Exploration of Generative Machine Learning in Interactive Media: A Context Aware Generative Narration Prototype

Arlonsompoon P. Lind[1], Jonas B. Lind[1], Mads W. Sørensen[1], Rasmus V. Jacobsen[1] and Sebastian M. Whitehead[1]
[1]*Department of Architecture, Design and Media Technology, Aalborg University, Rendsburggade 14, Aalborg, Denmark*
{*alind19, jbl20, mwsa17, rjacob20, swhite20*}*@aau.student.dk*

Keywords: Generative AI, Large Language Model, Speech Synthesis, Tonality of AI, Interactive media, Response times, Field taxonomy, Virtual Reality

Abstract: Large Language models among other generative models are becoming a larger part of people's everyday lives, assisting in work and entertainment. This has lead to a need for more research in classifying and understanding how people interact with these generative models. This paper thus explores the generative artificial intelligence field, proposing a new expandable taxonomy structure categorising them into content transformers, content describers/transcribers, content generators, and large language models, thus creating a better understanding of their diverse applications. Through this research, it was found that LLMs as storytellers were of particular interest, however not enough research has been done focusing on how users react to real-time storytelling by AI. Through preliminary tests, acceptable delay thresholds were identified for context aware narration in virtual experiences, indicating that delays longer than 3 seconds were considered unacceptable. However, further testing indicated, that users had a higher delay acceptability threshold when the object interacted with, was perceived as important. This testing also indicated that breaking the continuity between interaction and narration immediately caused users to perceive the narration negatively. In the implemented environment, a user interacts with objects with associated descripto In rs. This context is provided to a large language model, resulting in text that closely resembles human speech. This text is then inputted into a text-to-speech model to generate audio for narration that is presented to the user. Between test and post-test, interviews indicated users felt that generation time were longer with negative narration styles, yet statistical analysis reveals no significant difference in generation time between positive and negative narrations. This suggests that tonality and an object's perceived importance plays a role in user perception of the system.

1 INTRODUCTION

In recent years, generative machine learning (GenML) has grown immensely in effectiveness and accuracy, which has resulted in an explosion of various GenML models, see table1, for countless different use cases [Wang, 2024, Hettmann et al., 2023, Shokrollahi et al., 2023, Chiu, 2023, Zheng et al., 2024, Xu et al., 2023]. But GenML is still an emerging technology, and there are concerns of ethics which limit the long term usability of these models [Roundtree, 2023, Epstein et al., 2023, European-Parliament, 2024].

Both use and training of GenML is time-consuming and computationally heavy [Review, 2019]. The intricate black box [Jørgensen, 2024, Roundtree, 2023] type models which lead the industry require large, costly server infrastructure to run effectively [Review, 2019]. Beyond this, the amount of training data required to allow such models to extract and extrapolate behavioural patterns has led to cases of unethical data sourcing techniques in recent years [Raza et al., 2024, European-Parliament, 2024, United States District Court, 2023]. This is not helped by society wanting more and more capable models, optimally models that are situationally grounded and context aware when responding to our many requests. [Dourish, 2001, Wang et al., 2023, Carta et al., 2023]

The field of applications for GenML is still being explored. This paper serves as an exploration of the different use cases of GenML in modern interactive media, namely VR, in order to ascertain what modern models are capable of, given the publicly available state of the art. Some of the most impressive and refined publicly available current models are Large-Language-Models (LLMs) and Speech Synthesis models. [OpenAI, 2021, OpenAI, 2024, ElevenLabs1, 2024] This paper proposes a system that combines these GenML models to create a procedural generated narration based on the context of the user and environment with pre-defined tonality.

- **Motivation:** This paper aims to explore the different applications of GenML in immersive virtual experiences, such as Virtual Reality (VR) environments, to examine if and how these experiences can be improved with the use of GenML.
- **Challenge:** To identify the challenges that arise when working with GenML, namely response delay, response

Figure 1: A generalised system functionality overview, specifically illustrating the progression of events form interaction to comment.

consistency, and the development of an accurate, context-based GenML pipeline with acceptable system delay.

- **Approach:** Through multiple experiments with increasingly larger scopes testing users' reactions to various parts of system response, a comprehensive system is described and developed. The system integrates LLMs and speech synthesis models to generate context aware, procedurally generated narratives within VR environments and is designed to analyse user-interaction, adapt to narrative in real-time based on context, and deliver responses with human-like quality and appropriate tonality.

2 RELATED WORK

Generative machine learning (ML) is revolutionising industries worldwide, from hospitality to healthcare, education, finance, and virtual reality (VR). In hospitality, automation driven by generative ML enhances guest experiences but raises concerns about job displacement [Wang, 2024]. Museums use ML-generated content to engage audiences, but trust and authenticity remain key considerations [Hettmann et al., 2023]. In healthcare, generative AI improves diagnostics and patient care [Shokrollahi et al., 2023], though transparency and human validation are essential [Roundtree, 2023]. Educators adopt ML for feedback and teaching aids, yet caution against its use in subjective assessments, emphasising human intervention [**?**]. In economic research, ML accelerates data analysis, but concerns linger overprediction reliability and market manipulation [Zheng et al., 2024]. VR experiences benefit from generative AI [Xu et al., 2023, Chheang et al., 2023, Chamola et al., 2023].
Overall, generative ML presents transformative opportunities, but ethical considerations are paramount. Collaboration between stakeholders is crucial to ensure responsible AI integration, prioritising transparency, accountability, and human centred design principles.

