects meant that important contextual cues — such as items held by the player — were lost.

We then tried using image-captioning models for the observer. Local captioning models, such as BLIP [28], provided good understanding of the environmental context, correctly identifying the player and the location the NPC was in, as well as recognizing specific in-game items such as guitars and swords. However, such models often failed to capture the players' gestures and intentions. Large local models also competed with the VR game for resources, causing degraded performance for both.

The key takeaways from these experiments were that a larger model could better understand the VR environment, and that it would need to be deployed on a separate system to avoid destabilizing the game. This led us to finally select cloud-based state-of-the-art VLMs. These VLMs integrate visual and reasoning capabilities, enabling us to merge the "observer" and "reasoning" stages into a single cloud model API call. After testing various hosted models, we opted to move forward with OpenAI's GPT-5 [1] and Google Gemini Flash 2.5 [2] due to their strong performance in image reasoning tasks.

[1] https://platform.openai.com/docs/models/gpt-5-mini
[2] https://deepmind.google/models/gemini/flash/

3.2 Model Input

The full prompt structure provided to the VLM model is shown in Table 1.

Table 1: Structure of the prompts, with descriptions and examples.

Component	Example
`npc_type` Specifies the type of NPC	stern_guard
`animations` A numbered list of available gesture options for the NPC.	0: Wave 1: Dismiss 2: Dance 3: ...
`response_format` Defines the required structure of the reply	gesture_index, statement, optional_pose
`prev_response` Provides the last statement this NPC made for context. Defaults to `None` if first interaction. Ensures varied responses to similar consecutive actions.	No performances here musician!
`images` 3 images from the VR environment	
`System message` Instruction to the model.	You are an NPC in a VR RPG game. Given an NPC type, images, animation list, and the last statement it made, reply with the index of the most appropriate animation strictly according to the NPC type and a new under 5 word statement the NPC makes. Optionally, include a better response pose name. Do not repeat statement. Do not explain.

One of our goals with *Look2React* was to ensure that the system requires minimal re-configuration for different NPCs, which also promotes caching of the prompt and context. To that end, we ensured that only one field sent to the VLM was NPC-specific. We abstracted model calls using a model-agnostic pipeline that works for both chosen VLM models, ensuring portability and the ability to rapidly experiment with different models.

3.3 Interacting with *Look2React* NPCs

Triggering the Interaction: In order to provide useful context to the VLM, without excessive model calls, we needed a way for the player to reliably and intuitively "trigger" NPCs, prompting them to begin observing the player's actions and responding to them. Such a trigger needs to feel realistic while avoiding accidental triggers (e.g. NPCs reacting to actions not directed at them). Triggers can be either implicit, triggering when players satisfy certain conditions, or explicit, triggering when players perform a certain action.

We chose a hybrid approach to triggering, using a combination of an implicit trigger — proximity to the NPC — combined with an explicit trigger — eye contact with the NPC. We considered voice and gesture triggers as well, but decided against them as player voice interactions are rare in VR, and gestural triggers would conflict with and interrupt player actions.

Player Feedback: To make this system intuitive and relatable, and to provide clear feedback of NPC state to the user, we took inspiration from stealth-based video games where player 'detection' is a major

component. Typical implementations provide NPCs which are either unaware of the player, fully aware of the player (and thus, for instance, chasing them or raising an alarm), or *partially aware*: in a state of heightened awareness due to player activity, but not yet fully aware of the player themselves.

Games employ various visual methods to indicate a partially aware state, but a common implementation is a loading-style indicator that gradually fills as the player continues to raise suspicion. The player must correct their behavior or evade the NPC before the bar fills completely, causing the NPC to enter the fully aware state. This visual feedback is easy to understand, and we adapted a similar approach for our system.

Interaction Model: Based on these design principles, we designed a four-step interaction system:

1. **Idle**. In the idle state, the NPC has a closed eye icon above their head (Fig. 2 (a)), signifying that they can be interacted with. The player moves towards the NPC; when they are close enough (within 5 meters), the NPC will turn to face them.
2. **Looking**. Next, the player gazes at the NPC. We use a raycast from the camera to the NPC's body to estimate gaze. This causes the eye icon to open and turn white. It will begin filling up at a fixed rate, during which the NPC records the player's actions (Fig. 2 (b)). If the user walks away or breaks their gaze before the eye fills up, the NPC reverts to the idle state. To make this more natural, gaze directed at any part of the NPC is registered as a valid interaction.
3. **Thinking**. After the eye fills completely (Fig. 2 (c)), which takes 3 seconds, the NPC strikes a thinking pose for a few seconds while it prepares a response to the player's action.
4. **Reacting**. Finally, the eye turns red (Fig. 2 (d)) and the NPC enacts the response: performing a gesture and displaying an accompanying textual message. After a short delay, the NPC reverts to the idle state (step 1), ready to interact further.

Fig. 2: *Look2React* interaction sequence. (a) Idle, (b) Looking, (c) Thinking, (d) Reacting.

3.4 Environment Images

To capture images of the user and the environment, we added an additional virtual camera to the scene, which captures what the NPCs would see from their perspective. We capture 1 image every 500 ms, starting from when the player initiates step 2 ("Looking") till they look away, or reach step 3 ("Thinking"). To balance environmental context and resource use, only the last 3 images are sent to the VLM. In our testing, we occasionally waited to confirm the interaction trigger (eye opening, turning white), thus not starting the gesture immediately after the trigger, so we decided to use the last three images to allow players with a short buffer if they want to reposition themselves from the interaction trigger to the action.

The camera has a wide field of view. To reduce token costs, we run an algorithm to crop the images to focus on the player. To prevent player clipping and ensure a consistent view of the player in the frame, we avoided static measures like centering the player, and instead use Google MediaPipe Pose [3] to detect the player in the frame. We add a padding of 100 pixels to all 4 sides of the player bounding box to cover sufficient environmental context. Running MediaPipe is relatively inexpensive, with the complete cropping process taking 3–4 ms per image. This cropping results in images that are at typically 3 times smaller than the original images, reducing token costs. With 3 full-size images, one *Look2React* API call to `gpt-5-mini` takes roughly 3500 tokens, whereas with 3 cropped images, this number reduces to around 1400 tokens.

