Exploring the Application of AI in Sound Design for Video Games

Abstract

By assisting in the creation of video games, artificial intelligence (AI) is revolutionising the manufacturing process. This is particularly true for sound design, which is crucial for giving the game's characters a sense of immersion and for giving the narrative depth. Though they work well, traditional sound design methods can be tiresome and don't always fit in with the dynamic environments of modern games. This thesis presents and assesses a new pipeline that uses generative AI to improve and automate game audio production. The program primarily converts soundless video game footage into detailed, descriptive text using the LLaVA (Large Language and Vision Assistant) architecture. This text instructs a new latent diffusion model called AudioLDM 2 to create scene-appropriate sounds and soundscapes. This work investigates retrieval-augmented methods that improve the variety and precision of the generated audio in order to address important issues in text-to-audio generation, particularly the long-tail problem of inadequately represented sounds. We employ a mixed-methods comparative analysis of soundscapes created using fully AI-generated, AI-assisted, and conventional manual design methodologies in order to evaluate the efficacy of this pipeline. In order to provide empirical evidence on the feasibility of AI-generated audio in meeting the aesthetic and functional standards of contemporary video games and to investigate its potential as a collaborative tool for sound designers, the evaluation focuses on metrics of immersion, emotional engagement, and audio-visual coherence.

Chapter 1: Introduction

1.1 The Sonic Dimension of Interactive Worlds

Sound is a vital part of playing video games that many people forget about. In conjunction to audio feedback [1], it is an important tool for developing stories, making players feel connected to them, and getting them to fully engage with the game. The noises in a game, like the conversation, sound effects, music, and background noise, help to show off the personality of the virtual world, focus the player's attention, and give essential information about the gameplay [2]. Sound design is an important part of gameplay in first-person shooter (FPS) games, not only an atmospheric feature. Directional audio cues from footfall and gunfire tell gamers where enemies are and what critical things are happening in ways that pictures alone can't [29]. This acoustic input has a direct effect on how well a player does and how they make decisions [14].

The deep psychological effects of sound are shown by how audio cues change a player's mental and emotional state, from the scary music in a horror game to the satisfying "clink" of getting a reward [3]. Well-designed audio can make a player feel more present, realistic, aware of their surroundings, and able to locate sounds [30]. This deep integration of sound and movement creates a strong link between what the player does and what they hear. Scholar Karen Collins calls this "kinesonic synchresis" [4]. Audio is the major means for visually impaired players (VIPs) to get around and enjoy the game environment, but for many other players, it's just an added feature [29]. So, sound design is an important part of making interactive media, not something that comes after the fact.

1.2 How Contemporary Games Present a New Difficulty for Conventional Sound Design

Traditional video game sound design is a time-consuming and painstaking process that requires thousands of individual audio assets to be created, selected, mixed, and implemented by hand [5]. While this personalised method can produce high-quality, emotionally impactful outcomes, it is laden with difficulties that have been exacerbated by the magnitude and intricacy of contemporary games. The conventional method is becoming more and more memory-and time-intensive due to the ever-increasing quantity of interactive media, especially massively multiplayer open-world and procedurally generated environments [11].

In addition, conventional approaches frequently use a limited collection of audio loops that have already been recorded and a set of triggers that are based on events. This might cause the audio to become repetitive and unresponsive to the ever-changing, unexpected aspects of emergent gameplay, which can disrupt the player's immersion [8]. Development resources are greatly taxed by the complexity of contemporary audio pipelines, which are tasked with handling many tasks such as spatialization, digital signal processing (DSP), dynamic mixing, and voice communication protocols [28]. Modern interactive entertainment relies on real-time, context-aware experiences, which pre-authored audio simply cannot provide. Because of this, there is an immediate and obvious need for an intelligent, scalable, and efficient method of making game soundscapes.

1.3 The Potential of Generative AI for Music Composition

A new age in sound design has dawned with the advent of sophisticated Generative Artificial Intelligence (GAI), which provides formidable answers to the problems posed by scalability and dynamic complexity [9]. GAI is an umbrella term for a group of artificial intelligence systems that can learn the underlying patterns in massive amounts of data and then use those patterns to generate new content like text, images, and audio [9], [27]. Transformers and diffusion models, the foundational components of modern GAI frameworks, outperform their predecessors in terms of content generation capabilities [13].

This technology allows for an unprecedented amount of procedural audio generation (PAG) in sound design. As the player interacts with the game and its surroundings, AI can synthesise sounds in real-time using in-game settings, resulting in infinitely varied and responsive soundscapes [10]. This paves the way for a future where acoustic environments are genuinely dynamic, rather being limited to pre-recorded assets. The field has progressed to the point where universities provide specialised courses that teach students how to create modern game audio using DSP, spatial audio, and AI [11].

Using two kinds of state-of-the-art GAI models, this research investigates a new pipeline that operates at the cutting edge:

Machine Learning for Vision and Language: Large Language and Vision Assistant (LLaVA) and similar systems can interpret visual input and produce textual descriptions of it [13], [15]. In order to create contextually rich prompts that depict the on-screen action, setting, and mood, LLaVA analyses game graphics.

Translational Audio Models (TTAs): Latent diffusion models are used by systems such as AudioLDM 2 to convert textual descriptions into comprehensive, high-fidelity audio, which includes speech, music, and sound effects [12].

This thesis investigates a technique that may automate the development of contextually aware soundscapes directly from visual gaming by integrating these two powerful GAI technologies. This represents a substantial improvement over existing methods.

1.4 Primary Objectives and Methodology

This thesis delves at the artistic and practical possibilities of an AI pipeline that can generate video game soundscapes from start to finish. The main objective is to assess how well AI-generated sound can match the quality and immersion of human-led design. Additionally, the study will investigate whether this pipeline may serve as a collaborative tool that complements the human sound designer instead of replacing them.

In order to accomplish this, the research presents and executes a unique two-step pipeline that is based on generative AI:

First Step: Generating Text from Video: A Large Language and Vision Assistant (LLaVA) processes gameplay footage without the original soundtrack. A series of visual frames are analysed by the LLaVA architecture, which links a visual encoder to a Large Language Model (LLM) [13], in order to generate detailed text prompts. The model creates a story that depicts the scene's objects, events, and environment by utilising its ability to handle changes in time [18].

The second step is text-to-audio synthesis, which involves feeding the produced text prompts into AudioLDM 2, a latent diffusion model that converts the textual descriptions into ambient music and sound effects that are contextually relevant and of high quality [12]. In order to create a full soundtrack that corresponds semantically to the commands extracted from the game footage, the model employs its internal "Language of Audio" (LOA).

Each video clip's sound design will be created in one of three ways utilising this pipeline: one completely generated by AI, one with AI assistance (using AI to choose from a library), and the other one that is manually developed in the traditional sense. In order to compare and contrast the efficacy and reception of each strategy, a mixed-methods study will be carried out.