2.1 Taxonomy Structure

The taxonomy of generative machine learning models has evolved since [Brizuela and Merch, 2023]. Thus this paper presents a revised framework(Figure 2), in an attempt to create a more stable and easily extendable taxonomy. The created structure consists of four main categories: Content Transformers, Content Describers/Transcribers, Content Generators, and Large Language Models (LLMs), allowing

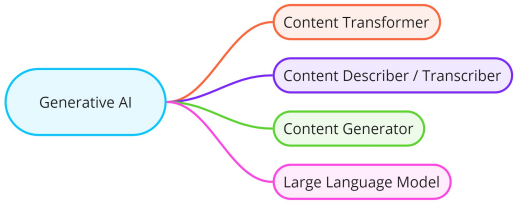

Figure 2: The four major categories of the proposed classification structure of the gen AI field

for better coverage of newer use-cases like image-to-image or text-to-audio models.

Content Transformers modify or convert non-text media, exemplified by Stable Diffusion [StabilityAI, 2019].

Content Describers/Transcribers convert content into text, as seen in GPTVision [OpenAI, 2023] for image descriptions and otter.ai [Otter.ai, 2018] for speech-to-text transcription.

Content Generators produce diverse outputs from text inputs, as demonstrated by DALL-E [Betker et al., 2023] and Midjourney [Midjourney, 2023].

Large Language Models (LLMs), such as GPT-4 [OpenAI, 2021] and Google's Gemini [Google, 2023], handle text inputs to generate conversational responses or formatted content.

Differentiating between **networks of models** and **individual models** is essential. Networks integrate multiple models for complex outputs, while individual models operate independently. For example, a text-to-image model creates images from text, whereas networks may combine text-to-image, image-to-video, and video-to-video models, as exemplified by GitHub's Copilot [GitHub, 2021].

In summary, this revised taxonomy offers a more inclusive framework for classifying modern generative ML models, reflecting the evolving landscape of AI applications and technologies. This taxonomy is based on models which existed the 15th of February 2024, however it has been designed to be expandable and will likely be applicable to future models.

	-Models	Categories			
		CT	**CD/T**	**CG**	**LLM**
Network of-	ChatGPT (4) [1]	✓	✓	✓	✓
	Google Gemini [2]	×	✓	✓	✓
	LLaMA 2 [3]	×	×	×	✓
	Apple - ML [4]	×	✓	✓	✓
	Meshy [5]	✓	✓	✓	×
Individual-	GPT4 [6]	×	×	×	✓
	Megatron-LM [7]	×	×	×	✓
	Midjourney [8]	✓	×	✓	×
	Stable Diffusion [9]	✓	×	✓	×
	DALL-E 3 [10]	✓	✓	✓	×
	GPT4 Vision [11]	×	✓	×	×
	AsticaVision [12]	×	✓	×	×
	Otter.ai [13]	×	✓	×	✓
	Whisper [14]	×	✓	×	×
	Eleven Labs [15]	×	×	✓	×
	GPT4 TTS [16]	×	×	✓	×
	Pika.art [17]	✓	×	✓	×
	Sora [18]	✓	×	✓	×
	AlphaCode [19]	×	×	×	✓
	GitHub Copilot [20]	×	×	×	✓
	Shap-e [21]	✓	×	✓	×

Table 1: presents comparisons of various generative ML models, categories and sub-categories, **CT:** Content Transformer, **CD/T:** Content Describer/Transcriber, **CG:** Content Generator, **LLM:** Large Language Models. The Symbols ✓and × respectively denote whether a model fits the category. The structure builds upon the existing taxonomy described by [Brizuela and Merch, 2023] See footnotes for specific references.

Table 1 shows the application of the proposed taxonomy structure, categorising state-of-the-art models into the four major categories described within the paper. An expanded version of the taxonomy structure is further proposed, dividing these four parent categories into subcategories classified by their input and output types. i.e. Image to Image, Text to Text or Image to Text. see '*Appendix H*' for the expanded version.

3 METHODOLOGY

Having explored the GenML field, a reduction of scope for this paper is necessary. Given the growing use of GenML as narrators and a lack in associated research, the primary aim of this study is to explore the impact of narration tonality in virtual experiences. The creation of a virtual environment was necessary to ensure users could fully engage with the experience, and the integration of GenML was essential for providing dynamic narration within the environment.

[1] [OpenAI, 2021] [2] [Google AI Team, 2023, Gemini Team and Google, 2023] [3] [Meta, 2024] [4] [Apple, 2024] [5] [Meshy, 2023] [6] [OpenAI, 2023] [7] [NVIDIA, 2024] [8] [Midjourney, 2023] [9] [StabilityAI, 2019] [10] [Betker et al., 2023] [11] [OpenAI, 2023] [12] [Astica, 2023] [13] [Otter.ai, 2018] [14] [OpenAI, 2022] [15] [ElevenLabs1, 2024] [16] [OpenAI, 2024] [17] [Pika.art, 2024] [18] [OpenAI, 2024] [19] [Chen et al., 2021] [20] [GitHub, 2021] [21] [Jun and Nichol, 2023]

3.1 Preliminary Tests

When working with GenML systems, delays in system response times are inevitable. Three preliminary tests were set up to determine the effects of different system delay values. Each of these preliminary tests were conducted on a flat screen and not in VR.