3.5 Gestures

Body language, gestures, and physical expressions contribute significantly to human interactions. Players often report that NPCs appear stiff or rigid when their body movements are disconnected from their dialogue. Prior work has addressed this by having a predefined set of basic gestures and postures, from which LLMs can select the most contextually appropriate option to accompany conversational exchanges in narrative-centered games [25]. These results highlight the capacity of LLMs to map dialogue to suitable nonverbal behaviors when constrained to a finite list of actions.

In *Look2React*, we provide the VLM with a curated list of animations to choose from, enabling more expressive and contextually appropriate responses. These animations are sourced from Mixamo's catalog [4]. This list of animations remains the same for all NPCs, giving the VLM responsibility for not only choosing the most situationally accurate gesture, but also considering the NPC's personality in its decision. In our prototype, we provided 12 gestures in total: Arm Stretching, Intense Cheering, Clapping, Dance, Dismissing Gesture, Tasty Eating, Aggressive Fight, Greeting Wave, Informal Bow, Scared Shock, Terrified, and Thankful.

3.6 Gesture Suggestions and Generation

Although VLMs can employ the existing list of gestures to create convincing responses, we are aware that a limited set of gestures may become monotonous and repetitive over time. To mitigate this, we asked the VLM to include an additional "suggested gesture" parameter in its response. In this field, the model can optionally suggest a gesture it considers more suitable for the given context, beyond the available set of animations. Developers can then use this list to select and implement the most frequently appearing suggested gestures in future game patches.

We initially planned to use these suggested animations to directly generate new gestures on the fly within *Look2React*. Recent research has shown success in generating 3D animations from textual descriptions, and some systems also offer real-time poses [3, 16, 69]. We tried two such systems, Text2Motion.ai [5] and EMDM [69]. We found that generating complex and situational gestures required multiple iterations and prompt engineering, and would require a human in the loop. For this reason, we dropped the generative pose aspect in the current version.

With human intervention, we were able to create high-quality animations using Text2Motion.ai, needing around 5 minutes per animation. The Unity integration meant that animations could be seamlessly plugged into Unity. Alternatively, we also tried extracting pose data by finding videos of people doing our target gestures and converting that data into 3D animations. We achieved similar performance in some cases, but this method relied heavily on video data, which may not be open-source. Although Look2React currently doesn't incorporate dynamic gesture generation, this exercise allows us to estimate the time it will take to retroactively add additional animations to the game, making the exploration valuable.

[3] https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker
[4] https://www.mixamo.com/
[5] https://www.text2motion.ai/

4 User Study

4.1 Stimulus - RPG Game

To test our NPC system, we created a VR game set in a medieval village. The main mission is a typical fetch quest in which one of the NPCs sequentially asks the players to bring him two items (ordered randomly) — a guitar (lute) and a sword.

The players start about a third of the way down a path with fences on both sides and an obstacle on both ends. The players are told that everything they need to complete the quest can be found within this perimeter. However, they are not restricted and are free to explore the hilly jungle that surrounds the area. On the path, there are 4 points of interest:

1. **Village Girl:** The first NPC the players encounter, positioned nearest to the player's starting point — a playful village girl in front of her house (Fig.3 (a)). Prompt: *young_girl*
2. **Town Guard:** The second NPC — a stern and tall guard monitoring the village (Fig.3 (b)). Prompt: *serious_guard*
3. **Old Man:** The final NPC, located at the end of the path — an old man that gives the quest to the players (Fig.3 (c)). Prompt: *mysterious_old_man*
4. **Table with Items:** A table with several different items is placed at the beginning of the path. Players can equip one item at a time. The table has a lute/guitar, a cake, a bouquet of flowers, and a sword (Fig.3 (d)).

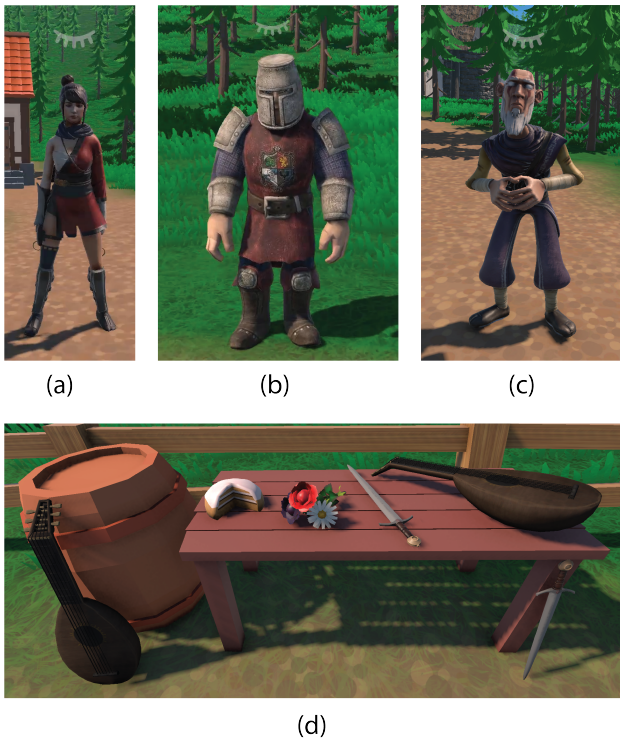


Fig. 3: Points of interest in the RPG game. (a) village girl, (b) town guard, (c) old man, and (d) table with items.

Players start near the Village Girl, and need to traverse the path to the old man to initiate the fetch quest. They must then walk back to the table, pick up the requested item, bring it back to the old man, and repeat the process for the second requested item. In this way, the participant is expected to travel across the path at least five times, meeting each of the other NPCs on the way.

For movement in the space, we decided to use teleportation, since research has shown that teleportation is less likely to induce nausea in VR [7, 40], and most major VR RPG games either use teleportation or give an option to choose between teleportation and joystick locomotion. In our game, pushing forward on the joystick on either Quest controller brings up an arc that the players can aim at the ground, and letting go will teleport them to their target. For ambient sound, there is medieval-themed music with sounds one would expect in a village (birds chirping, dogs barking, etc.).

4.1.1 Control Version

For the control (baseline) version of the game, we wanted NPCs' interactions to be similar to those of contemporary popular RPGs. We looked at the top VR and flatscreen RPG games, including Skyrim VR [6], Fallout 4 VR [7], Asgard's Wrath 2 VR [8], The Legend of Zelda - Tears of the Kingdom [9], and Dragon's Dogma II [10]. In most games, interactions with NPCs are triggered by pressing a specific interaction key. Optional NPCs within these games typically have fixed, short messages they cycle between at every interaction, in the absence of any plot-driven content. NPCs sometimes look at the player, but don't have any other gestures or animations, often not even addressing the player unless scripted to otherwise. In some games, NPCs will respond with a generic frightened animation if they are attacked.