1.5 Research Questions

- In order to assess the proposed pipeline comprehensively, this study is led by a main research question and multiple sub-questions:
- First Research Question: What is the best way to create context-aware and immersive audio for video games using an AI-driven pipeline that uses vision-language and text-to-audio models?
- This inquiry seeks to ascertain whether or not the suggested LLaVA-to-AudioLDM 2 workflow is both practical and efficient in its entirety. This study aims to find out whether such a system can accurately convert in-game visual events into comprehensible and convincing audio.
- Question 1's subquestions concern the following: how does the quality of soundscapes made entirely by AI stack up against AI-assisted and more conventional human-crafted soundscapes in terms of immersion, emotional impact, and audio-visual flow?
- The heart of the assessment is being probed by this query. This research will experimentally evaluate the perceived quality of AI-generated audio compared to human craftsmanship by comparing the three versions of the sound design. The criteria used in this assessment have been confirmed in studies on player immersion [8].
- Question 2: In the context of game production, what are the advantages and disadvantages of the LLaVA-to-AudioLDM 2 pipeline in terms of efficiency, semantic correctness, and creative control?
- Enquiring about the pipeline's feasibility is the purpose of this question. Issues like the "long-tail problem" (i.e., the challenge of producing unusual sounds) in TTA creation [25] and the possibility of "hallucinations" (i.e., semantic incompatibilities) are examined, along with the substantial opportunity for improved production efficiency.
- Question 3: How can this AI pipeline work in tandem with human sound designers to enhance their work, rather than supplant them?
- The human element implications of GAI in creative domains are addressed by this question. It delves further than a basic "AI vs. human" comparison to imagine a world where AI tools take care of mundane jobs, allowing designers more time for strategic thinking and fine-tuning.

1.6 Thesis Contribution

Game design, sound design, and human-computer interface are three areas that this dissertation intends to significantly advance. To start, it offers a realistic framework that researchers and developers can adapt by proposing and validating a revolutionary end-to-end pipeline for AI-driven sound design. Additionally, it will provide important information on the strengths and weaknesses of GAI in the creative realm by conducting one of the first empirical comparisons of AI-generated, AI-assisted, and traditionally created soundscapes in a video game setting. In conclusion, this study will shed light on the potential of AI and humans working together in the field of sound design in the future, pointing the way towards a future where generative technologies foster artist agency, innovation, and the creation of immersive, interactive experiences.

Chapter 2: Literature Review

2.1 How Game Audio Has Changed Over Time

Through the years, video game audio has progressed from simple, static noises to elaborate, interactive acoustic environments, a story of constant technical and artistic development. The early arcade games made use of basic synthesised sound effects, such as "beeps and boops," which were a result of technology constraints but were crucial in laying the groundwork for interactive auditory feedback [4]. Technological advancements allowed sound designers to start using prepared music and pre-recorded samples. Nevertheless, these parts were frequently used in a static, unresponsive loop; the music didn't vary no matter what the player did or what happened on screen. As game worlds become more complicated and player agency increased, this static method became unsuitable. It was fine for linear experiences, though.

Adaptive audio, which refers to sound systems engineered to respond in real-time to occurrences in gameplay, was developed in response to the industry's quest for more realism and responsiveness [10]. With this new way of thinking, dynamic soundscapes are possible, enabling the seamless transition of musical segments from exploration themes to combat anthems through methods such as horizontal re-sequencing and vertical re-orchestration [14]. Weather and time of day can also affect ambient noises, and sound effects can be fine-tuned to better represent the game's mechanics and the context in which the player is interacting with it. The precise form of this adaptation is crucial, according to research; synchronising sound with broad game states (like level changes) can affect performance and cognitive load differently than synchronising it directly with fine-grained player actions (like button presses) [14].

The role of sound in contemporary video games is complex, encompassing both the enhancement of immersion and the provision of crucial input on gameplay. For example, sound design is an essential gameplay mechanic in first-person shooter (FPS) games. When compared to images alone, directional audio cues such as gunfire, enemy vocalisations, and footfall provide players a much better idea of adversary positions, threat levels, and important events [29]. A player's performance, situational awareness, and decision-making are all affected by this audio feedback. In what scholar Karen Collins calls "kinesonic synchresis," the player's physical actions become fused with the resulting sound, giving them a strong sense of agency and embodiment, and the connection between player input and auditory feedback is inextricable [4].

One of the many facets of player immersion—a multi-stage psychological state that can be broadly classified into stages of engagement, engrossment, and absolute immersion—is the use of sound [8]. A well-designed soundscape helps to create a more realistic, responsive, and emotionally engaging gaming world, which in turn strengthens the bond between the player and the digital setting. One of the most important aspects of accessibility is sound design, since for many players—particularly visually impaired players (VIPs)—audio is not merely an aid but the main way to experience and navigate the game environment [29]. The technological difficulties are heightened by the fact that contemporary audio pipelines must handle a great deal of information, including spatialization, digital signal processing (DSP), dynamic mixing, and standardised in-game speech communication [11], [28]. It is evident that more intelligent and automated methods are needed to address the challenges posed by procedurally produced content and emergent gameplay, as the human, trigger-based approach to adaptive audio continues to be a major bottleneck.

2.2 A New Era of Generative Artificial Intelligence for Media Synthesis

Generative Artificial Intelligence (GAI) is a cutting-edge field of study that can produce new media types including text, photos, audio, and video [9]. To generate original, innovative content in response to input, GAI models learn the distribution of training data, in contrast to discriminative AI that focuses on data analysis and classification [9]. Diffusion Models, Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs) are a few of the important architectures that are driving this paradigm change. Diffusion models are now considered state-of-the-art for high-fidelity media generation, while GANs and VAEs have demonstrated success. Using a guided, step-by-step approach, they learn to progressively add noise to data in a "forward process" before training a neural network to do the opposite, so producing clean data from pure noise [9], [33]. Particularly well-suited to complex media like audio, this systematic denoising approach generates outputs that are both detailed and cohesive.

A watershed moment in AI's development has occurred with the emergence of large-scale, pre-trained foundation models [13]. These models can be used as a foundation for several downstream tasks by utilising transfer learning; they were trained on large and diverse datasets [34]. The GAI systems utilised in this thesis are built on top of this methodology. The models utilised include Vision-Language Models (VLMs) for visual content understanding and Latent Diffusion Models (LDMs) for audio synthesis. A paradigm shift from "information recovery" (i.e., faithfully recreating a pre-made asset) to "information regeneration" (i.e., creating something from scratch) is shown by this integration [27]. The new paradigm involves compressing a semantic representation and then generating the desired material on demand. One major goal of this thesis is to apply this capability to video game sound design so that it can transform high-level, cross-modal semantic information into rich media.