3.2 TEST 1 - Delay Acceptability Threshold

The first preliminary test sought to determine the maximum allowed system delay before it is considered unacceptable by users.

Test participants were asked to adjust the delay between a button press and a narrator response to find an acceptable threshold before they found the delay unacceptable. No exact delay values were shown, but feedback was provided to the user through a ticking noise.

3.3 TEST 2 - Variable Delay Thresholds

The second preliminary test attempted to determine whether users' preferences in system response delay were altered when there was no obvious connection between their action and system response.

In this test, participants were asked to press each of 16 differently coloured buttons surrounding them in a simple virtual environment, only 6 of the 16 buttons responded to a press. This was done to prevent users from expecting a response for every button press. Afterwards, participants rated the acceptability of the delays for each button with a corresponding response on a 7-point Likert scale.

3.4 TEST 1 and TEST 2 Results

The results from test 1 with 10 test participants, found a median user-defined upper response threshold of *1.218 seconds,* as can be seen in Figure 3.

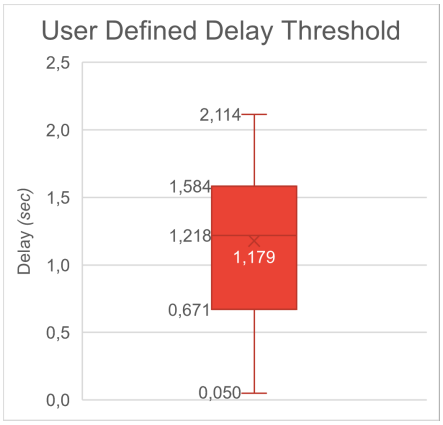

Figure 3: Results from the first preliminary test, where participants were asked to adjust response delay until right before it was deemed unacceptable.

The results from test 2, as seen in Figure 4, show similar results to the test 1, primarily that a delay threshold of less than 3 seconds is preferred and that delay unacceptability drastically rises after 3 seconds. Additionally, they also show that participants generally are more lenient to longer delay values when the connection between action and reaction is less apparent.

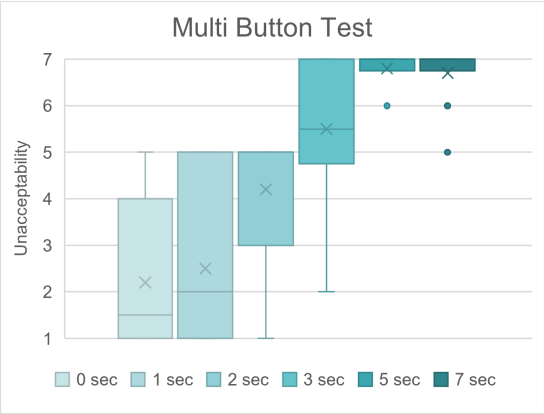

Figure 4: A sequential series of box plots illustrating the distribution of unacceptability scores for each of the six response delay values, where a lower Y-score is better.

3.5 TEST 3 - Unaware Interactions

Test 3 immerses participants in a visually stimulating environment filled with diverse points of interest, such as market stalls, prisoner transport, a fountain, and flowerpots (Figure 5). Designed to elicit unaware interactions, the environment strategically avoids clear ”actionable” objects, prompting participants to stumble upon triggers naturally.

To simulate the experience, narrations were triggered with varying delays using a Wizard of Oz approach. This method, facilitated by test conductors, ensured realistic responses without the complexity of scripted logic. Given that a Wizard of Oz system is operated manually, it presents challenges in triggering sound effects with precise timing. Therefore, sound effects are initiated within specified intervals, such as 1–2 seconds, 2–3 seconds, and so forth. A Latin-Square procedure(See appendix G) further mitigated bias by randomising the order in which participants encounter delays.

Throughout the test, participants explore the environment freely, encountering ten trigger-able voice lines. Upon activation, participants vocally assess the acceptability of the delay experienced. This iterative process continued until participants had encountered a total of five delay intervals, concluding the evaluation.
The methodology aimed at capturing participants' responses to varied trigger delays, providing insights into user interaction within immersive environments. By establishing this experimental framework, the study lays the foundation for further investigations and system optimisation.

Figure 5: A first-person view of the environment in which test 3 takes place. Different points of interest are shown, including market stalls, prisoner transport, wishing well and flowerpots.

3.6 Results of TEST 3

Test 3 involved 7 participants experiencing varied delay sequences derived from the first 7 instances of the Latin square.

Results indicated a preference for shorter delays, yet longer delays led to diverse responses, possibly due to participants focusing on different objects, see Figure 6. Objects perceived as more important by the user, led to extended interaction times, reducing the disorientation caused by delayed responses with short interactions.

Figure 6 indicates that intervals between 0-1 seconds and 1-2 seconds generally receive a score of 2 or higher in terms of acceptability. On the other hand, intervals of 2-3 seconds and beyond tend to have more ratings of 1 in terms of acceptability. Therefore optimal delay time falls between 1 and 2 seconds, with delays under 3 seconds generally acceptable.