For our game, we implemented a similar system, where the interaction is triggered by pressing the 'A' key on the right Quest controller. For the two optional NPCs (Village Girl and Town Guard), text responses are selected randomly from a list of phrases taken from similar NPCs in the reference RPGs. For the Old Man, the dialogue is related to the quest. Text responses are given instantly upon interaction. There are no pose responses in this version. The NPCs turn to face you when you are close to them. Everything else remains the same in the two versions.

4.1.2 *Look2React* Version

In the *Look2React* version, the main difference was that all three NPCs were now using *Look2React* interactions. Consequently, this meant that interactions with the NPCs were no longer triggered by key press, but by the *Look2React* triggering mechanism (§ 3.3). To facilitate the fetch quest, the Old Man NPC's dialogue remained the same, although he could respond physically using VLM-recommended gestures.

The VLM model chosen for the user study was OpenAI's `gpt-5-mini`, due to more consistent timing and diversity of responses. For more details, please refer to §5.1.

4.1.3 Apparatus and Setup

The game is built in Unity Engine (v6000.1.3f1). The system used has an Intel i9-12900K CPU, Nvidia 5090 GPU, and 64 GB of RAM. We used the Meta Quest 3 VR headset for VR with a refresh rate of 72 Hz. Participants with prescriptive glasses could wear the VR headset over them and did not need to remove their glasses.

4.2 Ethics

The study was approved by the Behavioral Research Ethics Board at the University of British Columbia (H19-02782). All participants were provided with an overview of the study, and written consent was obtained before the study was conducted.

4.3 Participants

Participants were recruited through word-of-mouth and social media posts across multiple channels. A total of 21 participants took part in the lab study; however, one participant felt nauseous during the first VR session, which led to an early stop and their data being dropped. A total of 20 participants completed the study (6 female and 14 male). 9 participants were between 20-24 years of age; 10 were 25-29 years of age; and 1 was 40-44 years of age. One participant (P13) had no prior experience with either flatscreen games or VR, but expressed a

[6] Skyrim VR – Bethesda
[7] Fallout 4 VR – Bethesda
[8] Asgard's Wrath 2 official page – Sanzaru Games
[9] The Legend of Zelda: Tears of the Kingdom – Nintendo
[10] Dragon's Dogma II – Capcom

keen interest in trying VR experiences; all other participants had some level of VR experience. Two participants had no experience with RPG games.

4.4 Study Procedure

Each participant played both versions of the RPG game, as described in § 4.1, with the order counterbalanced to avoid ordering effects. Participants were explained the controls of the game and the general objective — to explore and complete the fetch quest. Participants then adjusted and wore the VR headset so that they were comfortable. For participants new to VR or RPG games, or on request, an in-game walkaround was also provided so that the participants could familiarize themselves with the controls and ensure that there were no immediate motion sickness symptoms.

After this, the participants were allowed to play the game for as long as they wished, with the researchers only informing them when the main goal of the game was achieved and when they had passed 10 minutes and had no further content to explore. After each VR session, participants were asked to fill out the Game Experience Questionnaire (GEQ), which has two components, the In-game module (14 questions) and the Social Presence module (17 questions) [17]. Participants were offered an optional 2-minute break between sessions.

After both sessions, the lead researcher conducted a semi-structured interview to gather the participants' views and opinions on the two versions of the game, as well as anything that stood out to them. Finally, we prompted the participants to ask any questions they may have had about any part of the study. Following the interview, participants were thanked for their participation and compensated for their time ($16 CAD).

5 Results

5.1 VLM Performance Metrics and Selection

To select the best VLM for our user study, we compared 2 models: OpenAI `gpt-5-mini` and Google `gemini-2.5-flash`. Both models are known for their strong vision-reasoning capabilities at the time of the study, and we sought to select one that balances the trade-off between latency, performance, and stability.

For the pose selection and text generation tasks, both models performed well and were comparable across all the input actions we tried. However, response times with Gemini were not consistent. Most Gemini API calls would take around 6–7 seconds to complete, which is similar to GPT; however, frequently the response times would go over 20 seconds, going as high as 30 seconds in some instances. This happened often enough that we could not consider Gemini as a reliable VLM for conducting user studies, hence `gpt-5-mini` was used. Response times for `gpt-5-mini` were between 5–9 seconds (M = 7.64 s, SD = 1.40 s).

Since these are API-based models, token usage is a significant metric. One *Look2React* API call to `gpt-5-mini` takes around 1400 input tokens (M = 1347.0, SD = 116.5) and 11 output tokens (M = 11, SD = 1.4); at current prices, and without assuming any caching, this would cost about $1 USD per 3000 NPC interactions.

5.2 Interaction Frequency

All players played the *Look2React* version of the game for much longer compared to the control version. The average *Look2React* session was 10 minutes and 39 seconds long (M = 639.15 s, SD = 232.00 s), and the average control session was 4 minutes and 35 seconds (M = 275.85 s, SD = 99.82 s); note that we gently reminded participants after 10 minutes to keep the study running on time.

However, since each *Look2React* interaction took significantly longer than control interactions (which were instantaneous), we can't draw any insights from session lengths alone. To supplement this, we also performed an interaction analysis on the total number of interactions and repeat interactions. We define a repeat interaction as a case where the participant returned to the same NPC after having walked away to interact with another NPC, collect a prop, or for any other reason.

Participants in the *Look2React* version had more total interactions (M = 20.8, SD = 6.7) on average compared to the control version (M = 10.8, SD = 5.2). Repeat interactions followed a similar trend, with an average participant returning to the same NPCs around twice as often as in the *Look2React* version (M = 11.1, SD = 3.2) than in the control version (M = 5.8, SD = 3.4).

5.3 GEQ - Social Presence Module

The social presence module has 17 questions that aim to measure how strongly players feel connected to, aware of, and engaged with other characters or agents in the virtual environment. All questions were scored on a 5-point Likert scale. For example, "I empathized with the other(s)," where `(0)` corresponds to "not at all" and `(4)` corresponds to "extremely."