2.3 The LLaVA Architecture: A Multimodal AI for Visual Understanding

Accurately understanding the visual events in a game in real-time is a key first step in generating context-aware music. For this, you'll need a model that can translate between numerical values and their semantic equivalents. By combining cutting-edge vision and language understanding, a subset of VLMs known as Large Language and Vision Assistants (LLaVA) has shown exceptional multimodal reasoning capabilities [13].

2.3.1 Core Architecture and Training Methodology

The main innovation of the LLaVA architecture is the efficient and straightforward way it connects a pre-trained visual encoder with a pre-trained LLM. For feature extraction, a common solution employs CLIP's visual encoder (ViT-L/14) in images. After learning a solid, shared embedding space from a large dataset of internet-sourced image-text pairs, CLIP's strong zero-shot capabilities led to its selection [13]. Vicuna, an LLM famous for its excellent conversational and instruction-following abilities[13], is trained using these visual cues input into a basic, trainable projection matrix (a multi-layer perceptron, or MLP) that maps them into the word embedding space of an LLM. Because of its architecture, the model may take advantage of the LLM's superior ability to follow instructions in order to process and reason about visual material.

To ensure this relationship is effective, the training process is staged strategically:

Prior to training, feature alignment: Initial training of the model is done on a massive dataset consisting of image-text pairs. At this point, we're only training the projection matrix and have frozen the visual encoder and LLM weights. Without completely wiping out the LLM's prior knowledge, this stage effectively trains the model to match visual features with their semantic representations in the vocabulary [13].

To fine-tune the model, visual instruction tuning is used on a smaller dataset of multimodal instruction that is of higher quality. Images in this dataset often accompany detailed question and response pairs based on conversation, description, or reasoning. At this point, the model is trained to follow human directions and to generate precise, context-aware textual answers from visual input by training the LLM and the projection matrix [15].

By completing these two steps, LLaVA is able to demonstrate remarkable general-purpose visual awareness, which allows it to respond to visual cues with free-form ideas and carry on sophisticated visual conversations. Improved reasoning, optical character recognition, and world knowledge capabilities have been achieved in subsequent iterations of this method, such as LLaVA-NeXT [35].

2.3.2 Modularity and Application-Dependent Tuning

The flexibility of the LLaVA architecture is one of its strongest points. Its architecture permits efficient fine-tuning on specialised datasets, even if it is trained on general-domain online photos, which results in an explosion of domain-specific variants. This has been proven in many different domains, which shows how reliable the model is.

For example, LLaVA-Med outperformed general-purpose models after being fine-tuned on biomedical image-caption pairings to answer queries on medical imagery [16]. Models like PA-LLaVA, which aims to understand images from human pathology [17], and LLaVA-NeXT-Med, which aims to provide more advanced medical VQA [35], have further developed this specialisation. Using domain-specific alignment, these models demonstrate how a generalised model can be grounded in the specialised field's complex lexicon and visual patterns.

Many other fields can also benefit from this versatility, such as:

In the field of remote sensing, LLaVa-RS was created to serve as a consistent paradigm for satellite image captioning and change detection [19].

For the purpose of agricultural plant pathology diagnosis, LLaVA-PlantDiag combines vision and language capabilities [20].

For use in industry, scientists have fine-tuned LLaVA using contextual instruction tuning to establish its knowledge on technical product manuals for oil and gas [22].

The Humanities and the Arts: LLaVA-Oil Painting Expert annotations helped refine appreciation to the point that it could offer AI-driven analysis and discussion of oil paintings [21].

In order to forecast popular tourist locations and produce image-based reviews, LLaVA-Tour drew on a massive dataset from Japanese travel websites [36].

While the basic LLaVA model does offer generic visual reasoning, domain-specific fine-tuning is necessary for optimal performance on particular tasks, as this substantial body of research shows. An example of this is the Quilt-LLaVA project, which used localised narratives from instructional films to build a histopathological instruction dataset [31]. The goal was to develop a high-quality dataset that could be used for instruction in the target domain. While it is possible to employ a generic pre-trained LLaVA, this thesis's methodology is based on the idea that more accurate and contextually relevant textual descriptions might be obtained by fine-tuning the model using a dataset of gaming footage.

2.3.3 Use with Sequential and Temporal Data

Understanding time-based data, like video, is a major obstacle for traditional VLMs. Understanding a video game necessitates processing a succession of frames to identify events, actions, and dynamic changes in the environment, in contrast to a basic LLaVA model's ability to explain a single static image. To solve this, the LLaVA architecture was successfully expanded. Models that incorporate techniques to process several frames simultaneously, such as Video-LLaVA and TX-LLaVA, expand upon the fundamental structure [18].

For instance, TX-LLaVA can examine the evolution of a set of chest X-rays over time by feeding the visual encoder a number of images and letting the LLM deduce their differences [18]. The process of analysing gaming footage is highly similar to this. The model can progress from static object recognition (like "a character is standing") to dynamic event description (like "a character is running towards a building while firing a weapon") by processing a series of frames. Since the main drivers of dynamic sound design are the game's actions and narrative flow, this capacity is crucial to the proposed methodology since it enables LLaVA to generate text prompts that capture these elements. For the initial step of the audio generating pipeline, LLaVA is a good fit because of its capacity to produce descriptive prompts for dynamic scenarios.

2.4 Random Sound Creation Using Latent Diffusion Models: AudioLDM 2

The semantic "what" is provided by Vision-Language Models, while the aural "how" is handled by Text-to-Audio (TTA) models. Latent diffusion models have completely transformed TTA generation, as they are very good at creating high-fidelity audio from textual cues [9]. One of the best options for constructing varied and extensive game soundscapes is AudioLDM 2, a state-of-the-art framework that offers a holistic approach to producing music, voice, and sound effects [12].

2.4.1 Fundamental Design: The Audio Language and Its Dissemination Mechanism

One global audio representation that the AudioLDM 2 framework is based on is the "Language of Audio (LOA)." An intermediary representation of the semantic and acoustic features of audio that is modality-agnostic, this idea is vital. The LOA is built using a self-supervised pre-trained AudioMAE (Audio Masked Autoencoder), which is genuinely general-purpose because it learns to represent audio data without depending on domain-specific inductive biases [12]. With this one, uniform "language," the model can manage all the different kinds of sounds found in a game environment, from the background noise and footfall to the dialogue between characters and the impacts of explosions.