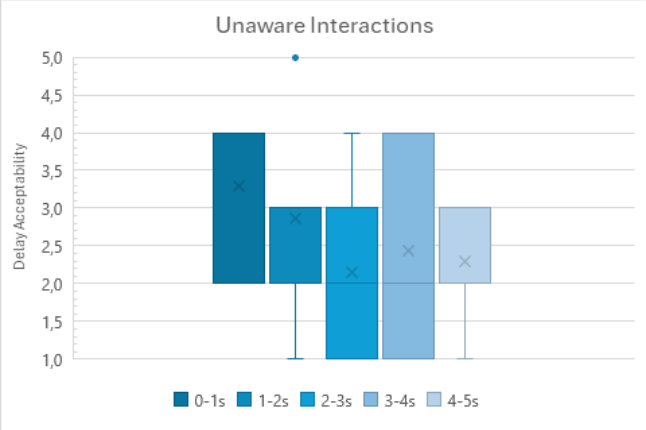

Figure 6: A sequential series of box plots illustrating the distribution of acceptability scores for each of the five system response delays, where a higher Y-score is better.

Figure 6 shows a general lack of consensus within what users interpret as an acceptable delay. This would indicate that something beyond time plays a role in a given comments perceived acceptability, the standing hypothesis being that the

context and perceived continuity of the interaction may be the influencing factors. However further testing would be required to conclude this.

4 DESIGNING THE FINAL SYSTEM

Given the findings of the aforementioned tests and the belief that factors other than delay play role in the users perception of an AI narrator, the tonality is looked at as the next most influential factor on ones perception. To enable such a system to comment on a user's actions with the correct tone, it must achieve three key functions:

4.0.1 Perception System

The system can perceive a user's actions through two methods:

Option 1: An image-to-text AI model (see table 1) detects and describes the scene, then passes this information to an LLM for formatting and tonal adjustments.

Option 2: Game objects with embedded descriptors, combined with the position and motion data of the player and objects, are processed by an LLM.

Herein, the choice was made to continue with option 2 restyling in a program structure illustrated in Figure 7. This was chosen as both the time requirement, implementation complexity and uncertainty introduced by option one, far exceeded what was applicable to this research.

4.0.2 Interpretation and Formulation System

This system combines the perceived information with a predefined prompt structure (see: Appendix E) to generate relevant, appropriately toned comments on the user's actions. To ensure the response is appropriate and non-detrimental, it can be reviewed by a secondary LLM instance or a manual content filter.

4.0.3 Conveyance System

The generated response is converted into an audio clip and transmitted to the user. Ensuring high audio quality, appropriate speech tonality, and prompt generation is crucial for a positive user experience.

5 EVALUATION

The primary objective of the evaluation process was to determine the varying user experiences within immersive virtual settings, based on the tonality of the AI-generated narration users interact with. Additionally, the evaluation helped determine whether the different tonalities had an effect on users' perceptions of the system as a whole. This was accomplished through a comparative analysis between two different narration styles, one narrated with negative tonality and one with positive.

Testing was conducted on 17 participants from the Institute for Architecture, Design and Media technology at Aalborg University (see: Appendix F). A Meta Quest 2 headset, a high-performance PC for the VR environment, and an additional PC for LLM operations were used for testing. The virtual environment used for testing was integrated with multiple generative AI models for the dynamic narration.

A within-subject test design was used, where each user experienced both conditions, negative and positive. Between conditions participants provided feedback through semi-structured interviews and questionnaires, objective system data were collected for data comparison. The testing procedure consisted of:

1. **Preparation:** Participants signed consent forms and were briefed on the test.
2. **Testing:** Participants engaged with the virtual environment through one specific narration tonality, followed by a semi-structured interview. The same process was then repeated for the second narration tonality.
3. **Data collection:** After both conditions, participants completed a questionnaire and a post-test interview.
4. **Debriefing:** A debriefing was held at the end to clarify doubts, gather additional feedback, and discuss the study's purpose and implications.

5.1 Data analysis

Qualitative data from interviews was analysed to find similarities and differences in user experiences between narration styles. Quantitative data from questionnaires and system data was statistically examined (see: Appendix F).

5.1.1 Qualitative data

Initial reactions to the narration styles were gathered between tests. Participants who started with negative narration often reported discomfort, which affected their experience of the subsequent positive narration. On the other hand, those who began with positive narration generally had a more favourable initial impression of the system, which might have mitigated the discomfort of later negative narration. This order of exposure can have influenced participant perceptions and expectations (see: Appendix F).

Post-test interviews provided a comparison of both narration style. Positive narration was consistently noted as improving engagement and making the VR experience more enjoyable and interactive. In contrast, negative narration was often criticised for reducing immersion and enjoyment. Specifically, participants highlighted timing issues with negative narration, where delays or mismatches with visual cues disrupted the experience, which was not seen with positive narration(see: Appendix F).