For social presence, all participants reported significantly higher scores across all questions for the *Look2React* version, as shown in Fig. 4 (a). Following standard practice, we split the 17 questions into three segments — behavioral engagement, empathy, and negative feelings [17]. To assess the statistical significance of these values, we first performed a Shapiro-Wilk normality test. Since all values were non-normal, we conducted non-parametric tests for all conditions. Effect sizes are also reported to indicate the practical significance of the differences. For our within-participants design study, we applied the Wilcoxon signed-rank tests for non-parametric data. All results and p-values are given in Table 2.

For Social Presence, all three subscales (Behavioral Engagement, Empathy, and Negative Feelings) showed significant increases in the *Look2React* condition. Wilcoxon signed-rank tests revealed strong effects for each subscale, indicating that participants consistently felt stronger social presence in *Look2React* compared to the control condition. The Rank-Biserial Correlation (RBC) was 1 for all three, and the Common Language Effect Sizes (CLES) ranged from 0.854 to 0.903, reflecting a clear and practically meaningful increase across all subscales. The increases in Empathy and Behavioral Engagement were particularly strong, highlighting that *Look2React* substantially enhanced social connectedness and interaction with NPCs.

5.4 GEQ - In-Game Module

The In-game module has 14 questions that aim to measure players' subjective experiences during gameplay, including aspects such as immersion, flow, competence, tension, and challenge. The questions are marked on a 5-point Likert scale, similar to the Social Presence module. The questions and participant average scores are given in Fig. 4 (b). All questions were found to be non-normally distributed, and we applied the Wilcoxon signed-rank test to them. All results are given in Table 3, with significant results highlighted.

Most notably, players reported higher contentment with the *Look2React* version. Players also reported higher challenge and effort required in *Look2React* version, but the significant increases in Q1, Q3 (decreased), Q4, and Q14 scores suggest that the system encouraged them to try new and different things, ultimately leading to a more dynamic and rewarding gameplay experience.

5.5 Qualitative Findings

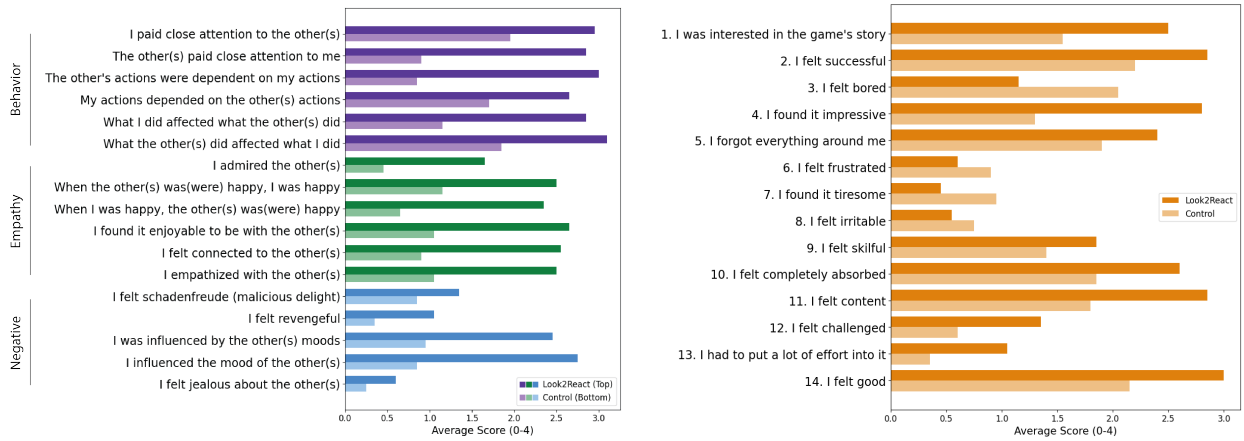
Our semi-structured interviews yielded approximately 5 hours of interview recordings. To analyze the data, we coded the interview transcripts and organized the feedback into affinity diagrams and higher-level categories. We then synthesized them into overarching themes and report them as follows. We begin with participants' immediate feedback on the NPCs and in-game interactions, then turn to their reflections on how *Look2React* might shift the gaming paradigm at a higher level. Finally, we conclude with participants' suggestions for potential improvements.

5.5.1 NPCs as "real people": Increased expressiveness, emotion, and personalities

Participants generally noted that they perceived the *Look2React* NPCs more as "real people". From their feedback, we identified three factors contributing to this resemblance: (1) the NPCs' expressiveness, (2) their perceivable emotions, and (3) their more distinctive personalities.

Table 2: Shapiro–Wilk and Wilcoxon signed-rank results for Social Presence subscales.

Variable	Shapiro–Wilk (W, p)	Wilcoxon (W, p)	RBC	CLES
SP_Behavior	0.939 (0.032)	0.0 (0.000002)	1.0	0.9038
SP_Empathy	0.938 (0.029)	0.0 (0.000142)	1.0	0.8575
SP_Negative	0.937 (0.027)	0.0 (0.000212)	1.0	0.8538



(a) Social Presence Module

(b) In-Game Module

Fig. 4: Game Experience Questionnaire Results

Table 3: Wilcoxon signed-rank test results for GEQ In-Game Module items. (** Highly significant ($p < 0.001$), * Significant ($p < 0.05$)).

Question	Wilcoxon (W, p)	RBC	CLES
Q1	0.0 (0.0007)**	1.000	0.763
Q2	26.5 (0.0541)	0.558	0.588
Q3	22.0 (0.0090)*	-0.712	0.281
Q4	0.0 (0.0003)**	1.000	0.836
Q5	27.0 (0.0495)*	0.550	0.624
Q6	25.5 (0.2991)	-0.346	0.455
Q7	4.5 (0.0150)*	-0.836	0.365
Q8	6.5 (0.4568)	-0.381	0.431
Q9	18.0 (0.0425)*	0.604	0.613
Q10	6.0 (0.0022)*	0.886	0.700
Q11	0.0 (0.0005)**	1.000	0.778
Q12	13.5 (0.0236)*	0.703	0.668
Q13	3.5 (0.0140)*	0.873	0.716
Q14	5.0 (0.0022)*	0.905	0.738

Regarding expressiveness, all participants observed that the *Look2React* NPCs reacted meaningfully to the players' behavior:

> *P6: I tried to play the guitar [to the girl]... She danced and said "singing my song". I think that was a kind appreciation.*