With its modular design that boosts its flexibility and power, AudioLDM 2's generation method is intentionally two-staged:

The translation of text into letters of authorisation: At the outset, a GPT-2 language model mediates between the two languages. The input text prompt is transformed into the matching LOA sequence by this process. The translation is guided by embeddings from a CLAP (Contrastive Language-Audio Pretraining) model to achieve a high semantic alignment between the text and the intended audio. CLAP has been pre-trained on a large dataset to embed a common embedding space for paired audio recordings and text descriptions. The GPT-2 model is able to produce a LOA sequence that faithfully reproduces the target audio properties because it obtains a detailed, audio-aware text representation from CLAP's text encoder [12], [26].

Generating Audio from LOA: The actual audio synthesis is carried out via a latent diffusion model (LDM) in the second step. In the first step, you construct a LOA sequence. This LDM relies on it. It takes unstructured noise as a starting point and uses an iterative process of reverse diffusion (denoising) to transform it into a structured representation of the target sound. Crucially, this operation takes place within the compressed latent space of a Variational Autoencoder (VAE) rather than in the high-dimensional waveform or spectrogram space directly. The computational cost is reduced while the fundamental elements needed for high-quality reconstruction are preserved when operating in this lower-dimensional space [12], [33]. The final, high-fidelity audio waveform is created by converting the mel-spectrogram that is produced by the LDM's target latent representation using a HiFi-GAN vocoder [12].

The core LDM may be self-supervisedly trained on large volumes of unlabelled audio using this modular, two-stage architecture, which provides a strong basis for further fine-tuning for specialised conditional generation tasks [12], [37]. This method of pre-training and fine-tuning has been tested and proven to be quite effective, particularly in cases when there is a lack of labelled data, or data scarcity [37].

2.4.2 Raising the Bar for Generational Diversity and Quality

The long-tailed generation problem is one of the major obstacles that TTA models must overcome. This is because, despite their power, TTA models are imbalanced in their training data and perform well on common sounds (like "dog barking") but poorly on rare, complex, or nuanced ones (like "a small metal object dropping on a ceramic tile in a reverberant room") [25]. In order to build a convincing world, video games rely heavily on such detailed, long-tail noises. Various initiatives for improvement have been devised to tackle this and other quality challenges.

Retrieval augmentation is a highly effective strategy. To improve upon the regular AudioLDM pipeline, the Re-AudioLDM model searches a database for audio-text pairs that are related to the input prompt and share semantic similarities. Diffusion is further guided by the features extracted from these recovered samples. The model's capacity to provide precise and comprehensive audio for under-represented or extremely specialised cues is greatly enhanced by the inclusion of tangible examples of the desired sound [25].

Assuring the resulting audio's perceived quality and integrity is another significant obstacle. Although post-filtering techniques can enhance outcomes, they are computationally inefficient [26]. These methods include creating numerous samples and choosing the best one based on its CLAP similarity score. A more refined approach would be to incorporate quality assurance within the training itself. To do this, the Latent CLAP Loss method incorporates an additional loss term into the model as it is being fine-tuned. Minimising the distance between the target audio's CLAP embedding and the LDM's latent output is the immediate goal of this loss. It improves overall quality and eliminates the need for inefficient post-filtering at inference time by encouraging the model to generate audio that is fundamentally more true to the target sound from the outset [26]. Another major issue in generative music and sound design is plagiarism, therefore data augmentation approaches are essential for making things more interesting and less likely to happen. Methods such as the beat-synchronous mixup techniques suggested for MusicLDM show how creative data manipulation can broaden the training distribution and motivate the model to generate more unique outputs [32]. These techniques recombine existing audio samples in a musically coherent manner.

2.4.3 Efficiency and Deployment Considerations

Since large diffusion models can be unsuitable for real-time applications like video games due to their high computational cost and enormous memory footprint, they have not been widely used [24], [39]. To solve this problem, optimisation experts recommend Post-Training Quantisation (PTQ). Following training, PTQ converts the model's parameters from their usual 32-bit floating-point values to their low-bit equivalents, such as 8-bit integers, therefore reducing the model's complexity. Without the need for expensive retraining, this drastically decreases the model's size and speeds up inference [24]. The quantisation mistakes may build up across the several denoising processes, reducing the end audio quality, making it difficult to directly apply PTQ to diffusion models. To address this, specialised frameworks are being created, such as PTQ4ADM, which employs temporal dynamic quantisation and coverage-driven prompt augmentation to efficiently quantise audio diffusion models while maintaining their generative performance [24]. Integrating such robust generative models into the rigorous, real-time settings of video games necessitates these efficiency considerations.

2.5 Bridging Generative AI and Immersive Gaming

Improved player immersion—a complex psychological state with many facets that can be generically classified into stages of engagement, engrossment, and total immersion—is the end goal of advanced sound design [8]. The credibility, responsiveness, and emotional resonance of a gaming world are all enhanced by its soundscape. This thesis proposes a generative AI pipeline that can accomplish this goal by integrating the visual understanding of LLaVA with the audio synthesis of AudioLDM 2. The result will be audio-visual experiences that are both highly interactive and constantly evolving.

One major shortcoming of conventional sound design is the absence of comprehensive, real-time visual context; this is remedied in the first step of the pipeline, which is LLaVA-powered. The system is able to provide narrative and descriptive text prompts that capture the action's progression and the environment's changes because it makes use of LLaVA's capability to process temporal sequences of frames [18]. This guarantees that the semantic data used to train the audio generation model accurately represents the ever-changing nature of the action, laying the groundwork for an aural environment that is intrinsically timed with what's happening on screen.

This rich semantic information is then transformed into high-fidelity audio in the second stage, which is driven by AudioLDM 2. A full and consistent soundscape can be generated via the model's comprehensive framework, which can provide everything from simple background noises to intricate Foley effects [12]. Importantly, the requirement for realistic sound detail is directly addressed by enhancement methods such as Latent CLAP Loss [26] and retrieval augmentation [25]. These techniques guarantee perceptually convincing and contextually relevant audio by making the model better at producing unusual sounds and increasing overall output fidelity.

An effective instrument for investigating the foundations of multi-modal perception in video games is this integrated pipeline. For a virtual world to be convincing, cross-modal perception—the brain's ability to combine data from several senses—must be present [1, 30]. According to studies, there is a strong correlation between our aural and visual perceptions of location; for example, we may use environmental noises as enough clues to create recognisable visual representations, and the same holds true for visual perceptions [34]. Instead of just superimposing audio over video, the suggested technique computationally extracts the audio from the video. As a result, the player feels even more immersed and empowered by the game's audiovisual elements. An illusion of realism in the virtual world is enhanced when, for example, the sound of footfall precisely matches the character's stride on a given surface, or when the boom of an explosion is semantically connected to its visual size and intensity.