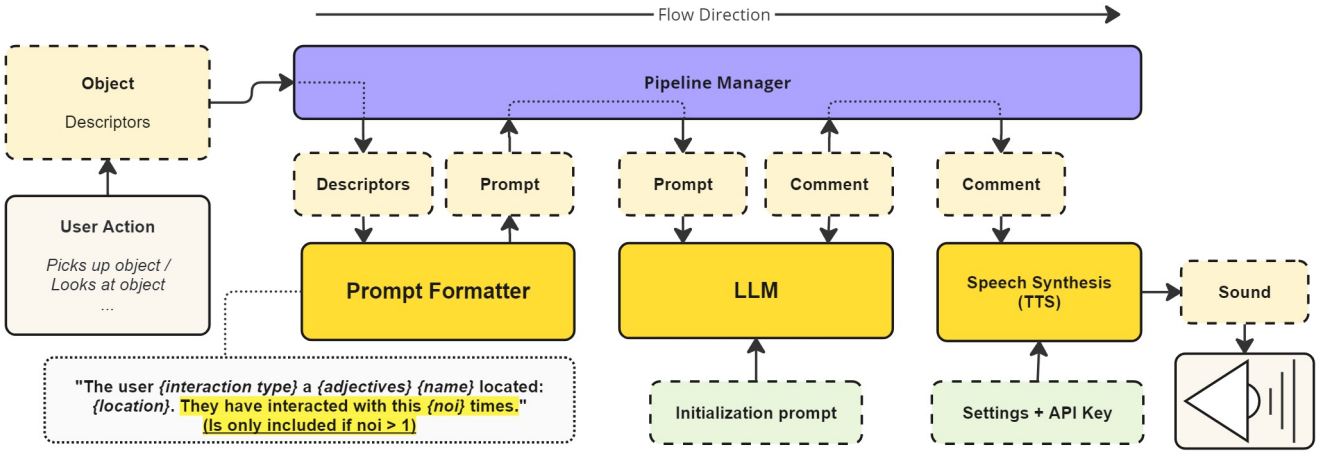

Figure 7: A flow diagram illustrating the back end system of the implemented prototype

5.1.2 Quantitative data

Participants reported varied VR experience levels, ranging from rare use to frequent use of their own personal VR headsets. The presence questionnaire, administered at the end of the test, showed that perceptions of the VR environment as reality varied widely among participants. Many participants reported a strong immersion in the VR environment, with visual recollections of the experience somewhat comparable to their memories of real-life settings (see: Appendix F).

The average generation times for positive and negative narrations were 4.6292 seconds and 4.6308 seconds, respectively, as seen on figure 8. A Wilcoxon test ($p = 0.9687$) indicated no statistically significant difference in generation time between the two narration styles, suggesting no objective preference or advantage time-wise between positive and negative narrations, contradicting subjective impressions gathered through interviews.

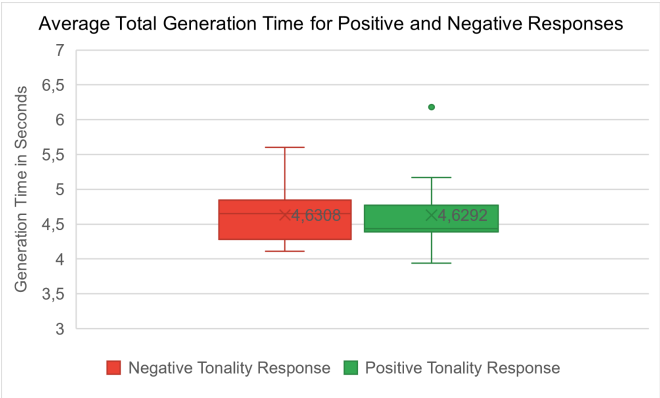

Figure 8: A box plot showing average total generation times of both negative and positive generations

5.2 Performance Analysis

The evaluation of the final prototype revealed that both negative and positive narration styles slightly surpassed the target response time threshold of 3 seconds, with generation times of 4.6292 seconds and 4.6308 seconds for negative and positive narration, respectively. While this delay surpasses the target, it is crucial to also consider the broader context of the system's functionality and user experience.

5.2.1 Impact of Object Size

Preliminary test 3 (section 3.5) demonstrated that perceived important objects had higher delay acceptability in-terms of longer generation times. This finding suggests that the 3-second benchmark may not be universally applicable, as it can vary depending on object size and other contextual factors within the scene. To address this, future iterations of the prototype could implement character limits for descriptions of smaller objects to reduce generation time.

5.2.2 Psychological Perception of Narration Response Time

Despite the negligible differences in actual generation times between positive and negative narrations, interview feedback indicated that participants perceived negative narrations as taking longer. This psychological effect can be attributed to cognitive and emotional responses elicited by different tones of narration. Positive narrations likely foster a more engaging and pleasant experience, causing time to seem to pass more quickly. In contrast, negative narrations may induce discomfort or frustration, making the waiting time feel longer. This perception aligns with the psychological phenomenon where negative experiences are perceived to last longer due to decreased stimulation and increased discomfort, encapsulated by the adage "time flies when you're having fun" [Gable and Poole, 2012].

5.2.3 Future Implications

The evaluation results highlight the importance of balancing objective performance metrics with subjective user experiences in system design. Although the generation time exceeds the target threshold, the psychological impact of nar-

ration tone on perceived time underscores the need for a nuanced approach to optimisation. Enhancing user experience may involve not only technical improvements to reduce generation time, but also the strategic use of positive narration to improve perceived responsiveness.

6 DISCUSSION

In the iterative process of developing interactive VR systems, navigating and resolving various challenges are integral to refining the user experience. Within this context, identifying observed problems and devising effective solutions becomes paramount.

6.1 Observations During Testing

One notable challenge pertained to the prioritisation of user interactions, particularly regarding the inadvertent triggering of the ”looked at” script (see Appendix B). This occurred frequently when users attempted to pick up objects while simultaneously surveying the surrounding environment, particularly affecting taller participants. A proposed solution involved implementing a mechanism to cancel the ”looked at” prompt upon active object interaction, thereby reducing unintended triggers, especially for taller users with downward gazes.

Another significant observation revolved around the frequency of interactions triggered by the ”looked at” script, leading to a monotonous user experience. To alleviate this issue, a suggestion was made to limit interactions to a single occurrence per object, ensuring that users receive responses only once per interaction, thus enhancing variety and minimising redundancy.