The richer expressiveness of the NPCs generally led to their emotions being more perceivable, as confirmed by our quantitative results ("I empathized with the others" and "I was influenced by others' moods" in Figure 2). Participants' feedback shed light on the moments when they empathized with the NPCs, sensing their happiness, fear, or hunger:

> *P3: She [the girl] said, "want some cake" with an eating gesture, so I felt her desire for the cake.*

> *P12: I can see my dance being mirrored by the girl... I was able to capture, say, her happiness very fast.*

Eventually, when comparing the NPCs' personalities in the two versions, all participants found that the NPCs in the *Look2React* version generally exhibited more pronounced personalities:

> *P21: The girl has like a very upbeat personality because like she would sing a song with you while you are playing the guitar... so like a very outgoing person.*

Most participants (N=17) also noted that the NPCs' personalities were more distinct in the *Look2React* version, particularly when they reacted differently to the same gesture:

> *P7: When I held up the sword to her (the girl), she got scared and backed away, when it's the guard [he] like wanted to fight so you could see that there's definitely a personality difference.*

5.5.2 Realistic Immersion: more fluid conversation, intrigued exploration, deeper engagement

Having established that participants perceived the NPCs more as "real people", we can now examine what this implies for the players' in-game experience. Because participants were able to interact with the NPCs in ways that resembled real life, they generally reported that the experience was more fluid, encouraged exploration, and was more engaging.

Increased fluidity in the NPC interactions mainly arose from being able to use real-world social protocols to interact with them:

> *P4: It's much easier to hand in the quest. Like I just walked there and the old man just say okay... It's more like in the real world.*

Given that most existing games require players to press a specific key (e.g., a button) to interact with an NPC, it was notable that, in the *Look2React* condition, participants spontaneously used real-life social protocols (e.g., walking closer and showing an item) to hand in the quest to the old man, without any explicit prompting from the experimenter. We believe this behavior in the *Look2React* condition indicates increased immersion and realism. For example, P12, who self-identified as a frequent gamer, commented:

P12: when I was holding the item [the old man requested], I knew exactly that I didn't need to press the buttons after returning to him... I can just grab stuff and just return back there and wave it at him.

Echoing the fact that participants all spent more time in the *Look2React* version, participants' feedback indicated increased willingness to explore and deeper engagement with the game. Because the NPCs provided unscripted responses (i.e., not explicitly programmed) that depended on players' spontaneous actions, participants reported feeling more curious about interacting with both the NPCs and the game overall:

P1: The NPCs are more capable, that intrigued me to do more things, to see what their reactions were.

P19 compared this to how players often skip through conversations with NPCs in existing games, and noted:

P19: ...because a lot of the existing games, it feels more like trying to skip through the dialogue... That response is varied, so I feel more willing to like try out different responses or even do the same action but do it like twice to see how how it might change.

Oftentimes, this willingness to explore allowed players to enjoy the game beyond the main quest. When asked whether they approached the game differently in the two versions, P3 and P9 explicitly noted that they focused less on the main quest in the *Look2React* version, which aligned with our observation:

P3: Other than finishing the task, I can explore more in the world by interacting with the NPCs.

P9: In the AI version I could spend hours other than the main story just interacting with the NPCs.

5.5.3 Heightened social anticipation: Satisfied intent vs. unexpectedness

In line with the quantitative results from the social presence module (Figure 2), we observed heightened social awareness and anticipation from the participants in the *Look2React* version. As participants treated the NPCs more like real people and engaged more deeply with the game, they became more aware of the intent behind their own actions and anticipated that the NPCs' responses would align accordingly. Overall, feedback was highly positive when the NPCs in the *Look2React* version responded exactly as players expected:

P16: I feel more immersed when the interactions were sort of in line with what I expected.

P17: I would do some gestures like the bouquet of flower or wave the cake, [I would expect] like sniffing... and something like taking a bite of the cake. So basically I would interact with the props and the reactions of the NPCs would be consistent with my gestures.

However, what we observed was even more intriguing: moments when the NPCs' responses did not align with players' expectations. For example, the young girl in the game typically responded warmly to gifts and would even dance along when a participant played the guitar—behaviors that often matched participants' anticipation. In contrast, we gave the guard a stiffer persona: he consistently rejected gifts and appeared more indifferent. A few participants considered the guard's indifference somewhat unexpected. For example, P1 tried to bow to the guard and mentioned he expected him to bow back. However, the guard brushed his hand through the air in a shooing gesture. Interestingly, when players received such unexpected responses, it often strengthened their perception of the character. For instance, P7 and P12 reflected on how they felt about the guard when their kindness was not reciprocated:

P7: He was mean. He didn't want any of my gifts and he only wanted to fight. So he's the big bad guard.

P12: ...the guard has a very strong personality. Yeah, he is not a person I think a player will like, very rude.

5.5.4 From "role-play" to "co-play": Unscripted and spontaneous plots add to the game's possibilities

Building on the above observations, some participants reflected on how *Look2React* might shift the gaming paradigm for NPC-driven games. An important takeaway was the possibility for players to spontaneously co-create their own stories with NPCs beyond the main narrative. Several participants described an increased sense of influence over the game:

P7: I felt like I actually had an influence on the environment... I am not just trying to work with what the game was giving me, but I am able to provide input to the game, and the game is responding based off what I was doing... sort of like I'm role-playing and the game is kind of role-playing back to me.

By enabling players to influence how NPCs "co-play" in response, *Look2React* went beyond simply providing human-like replies to isolated inputs. We observed that these more lively responses were often connected by players into emotional and memorable mini-plots. Echoing the theme of "unexpectedness" noted in the previous section, such mini-plots could even emerge from moments when players' intentions were left unfulfilled. Coincidentally, P6 and P7 shared similar stories about their "unpleasant" encounters with the guard:

P6: In front of the guard, first I showed friendliness. I wanted to say, Ok, I am not aggressive, I will give you some food. However, the guard was not showing the same friendliness to me. So I went back and picked up the sword and tried to attack him.

P7: So a good example is the big bad guard. At first I tried to make friends with him. I brought him a cake and flowers and he didn't want any of that. He didn't want anything to do with me. So that made me frustrated and angry. I grab a sword and I went up to him and threaten him with the sword. You know... then he wanted to fight. I think I would have won...