Using the theoretical framework of semantic communication, we can view this complete procedure [27]. According to this theory, LLaVA plays the role of the "semantic encoder," taking a high-dimensional visual input (the game footage) and translating its core meaning into a compressed, low-dimensional output (the text prompt). The "semantic information" that is sent along is this query text. After obtaining this data, AudioLDM 2 plays the role of the "semantic decoder," creating the associated audio data for the human listener. This method is highly congruent with the aims of procedural and dynamic sound design since it places an emphasis on meaning transmission rather than signal fidelity.

A distinct knowledge vacuum is revealed through the process of literature synthesis. Although video learning machines (VLMs) and time-series audio synthesis (TTA models) have each received extensive academic attention for their own purposes, the integration of these two technologies into a unified pipeline for the purpose of procedural video game sound design is an emerging field of study. Additionally, there is a lack of a thorough empirical comparison of this pipeline to conventional human-centric processes in the existing literature. This thesis sets out to rectify that situation by offering a workable methodology together with empirical evidence on the effectiveness of GAI in a rigorous creative field. By methodically studying the effects of these artificial intelligence (AI) generated, highly synchronised soundscapes on player perception, emotion, and behaviour, it will push the limits of interactive sound design [8].

Chapter 3: Research Methodology

In this chapter, we lay out the comprehensive, multi-stage research strategy that will be used to study how AI can be included into video game audio. This study's central experiment compares and contrasts AI-generated soundscapes with more conventional, human-led sound design methods in order to draw conclusions about their perceived quality and efficacy. We start by describing the research design, and then we go into detail about the atmosphere of the custom-built prototype game. From LLaVA's video-to-text generation to AudioLDM 2's text-to-audio synthesis, the following sections offer a detailed technical explanation of the novel AI pipeline. To wrap up, the chapter discusses the ethical considerations that guided this research and lays out the structure for collecting data, including the user study protocol and data analysis methodology.

3.1 Research Design

Using a controlled video game setting, this study uses a mixed-methods, comparative experimental strategy to investigate AI-generated soundscapes' efficacy and perceptual quality. The approach is based on a within-subjects design, which means that every participant uses the same gameplay prototype under three different experimental settings. In this setup, the sound design process is the sole independent variable, thus we can examine its effect on the player's experience with pinpoint accuracy.

The following are the three prerequisites:

Situation A: Conventional Sound Design (Control). An immersive environment that the researcher built from scratch utilising high-quality sound resources and the popular Digital Audio Workstation (DAW) program. This state is the starting point, symbolising the long-standing, human-led creative benchmark for game audio development.

- State B: AI-Generated Soundscape with Prior Training. A soundtrack created by the versatile, pre-trained cvssp/audioldm2 model. In this state, we are evaluating the performance of a basic text-to-audio model without domain-specific adaption; it is entirely automated and AI-driven.

- Situation C: Artificial Intelligence-Simulated Fine-Tuned Soundscape. An aural environment that simulates the results of an AI model that has been trained to work in a certain domain. The purpose of this condition was to see whether the outcomes would be better for a model that was trained on a curated dataset that was specific to the game.

The goals of the thesis depend on the use of a comparative experimental design. To determine that audio is the most important factor, we used all three of these sound design approaches to the same visual and interactive gameplay sequence. This methodology allows for a thorough evaluation of its impact on player perception inside a controlled environment, answering the main study issue of the feasibility of AI-generated audio in comparison to a human-crafted standard.

A more complex examination is made possible by including two separate AI circumstances, namely pre-trained and fine-tuned, rather than just a "AI vs. human" comparison. This two-pronged AI strategy permits investigation of the outcomes of domain specialisation and transfer learning. It shifts the focus from whether AI can competently produce sound to how the adaptation (fine-tuning) process affects the soundscape's quality, contextual relevance, and creative potential.

A mixed-methods evaluation framework is used in conjunction with this structured experiment. To fully grasp the findings, it is necessary to gather data in two different ways. Evidence of statistical significance about the quality, performance, and subjective evaluations of each condition is provided by the quantitative component, which consists of structured questionnaire data obtained using Likert scales. A player's "what" and "how much" preferences can be satisfied using this data. At the same time, the subjective intricacies of the player experience are captured by the qualitative component, which is based on open-ended survey questions. When it comes to figuring out the "why" behind the quantitative results, qualitative data is crucial. It provides valuable insights into the emotional impact, room for creative interpretation, and particular strengths or shortcomings that numbers can't express. The research confirms and elaborates upon these hypotheses on AI's potential in the artistic field of sound design by combining quantitative and qualitative datasets.

3.2 The Prototype Game Environment

A premade prototype game, purposefully found with technical and aesthetic simplicity to meet the research goals, serves as the empirical basis of this study. The player controls a solitary avatar through a linear world in this simplistic 3D FPS (first person shooter) game. To eliminate distracting visuals and keep the player's attention fully on the sound, the developers opted for a minimalist, "barebones" design. Study participants were more able to link differences in player experience to the impact of sound design when cognitive load was reduced due to simplified gameplay mechanics, simplified narratives, or visually overwhelming images. This controlled environment is perfect for studying the effects of sound on immersion and audio-visual coherence since the player's attention is not occupied by complicated visual cues, allowing them to notice and assess the subtleties of the game's audio.

The 3D FPS prototype game environment, featuring the player character and important interactive elements, is depicted in this screenshot (PLACEHOLDER FOR FIGURE 3.1).

A final gameplay sequence was devised and recorded to offer a uniform and repeatable experience for all participants and AI operations. This scene, which lasts about 60 seconds, shows a variety of the game's main features and how the environment interacts with them. The visual and interactive events are consistently presented to all participants and stages of the AI pipeline in this controlled order, guaranteeing consistency across all experimental settings.

A master video file was created at a resolution of 1920 x 1080 and a frame rate of 30 frames per second by capturing this sequence using screen recording software. A silent base video was created by later removing the original audio from this recording. All three experimental conditions relied on this silent master as their foundational visual layer, and the LLaVA analysis stage of the AI pipeline used it as its major input.

3.3 The Pipeline for AI-Powered Sound Generation

A unique two-stage pipeline that converts silent gaming footage into a full aural soundscape is the main technical contribution of this thesis. Using a combination of state-of-the-art generative models and GPU acceleration, the whole process was executed within a Google Colab Jupyter Notebook environment. By establishing a direct, data-driven connection between on-screen visual events and their accompanying auditory representations, the pipeline aims to automate what is typically an interpretive process that is focused on humans.

The whole AI-powered sound synthesis pipeline is shown in a high-level flowchart diagram, which goes as follows: Video Input -> LLaVA (Frame Extraction & Description synthesis) -> Text Prompts -> AudioLDM 2 (Audio Synthesis) -> Final Soundscape.

3.3.1 Stage 1: Video-to-Text Generation with LLaVA

Silent gameplay footage was processed using the Large Language and Vision Assistant (LLaVA), notably the llava-hf/llava-1.5-7b-hf model checkpoint, in the initial and critical step of the pipeline. The supplied LLaVA.ipynb script documents the full implementation for this step.