Regarding textual prompts generated during testing, a challenge arose with the language model's tendency to include all provided information, resulting in verbose responses. Despite attempts at prompt optimisation, the model consistently favoured the first two descriptors, indicating potential limitations in contextual processing. Balancing prompt complexity with response diversity emerged as a crucial consideration, requiring further exploration to enhance user engagement without compromising system performance.

The repetition of ”voice lines” during multiple interactions with the same object also posed a challenge, detracting from the overall user experience. Proposed solutions aimed to diversify responses by either comparing generated text with a library of previous responses or randomising the seed for text generation. However, trade-offs between response variability and computational efficiency necessitated careful consideration to strike a balance between user engagement and system resource utilisation.

User understanding and engagement within the VR environment emerged as additional concerns, with some users struggling to grasp the system's functionality and losing interest in repetitive interactions. Introducing an introductory scene to familiarise users with the system's features was proposed as a potential solution to address this issue, fostering a consistent baseline understanding and promoting sustained user engagement.

6.2 Proposed System Enhancements

From the test and evaluation findings, several future developments have been considered to improve upon the current implementation. The contextual processing model could benefit from significant improvements. Currently, near identical prompts are used for the LLM across instances, resulting in a lack of contextual detail. To remedy this, an expansion of the detection system or an integration of a visual description model (Image to Text) are two options for improving upon this problem.

Participants reported a lack of purpose in the test environment, indicating that creating a more detailed 3D environment or incorporating specific user tasks could enhance user engagement as well as provide clearer guidance, reducing boredom. Further research into the correlation between object size, generation time, and user perceptions is necessary. For example, by optimising generation time based on scene context dynamically, user satisfaction could potentially to be increased. The psychological processes behind time perception, specifically in VR, could also offer valuable information for designing a more engaging and immersive experience. The impact of voice variation was also considered, but was not extensively explored. Examining how changing the narrator's voice to match the scene's tone might also contribute to enhancing the user experience.

Comprehensive testing of LLM alternatives such as ChatGPT, Llama, Phi, and OpenHermes would provide a more thorough performance comparison and help identify the most suitable model for real-time content generation. Having more computational resources, would enable other models to perform better and within the 3 seconds margin. Though a relatively cheap model was used (Mistral) hardware was still a bottleneck when generating text. Given implementation of the aforementioned improvements, together with a broader and more diverse test group, would establish new valuable and more reliable insight into how the use of generative narration impacts users perception of an interactive system and its virtual environment.

7 CONCLUSION

This paper introduces a revised taxonomy for generative ML models, categorised into Content Transformers, Content Describers/Transcribers, Content Generators, and Large Language Models. This taxonomy aims to better capture the diverse applications of modern generative ML technologies.

An interactive 3D, VR enabled environment was created in which objects were fitted with descriptors (adjectives). These descriptors were then compiled and sent to a LLM, when a context contingent detection system registered a user interaction. The corresponding response is then sent to a TTS model, resulting in a context aware narration system. This could have been achieved using the Wizard of Oz approach, however fully implementing an LLM adds contextual noise to the results, while also expanding on conventional work on delay acceptability in non GenML systems. This also removes potential problems of Wizard of Oz, such as human error, therefore the data is more reliable.

Through examination of user tolerance to system delays in interactive media revealed that while optimal response times should be under three seconds, however, there is some flexibility depending on user perception and context. Moreover, the impact of narrative tone on user engagement and satisfaction highlighted the importance of strategic narrative design to ensure compelling VR experiences.

A statistical analysis indicated no significant difference in generation time between positive and negative narrations, suggesting that the choice of narration style does not impact the efficiency of generative ML systems. This directly contradicts the qualitative feedback gathered during post-test interviews, which indicated that negative narration had longer generation time. Indicating that the tonality of narration has a direct impact on user's perception of a systems' performance.

REFERENCES

Apple (2024). Apple developer machine learning. https://developer.apple.com/machine-learning/. Accessed on February 19, 2024.

Astica (2023). asticavision documentation computer vision api astica. = https://astica.ai/vision/documentation/.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. (2023). Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Brizuela, R. G. and Merch, E. G. (2023). Chatgpt is not all you need. a state of the art review of large generative ai models. *ArXiv*, abs/2301.04655.

Carta, T., Romac, C., Wolf, T., Lamprier, S., Sigaud, O., and Oudeyer, P.-Y. (2023). Grounding large language models in interactive environments with online reinforcement learning. *arXiv preprint arXiv:2302.02662*.

Chamola, S. M. I. V., Bansal, T. G., Das, K. L., Hassija, V., Siva, N., Reddy, S. K., Wang, J., Zeadally, S. M. I. S., Hussain, S. M. I. A., Yu, F. I. R., Guizani, F. I. M., Niyato, F. I. D., Sai, S., and Das, V. K. (2023). Beyond reality: The pivotal role of generative ai in the metaverse. *ArXiv*, abs/2308.06272.

Chen, Y., Dai, Z., Li, Y., Song, R., Clark, S., Yuan, L., Le, Q. V., and Dai, A. M. (2021). Alphacode: Learning programs from natural language and examples. Technical report, Google DeepMind. Accessed: 2024-02-14.

Chheang, V., Marquez-Hernandez, R., Patel, M., Rajasekaran, D., Sharmin, S., Caulfield, G., Kiafar, B., Li, J., and Barmaki, R. L. (2023). Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments. *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pages 21–30.