6 Discussion

6.1 Summary

To address RQ1, we experimented with several state-of-the-art methods and finally concluded that vision-language models are the best suited to balance relevant context with minimal pre-preparation. Building on this, we created *Look2React*, a model-agnostic system that uses such VLMs to select relevant gestures and generate accompanying text as NPC responses to user actions in VR environments. For RQ2, we report both objective performance (response latency, token usage, and interaction analysis) and subjective viability (perceived personality, immersion, and engagement) of *Look2React* to help create a more complete understanding of the system's capability. Finally, for RQ3, we explored how the participants felt by analyzing their game experience questionnaires and interview responses.

6.2 User Experience

We found a strong and statistically significant increase in presence, immersion, engagement, and satisfaction with *Look2React* interactions when compared to the control version from the GEQ and the interviews. The In-Game module highlighted that dynamic NPC reactions increase interest and impressiveness, at the cost of increased effort and challenge. In the interviews, participants preferred the dynamic reactions for greater repeatability and spontaneous interactions. At the same

time, some participants noted the importance of quick, efficient interactions when seeking story-related information from NPCs, suggesting a trade-off between engagement and efficiency. The Social Presence module also reflected the participants' increased connection with the NPCs, with significant increases across all three parameters. These reports were echoed in the interviews, which helped us understand how the participants perceived the NPCs as different people with their own personalities, and how this perception motivated them to return and engage frequently, hoping to uncover additional facets of these personalities. Players with previous experience with RPG games also drew parallels between the control version and games they have played before, with P7 commenting that the control version interactions *"felt like the ones in Pokemon games [popular RPG game series]"*.

6.2.1 Triggering Mechanism

The 4-step interaction we designed is a critical component of *Look2React*. We tried to merge real-life interaction cues with gameplay style iconography to ensure that the NPCs' transition from traditional scripted characters to more perceptual agents with distinct personalities was conveyed effectively. In the user study, participants with prior gaming experience quickly recognized the reference; for example, P9 remarked, *"... it reminds me of Assassin's Creed [a popular RPG game]"*. Interestingly, P13, who reported no previous experience with either VR or RPGs, also understood the mechanism immediately. This diverse acceptance suggests early success of the interaction design.

However, some players raised concerns (P16, P19). Specifically, P16 commented on its reliance on eye contact and how they personally felt awkward making prolonged eye contact, suggesting that we look into other body cues. Future work could incorporate the direction the user's body is facing to complement eye gaze detection.

6.3 Limitations and Future Work

6.3.1 User Study

For our user study, we created and compared two versions of the same game, with the only difference being the NPC interaction system (§ 4.1). For the control version, although we created an experience representative of the current RPG games, future work can compare against more elaborate systems that may identify the user's equipped items and respond accordingly. This would help highlight the importance of understanding dynamic actions compared to static identification of in-game items. Furthermore, future work can focus on evaluating the effectiveness of *Look2React* with additional customizations and scenarios, such as the player's clothes, age, hairstyles, and group dynamics in multiplayer. Future studies can also involve more participants and multiplayer capabilities to increase generalizability and to better understand group dynamics.

6.3.2 Alleviating Latency

The delay between 'Looking' and 'Reacting' was the most frequent topic of concern in the interviews. Part of this latency is an inherent consequence of working with VLM/LLM systems. In *Look2React*, we attempt to hide this latency with an intermediate 'thinking' pose. But we also introduce additional latency via the triggering and recording mechanisms. Future work could employ a "fast-slow" architecture [5, 20], using methods such as physics-based interactions or fast local models to continuously monitor for triggers and provide an immediate "instinctual" response, while a larger, slower VLM provides richer, more dynamic responses. Such an approach could also have other benefits: as two participants noted, the VLM model is unable to identify subtle details in gestures, like giving a flower vs. hitting with it, which may improve with additional physical context.

6.3.3 Scalability

A major consideration when designing NPC systems is the scale of current RPG games, which can easily have several different NPCs and factions. We designed *Look2React* to be highly scalable, with the `npc_type` (and `prev_response`) parameter (Table 1) being the only change between NPCs. Since the rest of the *Look2React* calls are the same (animations and prompt), they can be efficiently cached. Adding more animations is also straightforward, as only the names are shared with the VLM, and the animation files are common for each NPC, optimizing storage. Future work could create a more elaborate game with more NPCs to quantitatively measure the performance gains afforded by *Look2React* and the effectiveness of similar animations across different NPCs.

6.3.4 Technical Improvements

During the VR session, several participants used some form of speech while interacting with the NPCs, ranging from thinking aloud (*"Maybe I'll give this cake to the guard"*) to talking to them (*"You have a nice house!"*, to the girl). Speech interactions also came up often as a feature participants wanted during the interview (P1, P5, P7, P12). Future work should incorporate speech as an input modality for further enhanced freedom and immersion.

An intrinsic aspect of AI-based interaction is the stochasticity in the system. This randomness allows for unexpected and varied responses, which can make NPC interactions feel more natural and less scripted, but it also introduces unpredictability that may affect consistency and player trust, as some participants also pointed out (P1, P2, P21). To make interactions more consistent, a more detailed per-NPC interaction history can be shared with the VLM.

6.3.5 Flatscreen games

We built *Look2React* as a VR NPC system due to the increased control; however, it can also be used in flatscreen games. Players in flatscreen games frequently use traditional controls to communicate, such as jumping or crouching rapidly to say that they are friendly, and many games offer rich emotes for expression. However, these gestures are largely ignored by NPCs. *Look2React* presents an opportunity to bring more responsive, dynamic NPC behavior to flatscreen RPGs, enabling characters to recognize and react to player gestures in meaningful ways.

7 Conclusion

In this work, we proposed *Look2React*, a system capable of creating dynamic NPC responses by selecting an appropriate pose and generating accompanying text that feels natural and personalised to the NPCs, based on visual inputs from the virtual environment. We discuss the ideas we explored, explain why certain options were not selected, and detail our final architecture, laying down a crucial foundation for future work to build on. Additionally, we create an intuitive and reliable triggering mechanism that reduces accidental triggers, while giving users a natural way to initiate interactions. To evaluate *Look2React*, we created two versions of an RPG game, one with NPCs similar to popular commercial RPGs and one with *Look2React* NPCs. We invited 20 participants to play both versions of the game, and our findings show that *Look2React* NPCs increase presence, immersion, engagement, and satisfaction. With *Look2React*, NPCs can become more dynamic, expressive, and responsive to users, representing a meaningful advance towards creating more human-like, realistic in-game characters.