Here, we aimed to transform the video's raw pixel data into a structured, semantic, and textual representation that may act as good inputs for the model that generates sounds.

Several essential steps were involved in the process:

1. Initialisation of the Environment: First, we activated the GPU runtime and configured the Google Colab environment. Python libraries such as opencv-python, torch, accelerate, and transformers were installed. To facilitate quick inference, the Hugging Face Hub's pre-trained LLaVA model and its corresponding processor were subsequently transferred to the GPU.
2. To extract frames from a video, we used Python's OpenCV package to programmatically import the source file. Then, frames were extracted at a constant rate of 1 FPS. To avoid processing overhead and repetitive audio, we sparingly sampled at intervals that were neither too frequent nor too infrequent, ensuring that all major visual events in the gameplay were captured (e.g., the beginning of a jump and the moment of item collection). The Pillow (PIL) picture object format, the standard input format for the Hugging Face transformers library, was used to convert each extracted frame.
3. Generation of Prompts: The LLaVA model was used to process each extracted PIL image separately. At the beginning of each frame, the model was given a general, open-ended directive: USER: Please describe the scene in the picture in a few words, like a sound effect prompt. In contrast to more generic descriptions, this prompt structure directs the model to concentrate on specific items, character activities, and environmental elements in order to produce action-oriented, brief descriptions of the visual picture. One may say that the model's raw text output is a "snapshot" of the game's semantic state at that particular moment.

   Here is a code snippet from LLaVA.ipynb that shows the Python function that extracts frames and generates a description using the LLaVA model. It is placed in the placeholder for Screenshot 3.3.

4. Human-in-the-Loop Prompt Curation: During the first run, it was clear that LLaVA excels at detecting main actions (like "a character is jumping"), but it fails to reliably record all the audio events needed to create a full soundscape. The rustle of the character's garments, the background hum of the setting, or the distinct material sound of a footstep are all examples of subtle but important sounds that it could miss. A human-in-the-loop curation stage was incorporated into the process to tackle this problem. The first set of prompts provided by LLaVA was a solid starting point. The researcher then went back over this list and added to it by hand. In order to fill in the silences and make sure that every important event received a written hint, we made some extra prompts and put them in at the right times. A thorough audio brief could not have been produced without this involvement, which also serves as a significant discovery on the existing status of automated video-to-text systems in this context.
5. Data Structure: A Python list was methodically filled out with all the generated and manually curated text descriptions. The sequence in which these descriptions appeared in the list served as an implicit time stamp; for example, the first description appeared at

the one-second mark of the video, the second at the two-second mark, and so on. A script of the whole gameplay session that was time-aligned was the end result of this operation. Second stage of pipeline: text-to-audio synthesis received its direct input from this structured list of prompts.

3.4 Stage 2: Text-to-Audio Synthesis with AudioLDM 2

Data Structure: A Python list was methodically filled out with all the generated and manually curated text descriptions. The sequence in which these descriptions appeared in the list served as an implicit time stamp; for example, the first description appeared at the one-second mark of the video, the second at the two-second mark, and so on. A script of the whole gameplay session that was time-aligned was the end result of this operation. Second stage of pipeline: text-to-audio synthesis received its direct input from this structured list of prompts.

3.4.1 Condition B: The Pre-trained AI-Generated Soundscape

Using the Hugging Face Hub's standard, general-purpose cvssp/audioldm2 model checkpoint, the first AI-generated soundscape was constructed. The AudioLDM2.ipynb script details the implementation, which is a basic application of a generalist TTA model.

Following is the workflow: the AudioLDM2Pipeline was loaded using the diffusers Python library. This automatically instantiated all the required components, such as the variational autoencoder (VAE), the vocoder, and the CLAP model, which were used for text-audio embedding alignment and waveform synthesis, respectively. In Stage 1, the script generated a time-stamped list of textual cues, which it then iterated through. A matching audio waveform was generated for each prompt by invoking the pipeline. An important parameter for generating consistent and high-quality results was set to 7.0 for guidance scale, which encourages the model to produce audio that closely follows the text prompt, and to 200 for num_inference_steps, which provides a good balance between generation quality and computation time. The created audio clips were stored as separate.wav files, with names that matched their corresponding timestamps in the gaming video. You may now manually insert these sound effect and ambience files into a digital audio workstation (DAW) thanks to the directory they were created in.

3.4.2 Condition C: The Fine-tuned AI-Generated Soundscape

To explore how domain-specific adaptation affects generation quality, the second AI soundscape was conceptualised. While pre-trained models are well-versed in generalities, they might not have the refined aesthetic or musical taste needed for a niche creative endeavour. As a kind of transfer learning, fine-tuning improves the output's relevance, coherence, and quality within a given domain by tailoring a big foundation model to a smaller, domain-specific dataset.

Unfortunately, the technological hurdles that come with fine-tuning large diffusion models like AudioLDM 2—including high VRAM requirements, lengthy training durations, and hyperparameter tweaking sensitivity—were too great for the Google Colab environment. Model instability and convergence concerns were observed in the initial attempts at fine-tuning, which are recorded in the AudioLDM2Finetuned.ipynb script.

These technical obstacles prompted a change in methodology. A simulated fine-tuned condition was developed to test the basic hypothesis about domain specificity while preserving the three-condition experimental design. By depicting the result of a perfectly tuned model, this method keeps the experiment's conceptual integrity intact. Here is the procedure:

1.The process of creating a domain-specific dataset involved selecting 341 audio samples by hand. The prototype game's aesthetic and mechanical requirements informed the selection of these sounds, which include object gathering bells, ambient tones, event-triggering sound effects, and particular footstep textures. The intended "sonic palette" for the game was provided by this dataset.

2.Preparing the Dataset and Creating the Manifest: All 341 audio files were converted to a uniform format (16-bit WAV, 16kHz sample rate). This was followed by creating a manifest.csv file using the AudioToCSV.ipynb script. Important for tuning, this manifest has two columns: one for the relative path to each audio sample (e.g., "soft footsteps on stone floor," and another for a high-quality, descriptive description (e.g., "bright, magical chime for item collection"). The program learns new text-audio associations based on these captions.

3.To mimic the fine-tuned output, we used a bespoke 341-sound dataset to find the best audio file for each of the LLaVA-generated prompts in Condition C, rather than a technically fine-tuned model. When given the command "a character jumps," for instance, the researcher would access the custom dataset and choose the "jump.wav" file that had already been designed. This procedure mimics the ideal result of a highly tuned model that, when presented with a certain stimulus, would produce the exact, stylistically-aligned sound based on its training data. In order to conduct the comparison unaffected by the training process's technical challenges, this "Wizard of Oz" method offers a legitimate portrayal of a domain-adapted soundscape for the user research.