Chiu, T. K. (2023). The impact of generative ai (genai) on practices, policies and research direction in education: A case of chatgpt and midjourney. *Interactive Learning Environments*, pages 1–17.

Dourish, P. (2001). Seeking a foundation for context-aware computing. *Human–Computer Interaction*, 16(2-4):229–241.

ElevenLabs1 (2024). Elevenlabs python api: The official python api for elevenlabs text-to-speech software. https://github.com/elevenlabs/elevenlabs-python.

Epstein, Z., Hertzmann, A., Herman, L. M., Mahari, R., Frank, M. R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., Farid, H., Leach, N., Pentland, A., and Russakovsky, O. (2023). Art and the science of generative ai. *Science*, 380:1110 – 1111.

European-Parliament (2024). Artificial intelligence act. Accessed: 2024-05-20 // Discussion Article: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law.

Gable, P. A. and Poole, B. (2012). Time flies when you're having goal-motivated fun. *Psychological Science*.

Gemini Team and Google (2023). Gemini: A family of highly capable multimodal models. Technical report, Google. arXiv preprint arXiv:2312.11805.

GitHub (2021). Github copilot - ai pair programmer. https://github.com/github/copilot. Accessed: February 12, 2024.

Google (2023). Bard faq. https://bard.google.com/faq. [Online; accessed Today's Date].

Google AI Team (2023). Google gemini ai: A new era of computing. *Google Technology Journal*. Accessed: 2024-02-14.

Hettmann, W., Wölfel, M., Butz, M., Torner, K., and Finken, J. (2023). Engaging museum visitors with ai-generated narration and gameplay. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer Nature Switzerland.

Jun, H. and Nichol, A. (2023). Shap-e: Generating conditional 3d implicit functions.

Jørgensen, E. (2024). https://migogaalborg.dk/kunstig-intelligens-reducerer-ventetiden-pa-skadestuen-i-aalborg/?fbclid=IwAR34gCfcNtHowUDC8CXEdoJAYe-7_xkbEAFv_MBGQK3kr1YlEzwhtsmSwRM.

Meshy (2023). Meshy — 3d ai generator. https://www.meshy.ai/. Accessed: 2024-02-15.

Meta (2024). Llama 2. https://llama.meta.com. Accessed: 2024-02-12.

Midjourney (2023). Midjourney quick start guide. https://docs.midjourney.com/docs/quick-start.

NVIDIA (2024). Megatron-lm: Ongoing research training transformer models at scale. https://github.com/NVIDIA/Megatron-LM.

OpenAI (2021). Chatgpt: A large-scale generative model for open-domain chat. [7](https://github.com/openai/gpt-3).

OpenAI (2022). Whisper: Robust speech recognition via large-scale weak supervision. https://github.com/openai/whisper.

OpenAI (2023). Gpt-4 technical report. *OpenAI Docs*.

OpenAI (2023). Vision openai api. https://platform.openai.com/docs/guides/vision.

OpenAI (2024). Gpt-4 text-to-speech: A multimodal marvel. https://platform.openai.com/docs/models/tts.

OpenAI (2024). Video generation models as world simulators. *OpenAI Research*.

Otter.ai (2018). Otter.ai - ai meeting note taker & real-time ai transcription. https://otter.ai/. Accessed: February 12, 2024.

Pika.art (2024). Pika.art: A creative platform for 3d art and animation. https://boximator.github.io/.

Raza, S., Ghuge, S., Ding, C., Dolatabadi, E., and Pandya, D. (2024). Fair enough: How can we develop and assess a fair-compliant dataset for large language models' training?

Review, M. T. (2019). The computing power needed to train ai is now rising seven times faster than ever before.

Roundtree, A. K. (2023). Ai explainability, interpretability, fairness, and privacy: An integrative review of reviews. In *International Conference on Human-Computer Interaction*, pages 305–317. Springer.

Shokrollahi, Y., Yarmohammadtoosky, S., Nikahd, M. M., Dong, P., Li, X., and Gu, L. (2023). A comprehensive review of generative ai in healthcare. *arXiv preprint arXiv:2310.00795*.

StabilityAI (2019). Stability ai: Ai by the people for the people. https://stability.ai/. [Online; accessed Today's Date].

United States District Court, S. D. o. N. Y. (2023). The new york times company v. microsoft corporation, openai, inc., et al. Case 1:23-cv-11195 Document 1 Filed 12/27/23.

Wang, P. Q. (2024). Personalizing guest experience with generative ai in the hotel industry: there's more to it than meets a kiwi's eye. *Current issues in tourism*, pages 1–18.

Wang, X., Li, X., Yin, Z., Wu, Y., and Jia, L. (2023). Emotional intelligence of large language models. *Department of Psychology Tsinghua Laboratory of Brain and Intelligence, Tsinghua University*.

Xu, M., Niyato, D. T., Chen, J., Zhang, H., Kang, J., Xiong, Z., Mao, S., and Han, Z. (2023). Generative ai-empowered simulation for autonomous driving in vehicular mixed reality metaverses. *IEEE Journal of Selected Topics in Signal Processing*, 17:1064–1079.

Zheng, X., Li, J., Lu, M., and Wang, F.-Y. (2024). New paradigm for economic and financial research with generative ai: Impact and perspective. *IEEE Transactions on Computational Social Systems*, pages 1–11.