Acknowledgments

This work was supported in part by the Natural Science and Engineering Research Council of Canada (NSERC) under Discovery Grant RGPIN-2019-05624. We also acknowledge Michael Yin for his feedback and suggestions.

References

[1] K. Aggarwal and N. Nandhakumar. On the computation of motion from sequences of images-a review. 2

[2] E. F. Anderson. A npc behaviour definition system for use by programmers and designers. 2004. 1

[3] S. Azadi, A. Shah, T. Hayes, D. Parikh, and S. Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15039–15048, 2023. 4

[4] A. Baffa, P. Sampaio, B. Feijó, and M. Lana. Dealing with the emotions of non player characters. In *2017 16th Brazilian Symposium on Computer*

Games and Digital Entertainment (SBGames), pp. 76–87. ISSN: 2159-6662. doi: 10.1109/SBGames.2017.00017 2

[5] M. Bergamaschi Ganapini, M. Campbell, F. Fabiano, L. Horesh, J. Lenchner, A. Loreggia, N. Mattei, F. Rossi, B. Srivastava, and K. Venable. Fast, slow, and metacognitive thinking in ai. *npj Artificial Intelligence*, 1(1):27, 2025. 9

[6] J. A. Bopp, L. J. Müller, L. F. Aeschbach, K. Opwis, and E. D. Mekler. Exploring emotional attachment to game characters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, 12 pages, p. 313–324. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3311350.3347169 2

[7] F. Buttussi and L. Chittaro. Locomotion in place in virtual reality: A comparative evaluation of joystick, teleport, and leaning. *IEEE transactions on visualization and computer graphics*, 27(1):125–136, 2019. 5

[8] F. R. Christiansen, L. N. Hollensberg, N. B. Jensen, K. Julsgaard, K. N. Jespersen, and I. Nikolov. Exploring presence in interactions with LLM-driven NPCs: A comparative study of speech recognition and dialogue options. In *30th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–11. ACM. doi: 10.1145/3641825.3687716 1, 2

[9] L. M. Csepregi. The effect of context-aware LLM-based NPC conversations on player engagement in role-playing video games. 1

[10] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. pp. 1110–1118. 2

[11] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai. Revisiting skeleton-based action recognition. pp. 2969–2978. 2

[12] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis. Large language models and games: A survey and roadmap. *IEEE Transactions on Games*, 2024. 2

[13] A. Ghosh, A. Acharya, S. Saha, V. Jain, and A. Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. doi: 10.48550/arXiv.2404.07214 3

[14] R. Ghzouli, T. Berger, E. B. Johnsen, A. Wasowski, and S. Dragule. Behavior trees and state machines in robotics applications. 49(9):4243–4267. doi: 10.1109/TSE.2023.3269081 2

[15] M. D. Griffiths and F. Nuyens. An overview of structural characteristics in problematic video game playing. 4(3):272–283. doi: 10.1007/s40429-017-0162-y 2

[16] C. Guo, X. Zuo, S. Wang, and L. Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pp. 580–597. Springer, 2022. 4

[17] W. A. IJsselsteijn, Y. A. De Kort, and K. Poels. The game experience questionnaire. 2013. 6

[18] J. Jeong and T. Y. Lee. LIGS: Developing an LLM-infused game system for emergent narrative. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM. doi: 10.1145/3706599.3720212 1, 2

[19] G. Johansson. Visual perception of biological motion and a model for its analysis. 14(2):201–211. doi: 10.3758/BF03212378 2

[20] D. Kahneman. *Thinking, fast and slow*. macmillan, 2011. 9

[21] G. V. Kale and V. H. Patil. A study of vision based human motion recognition and analysis. 7(2):75–92. doi: 10.4018/IJACI.2016070104 2

[22] D. Kim, H. Yeo, and K. Park. Effects of an avatar control on vr embodiment. *Bioengineering*, 12(1):32, 2025. 1

[23] M. Korkiakoski, S. Sheikhi, J. Nyman, J. Saariniemi, K. Tapio, and P. Kostakos. An empirical evaluation of AI-powered non-player characters' perceived realism and performance in virtual reality environments. doi: 10.48550/arXiv.2507.10469 2

[24] K. Y. Kristen, M. Guzdial, and N. Sturtevant. The definition-context-purpose paradigm and other insights from industry professionals about the definition of a quest. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 17, pp. 107–114, 2021. 2

[25] V. Kumaran, J. Rowe, B. Mott, and J. Lester. SceneCraft: Automating interactive narrative scene generation in digital games with large language models. 19(1):86–96. Number: 1. doi: 10.1609/aiide.v19i1.27504 4

[26] P. Lankoski and S. Bjork. Character-driven game design: Characters, conflicts and gameplay. In *GDTW, Sixth International Conference in Game Design and Technology*, pp. 59–66, 2008. 2

[27] P.-H. Leveau and e. S. Camus. Embodiment, immersion, and enjoyment in virtual reality marketing experiences. *Psychology & Marketing*, 40(7):1329–1343, 2023. 1

[28] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. doi: 10.48550/ARXIV.2201.12086 3

[29] Z. Li, H. Zhang, C. Peng, and R. Peiris. Exploring large language model-driven agents for environment-aware spatial interactions and conversations in virtual reality role-play scenarios. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 1–11. ISSN: 2642-5254. doi: 10.1109/VR59515.2025.00025 1, 2

[30] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King. A survey on vision-language-action models for embodied AI. doi: 10.48550/arXiv.2405.14093 3

[31] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. 200(1140):269–294. Publisher: Royal Society. doi: 10.1098/rspb.1978.0020 2

[32] R. Michaux, C. Stassart, and A. Wagener. Does the need for control hinder sense of presence in virtual reality? *Psychologica Belgica*, 65(1):104, 2025. 1

[33] R. Mon-Williams, G. Li, R. Long, W. Du, and C. G. Lucas. Embodied large language models enable robots to complete complex tasks in unpredictable environments. 7(4):592–601. Publisher: Nature Publishing Group. doi: 10.1038/s42256-025-01005-x 1, 3

[34] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *J. Soc. Issues*, 56(1):81–103, Jan. 2000. doi: 10.1111/0022-4537.00153 2