3.5 Establishing and Combining Experimental Environments

There were three separate iterations of the prototype game put together once the manual design (Condition A) and two AI generating processes (Conditions B and C) were finished. A DAW project was created for each circumstance and the individual audio clips were loaded into it. Using the timestamps linked to the LLaVA prompts, the clips were painstakingly timed to match the master gameplay footage. The different sounds were blended into a continuous and seamless soundscape through the use of small crossfades and level adjustments.

Loudness normalisation was the last and most important stage. All three mixed audio files were levelled to a consistent integrated loudness level of -23 LUFS (Loudness Units Full Scale), in accordance with the EBU R 128 standard, to remove perceived loudness as a confounding variable and guarantee the user study's validity. With these standards in place, we can be sure that any discrepancies that participants may perceive are due to variances in the actual sound design and not to fluctuations in volume. One of the three separate normalised soundscapes was applied to each of the three identical video files that made up the final result.

3.6 Data Collection and Evaluation Framework

A mixed-methods strategy was used that combined objective computational analysis with subjective participant feedback to assess the three situations.

3.6.1 Objective Audio Analysis

- A first technical examination of the audio assets was carried out utilising the AudioEvaluation.ipynb script as a supplemental measure to the user survey. This script calculates a set of quantifiable parameters for audio quality that are free of human bias in order to set a quantitative baseline. Important measures comprised:
- As a surrogate for the quality of artificially produced noise, the Signal-to-Noise Ratio (SNR) compares the strength of the target signal to that of the ambient noise.
- One typical metric in audio source separation is the Signal-to-Distortion Ratio (SDR), which measures the target signal's quality in relation to undesirable distortion artefacts and provides an overall score of audio fidelity.
- One reference-free metric is the Fréchet Audio Distance (FAD), which compares the generated audio's statistical distribution of embeddings to a real-world audio dataset in order to determine how realistic and natural the audio is. A lower FAD score suggests that the audio is more realistic.
- The subjective outcomes of the user study can be better understood in the technical context provided by these objective measurements.

3.6.2 Subjective User Study

Human volunteers in a user research served as the major evaluators.

To provide a balanced sample of casual players and persons with more developed critical listening skills, we recruited participants from varied populations, including university students, online gaming communities, and audio production forums.

• Method: All participants were exposed to and assessed all three conditions (A, B, and C) as part of a within-subjects (or repeated measures) design. The presentation of the three versions was counterbalanced using a Latin Square design to mitigate order effects, which occur when the sequence of exposure influences perception. An online survey tool was used to administer the study. After answering some basic demographic information, participants played the initial game version and then filled out a short survey. For the other two variants, this procedure was repeated.

• Survey Tool: Following each condition, a standardised questionnaire was given to the participants. Using 5-point Likert scales, with 1 representing Strongly Disagree and 5 representing Strongly Agree, this tool was developed to assess important constructs. Here were the structures:

Statements such as "The sound design was believable and consistent" and "The audio made me feel present in the game world" are used to measure immersion.

To gauge emotional engagement, we used statements like "The audio effectively conveyed the intended mood" and "The sound effects made my actions feel more satisfying."

"The sounds I heard were a good match for the visuals on screen" and "The timing of the sound effects felt perfectly synchronised." are examples of evaluation criteria for audio-visual coherence.

The overall quality was evaluated using expressions like, "Overall, the sound design was of high quality" and "I would be happy to play a full game with this sound design."

For each condition, the poll asked participants to explain what they liked and disliked, as well as offer any additional remarks, in addition to the Likert scores. The goal of collecting this qualitative data was to gain understandings that the quantitative measures could miss. The poll concluded with a section that asked participants to provide personal details, such as their age, gender, and game preferences.

[PLACEHOLDER FOR FIGURE 3.4: A sample of a Likert scale question from an online survey interface.]

3.7 Data Analysis Plan

• Quantitative Data Analysis: Statistical tools (like SPSS or R) will be used to analyse the data from the Likert scale questions. For the main purpose of determining whether there are statistically significant differences between the three conditions' mean scores across the core dimensions (Immersion, Engagement, Coherence, and Quality), a repeated-measures Analysis of Variance (ANOVA) will be utilised. If the results of the analysis of variance (ANOVA) are significant, we will apply post hoc tests (such as Bonferroni-corrected pairwise comparisons) to determine which conditions differ from each other.

• Thematic analysis will be used to the responses from the open-ended questions as part of the qualitative data analysis. A methodical reading, coding, and topic identification of the textual data is included in this procedure. By comparing and contrasting the various sound design approaches and providing context for the quantitative results, this analysis will provide light on the player experience and its merits and shortcomings.

3.8 Ethical Considerations

All individuals who were eligible to participate were given an informative document that explained the study's goals, methods, anticipated time frame, and data type. Before any participant could begin the study, they were asked to provide their informed consent. The privacy and confidentiality of the participants were guaranteed by anonymising all acquired data, including demographic information and survey responses. The participants were made aware that their participation was completely voluntary and that they might discontinue the study at any moment without facing any consequences.

Findings and Interpretation (Revised and Expanded) Chapter 4

4.1 Overview

The extensive results of the mixed-methods comparison study are presented and discussed in this chapter. The main goal is to analyse and compare three different approaches to sound design: conventional human-led design, artificial intelligence generation using a pre-trained model, and artificial intelligence generation using a simulated fine-tuned model. The main research concerns regarding the feasibility, quality, and creative potential of AI in video game sound design are sought to be answered empirically by this examination.

The outcomes are derived using an evaluation methodology that employs two distinct methods. To begin, a technical baseline is established by an impartial computational evaluation of the produced audio assets, which quantifies inherent attributes such as noise, clarity, and fidelity. Secondly, 11 people filled out a subjective user research that provides important perceptual data on how the gamers felt these soundscapes. Separate from the numeric scores derived from the structured survey is the rich qualitative feedback gleaned from the free-form questions that make up the user data set.

Using this information, the chapter is organised to form a unified story. We start by analysing the demographics of the study's participants to put the results in perspective. After that, to provide the groundwork technically, we present the findings of the objective audio analysis. Beginning with a comprehensive statistical presentation of the quantitative survey results, organised by the basic constructs of Immersion, Emotional Engagement, Audio-Visual Coherence, and Overall Quality, the major body of the chapter delves into a thorough examination of the user research. Afterwards, the quantitative tendencies are explained by doing a comprehensive theme analysis of the qualitative comments, which includes quoting specific participants.

The chapter concludes with a thorough discussion segment. Here we triangulate and synthesise the findings from the quantitative, qualitative, and objective sources. We will analyse the convergent evidence, answer the dissertation's main research questions, and place the findings in the larger context of AI-driven creative media.