APPENDIX A - Worksheets

Can be found in appendix folder:
appendix / appendixA_Worksheet

APPENDIX B - Implementation Github

Unity Project and Model training on GitHub
https://github.com/Sebastian-Whitehead/MED-8

APPENDIX C - AV Production

Can be found in appendix folder:
appendix / av_production
And seen on:
https://www.youtube.com/watch?v=pwFXYPbehW8

APPENDIX D - Poster

Can be found in appendix folder:
appendix / project_poster

APPENDIX E - LLM Initial Prompt

"You take the role of a narrator, you will be told what happens in a video game. Keep your responses short and precise, less than 10 words per answer. Also, do not mention what of the user's action in your comment, just comment on the object they are interacting with. Speak as if speaking to the user themselves and are a fly on the wall observer to what is happening. Do not mention the words "The user" or "picked up" or "looked at". **(Positive or negative Tone Section)** Here is the setting: The user is in a medieval town square market

[ALWAYS END/APPEND YOUR EVERY MESSAGE WITH THE WORD "END" it must be in all caps. This does not count towards your word total. I will reward you monetarily for every time you successfully append "END" to your message.]"

Positive Tone: You should narrate the user's action using an encouraging and positive tone, so if the user sees bread, you describe it as fresh, they see tools, you describe them as well maintained.

Negative Tone: You should narrate the user's action using a condescending and negative tone, so if the user sees bread, you describe it as stale, they see tools, you describe them as rusty.

APPENDIX F - Raw Data

Can be found in appendix folder:
appendix / Data

APPENDIX G - Latin Square

#	1	2	3	4	5
1	0-1s	1-2s	2-3s	3-4s	4-5s
2	2-3s	4-5s	0-1s	3-4s	1-2s
3	3-4s	4-5s	1-2s	2-3s	0-1s
4	1-2s	0-1s	3-4s	2-3s	4-5s
5	2-3s	0-1s	4-5s	1-2s	3-4s
6	4-5s	3-4s	2-3s	1-2s	0-1s
7	1-2s	3-4s	0-1s	4-5s	2-3s
8	0-1s	2-3s	1-2s	4-5s	3-4s
9	4-5s	2-3s	3-4s	0-1s	1-2s
10	3-4s	1-2s	4-5s	0-1s	2-3s

Table 2: This table shows the different delay sequences test 3 participants experience, in a latin-square format. Each row contains a different sequence of delays for each participant.

APPENDIX H - Full Table of Machine Learning Models

	-Models	Categories														
		CT				**CD/T**			**CG**				**LLM**			
		*I-I	*I-V	*V-V	*I-3D	I-T	*V-T	*A-T	T-I	T-V	T-A	T-3D	T-T	T-C	T-S	T-Alg
Network of-	ChatGPT (4)	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
	Google Gemini	×	×	×	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
	LLaMA 2	×	×	×	×	×	×	×	×	×	×	×	✓	✓	✓	✓
	Apple - ML	×	×	×	×	✓	×	✓	×	×	✓	×	✓	×	×	×
	Meshy	×	×	×	✓	✓	×	×	✓	×	×	✓	×	×	×	×
Individual-	GPT4	×	×	×	×	×	×	×	×	×	×	×	✓	✓	✓	✓
	Megatron-LM	×	×	×	×	×	×	×	×	×	×	×	✓	✓	✓	✓
	Midjourney	✓	✓	×	×	×	×	×	✓	✓	×	×	×	×	×	×
	Stable Diffusion	✓	✓	✓	×	×	×	×	✓	×	×	×	×	×	×	×
	DALL-E 3	✓	×	×	×	✓	×	×	✓	×	×	×	×	×	×	×
	GPT4 Vision	×	×	×	×	✓	✓	×	×	×	×	×	×	×	×	×
	AsticaVision	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×
	Otter.ai	×	×	×	×	×	×	✓	×	×	×	×	✓	×	×	×
	Whisper	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×
	Eleven Labs	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×
	GPT4 TTS	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×
	Pika.art	×	✓	✓	×	×	×	×	×	✓	×	×	×	×	×	×
	Sora	×	✓	✓	×	×	×	×	×	✓	×	×	×	×	×	×
	AlphaCode	×	×	×	×	×	×	×	×	×	×	×	✓	✓	×	✓
	GitHub Copilot	×	×	×	×	×	×	×	×	×	×	×	✓	✓	×	×
	Shap-e	×	×	×	✓	×	×	×	×	×	×	✓	×	×	×	×

Table 3: Table presenting comparisons of various generative ML models, categories and sub-categories, **CT:** Content Transformer, **CD/T:** Content Describer/Transcriber, **CG:** Content Generator, **LLM:** Large Language Models, **I-I:** Image-to-Image, **I-V:** Image-to-Video, **V-V:** Video-to-Video, **I-3D:** Image-to-3DModel, **I-T:** Image-to-Text, **V-T:** Video-to-Text, **A-T:** Audio-to-Text, **T-I:** Text-to-Image, **T-V:** Text-to-Video, **T-A:** Text-To-Audio, **T-3D:** Text-to-3DModel, **T-T:** Text-to-Text, **T-C:** Text-to-Code, **T-S:** Text-to-Science, **T-Alg:** Text-to-Algorithm. Symbols ✓and respectively denote whether a model fits the category. Categories marked with the prefix * have been appended to the existing taxonomy described by [Brizuela and Merch, 2023] See Table 1 for references