[35] Y. Nong, H.-T. Zhang, and J.-Q. Sun. Natural language processing for immersive game interactions: Improving NLP models for more natural conversations with AI-driven NPCs in AR/VR games. doi: 10.36227/techrxiv.173747364.43683969/v1 1, 2

[36] K. L. Nowak. The influence of anthropomorphism and agency on social judgment in virtual environments. *J. Comput. Mediat. Commun.*, 9(2):00–00, June 2006. doi: 10.1111/j.1083-6101.2004.tb00284.x 2

[37] L. Odelbrink and E. Dolgikh. Methodology within npcs and game narrative, 2024. 2

[38] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pp. 1–22. Association for Computing Machinery. doi: 10.1145/3586183.3606763 2

[39] H. Qiu and B. Hou. Multi-grained clip focus for skeleton-based action recognition. 148. doi: 10.1016/j.patcog.2023.110188 2

[40] S. Rangelova, S. Flutura, T. Huber, D. Motus, and E. André. Exploration of physiological signals using different locomotion techniques in a vr adventure game. In *International Conference on Human-Computer Interaction*, pp. 601–616. Springer, 2019. 5

[41] B. Reeves and C. Nass. 1996. 2

[42] B. Ren, M. Liu, R. Ding, and H. Liu. A survey on 3d skeleton-based action recognition using learning method. 5:0100. Publisher: American Association for the Advancement of Science. doi: 10.34133/cbsystems.0100 2

[43] G. Rivera, K. Hullett, and J. Whitehead. Enemy NPC design patterns in shooter games. In *Proceedings of the First Workshop on Design Patterns in Games*, DPG '12, pp. 1–8. Association for Computing Machinery. doi: 10.1145/2427116.2427122 2

[44] K. Rogers, M. Aufheimer, M. Weber, and L. E. Nacke. Towards the visual design of non-player characters for narrative roles. 2

[45] M. S. Ryoo and J. K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. 82(1):1–24. doi: 10.1007/s11263-008-0181-1 2

[46] S. Salimpour, L. Fu, F. Keramat, L. Militano, G. Toffetti, H. Edelman, and J. P. Queralta. Towards embodied agentic AI: Review and classification of LLM- and VLM-driven robot autonomy and interaction. doi: 10.48550/arXiv.2508.05294 1, 3

[47] I. Sánchez-Berriel, F. Pérez-Nava, and L. Pérez-Rosario. Natural interaction in virtual heritage: Enhancing user experience with large language models. *Electronics*, 14(12):2478, 2025. 2

[48] K. Sasabuchi, N. Wake, A. Kanehira, J. Takamatsu, and K. Ikeuchi. Agreeing to interact in human-robot interaction using large language models and vision language models. doi: 10.48550/arXiv.2503.15491 3

[49] Y. A. Sekhavat. Behavior trees for computer games. 26(2):1730001. Publisher: World Scientific Publishing Co. doi: 10.1142/S0218213017300010 2

[50] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. 22(6):371–391. Publisher: SAGE Publications Ltd STM. doi: 10.1177/0278364903022006003 2

[51] L. Song. Llm-driven npcs: Cross-platform dialogue system for games and social platforms. *arXiv preprint arXiv:2504.13928*, 2025. 1, 2

[52] A. Steed, Y. Pan, F. Zisch, and W. Steptoe. The impact of a self-avatar on cognitive load in immersive virtual reality. In *2016 IEEE virtual reality (VR)*, pp. 67–76. IEEE, 2016. 2

[53] L. W. Thygesen. Replayability: a structural approach to players and computer games. *Master), Aalborg, Aalborg University*, 2014. 2

[54] J. van Stegeren and J. Myśliwiec. Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. In *Proceedings of the 16th International Conference on the Foundations of Digital Games*, FDG '21, pp. 1–8. Association for Computing Machinery. doi: 10.1145/3472538.3472595 2

[55] R. Vatsal, S. Mishra, R. Thareja, M. Chakrabarty, O. Sharma, and J. Shukla. An analysis of physiological and psychological responses in virtual reality and flat screen gaming. *IEEE Transactions on Affective Computing*, 15(3):1696–1710, 2024. 2

[56] H. Warpefelt. The non-player character : Exploring the believability of NPC presentation and behavior. Publisher: Department of Computer and Systems Sciences, Stockholm University. 2

[57] H. Warpefelt and B. Stra. Breaking immersion by creating social unbelievabilty. 2

[58] H. Warpefelt and H. Verhagen. TOWARDS AN UPDATED TYPOLOGY OF NON-PLAYER CHARACTER ROLES. 2

[59] M. Wittmann and B. Morschheuser. What do games teach us about designing effective human-AI cooperation? - a systematic literature review and thematic synthesis on design patterns of non-player characters. 2

[60] A. Yahyavi, J. Tremblay, C. Verbrugge, and B. Kemme. Towards the design of a human-like FPS NPC using pheromone maps. In *2013 IEEE International Games Innovation Conference (IGIC)*, pp. 275–282. ISSN: 2166-675X. doi: 10.1109/IGIC.2013.6659132 2

[61] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 379–385. ISSN: 1063-6919. doi: 10.1109/CVPR.1992.223161 2

[62] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pp. 7444–7452. AAAI Press. 2

[63] S. Yan, Y. Xiong, J. Wang, and D. Lin. Mmskeleton. `https://github.com/open-mmlab/mmskeleton`, 2019. 3

[64] C. Yin. NPCs in video games: a reflective resource for sports coaches and participant engagement. 6. Publisher: Frontiers. doi: 10.3389/fspor.2024.1403829 2

[65] M. Yin, E. Wang, C. Ng, and R. Xiao. Lies, deceit, and hallucinations: Player perception and expectations regarding trust and deception in games. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–15. Association for Computing Machinery. doi: 10.1145/3613904.3642253 2

[66] M. Yin and R. Xiao. Press a or wave: User expectations for NPC interactions and nonverbal behaviour in virtual reality. 8:1–25. doi: 10.1145/3677098 1

[67] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5):1005, 2019. 3

[68] J. Zhang, J. Huang, S. Jin, and S. Lu. Vision-language models for vision tasks: A survey. 46(8):5625–5644. doi: 10.1109/TPAMI.2024.3369699 1, 3

[69] W. Zhou, Z. Dou, Z. Cao, Z. Liao, J. Wang, W. Wang, Y. Liu, T. Komura, W. Wang, and L. Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pp. 18–38. Springer, 2024. 4