I am the one who is being referred to here.

Section 4.2: People Who Took Part

Eleven people made it through the whole user research, giving us a comprehensive dataset to work with. In order to understand their comments and evaluate the results' generalisability, it is crucial to know what this group is made of. Table 4.1 summarises the demographic data that was collected in the final survey module.

The average age of the participants was 22.1 years old, and their ages varied from 19 to 32. The video game industry primarily targets young adults, which this generation falls squarely under.

Importantly, the sample showed a lot of interest in playing video games. A significant portion of the participants, 63.6% (n=7), described themselves as "regular gamers" who dedicate a few hours each week to gaming. The remaining one-third (n=4) stated that they were "dedicated gamers," meaning that they played for numerous hours per week. It is worth mentioning that out of all the participants who finished the survey, not a single one identified as "casual gamers." This goes to show that the people who gave the feedback had a solid grounding in modern video game design principles, especially when it came to sound design. Because of their familiarity, they probably have better hearing and more established standards for what makes good gaming audio.

In addition, 18.2% (n=2) of the participants claimed to have prior work or educational background in audio-related disciplines like sound design or music production. Their presence, however small, enriches the qualitative data with professional critique. Gaining insight into the normal user experience, the majority, 81.8% (n=9), addressed the work with a focus solely on the players.

Count (n) and percentage (%) of demographic metric categories

Participation Rate: 11 out of 12

Playing video games regularly (hours per week): 4.36 percent

Consistent (a few hours per week): 7.6 percent

Professional Audio Exp. In a majority of cases, 18.2%

Oh no! Only 8.1 percent

Table 4.1: Demographics of the Participants (n=11)

Save as a Spreadsheet

I am the one who is being referred to here.

4.3 Analysing Audio Objectively

An impartial technical assessment of the audio assets was carried out prior to exploring the participants' subjective impressions. In order to provide a fair comparison point, this research quantifies the intrinsic signal quality of the three different soundscapes. The AudioEvaluation.ipynb script was used to compute the metrics.

4.3.1 Quality Metrics for Pairs: A Measure of Integrity

As a "ground truth" or high-quality reference, the assets from the Traditional (Condition B) soundscape were compared pairwise with the AI-generated sounds to determine their fidelity. Ratios of Signal-to-Distortion (SDR), Signal-to-Noise (SNR), and Perceptual Evaluation of Speech Quality (PESQ) were the metrics employed.

Figure 4.1 shows the results, which were very clear. Contrasted with the reference, neither the pre-trained AI (Condition A) nor the fine-tuned AI (Condition C) assets fared well. The artificial intelligence waveforms have far more noise and distortion than the clean, traditionally created assets due to the persistently negative SDR and SNR values. Both the Pre-trained and Fine-tuned conditions were technically significantly less accurate than the human-designed sounds, albeit the Fine-tuned condition had slightly higher average SDR. This indicates that, as compared to more conventional methods of recording and editing, the signal purity is diminished due to artefacts introduced by the present implementation of the latent diffusion process.

As shown in Figure 4.1, the artificially created audio is compared to the traditional reference in a pairwise fashion.

4.3.2 SRMR: Reverberation and Clarity Measurements

One reference-free way to measure how clear an audio signal is is the Speech-to-Reverberation Modulation energy Ratio (SRMR). Reverberation is reduced and the sound is more crisp and distinct when the SRMR score is high. Figure 4.2 shows that the Traditional soundscape assets always had the best SRMR scores. This is in line with the standards of professional sound design, which call for mostly unprocessed audio with the occasional strategic use of reverb.

The pre-trained AI assets, on the other hand, had the worst average SRMR ratings. Participants' subjective feedback (discussed in Section 4.5) described these sounds as "muffled" or "underwater." The low SRMR scores imply the presence of unwanted, "baked-in" reverberation or a lack of definition in the generated sounds, which detracts from their clarity. These findings offer a technical explanation for the subjective feedback.

In Figure 4.2, we can see the SRMR scores for the assets under all three scenarios.

Two main points are drawn from the objective study, to sum up. As a first point, the audio assets that were created by humans are far more technically advanced in terms of clarity and fidelity. Furthermore, there was a technical issue with the Pre-trained AI model that was directly related to the bad perceptual evaluations given by users: the model produced the least clear sounds.

I am the one who is being referred to here.

4.4 Results of Quantitative User Studies: Player Perception Deconstruction

The numerical information obtained from the poll of users constitutes the backbone of the subjective review. Each of the three soundscapes was evaluated using a 5-point Likert scale across 14 separate questions. Overall Quality, Audio-Visual Coherence, Emotional Engagement, and Immersion were the four main constructions that were formed from these questions. Table 4.2 displays the means and standard deviations for each construct. A more nuanced look at how players felt is revealed by analysing these results in depth.

[! *IMPORTANT NOTICE: The initial study yielded the following Mean (M) and Standard Deviation (SD) values. Your final dissertation requires you to compute these values utilising solely the eleven participants' full responses. There is no change to the descriptive analysis.

Condition A: AI that has been pre-trained (M ± SD) Condition B: AI that has been traditionally trained (M ± SD) Condition C: AI that has been fine-tuned (M ± SD)

Percentage of total immersion: 3.61 ± 0.84, 4.24 ± 0.74, 3.73 ± 0.96

Affective Connection 3.70 ± 0.86 4.33 ± 0.69 3.49 ± 1.13

Three.82 ± 0.72, 4.63 ± 0.49, and 3.90 ± 1.05 for audio-visual coherence.

Quality as a Whole: 3.40 ± 0.80, 4.27 ± 0.81, and 3.53 ± 1.05

User ratings' means and standard deviations are displayed in Table 4.2. (Only when n=11 may the values be recalculated).

Save as a Spreadsheet

Section 4.4.1: Immersion

An essential component of good sound design is immersion, or the sensation of actually being in the game environment. With a mean score of 4.24, the Traditional soundscape triumphed in this category, indicating that players were most captivated by its polished and cohesive quality. With a mean score of 3.73 and a standard deviation of 3.61, the two AI conditions were very comparable in their performance. This suggests that the AI was unable to replicate the enhanced sense of presence offered by a human designer, even though it was able to provide some immersion.

Emotional Engagement, the Second Construct (4.4.2)

This component assessed the degree to which the music evoked a certain feeling and facilitated the completion of desired tasks. Participants found the sound effects and atmosphere of the Traditional condition to be the most impactful and successful, leading to its highest rating (M=4.33). Oddly enough, this category gave more points to the Pre-trained AI (M=3.70) than the Fine-tuned AI (M=3.49). Qualitative feedback suggested that the sounds in the Fine-tuned version were occasionally deemed "too loud" or "distracting," which could explain this surprising outcome by suggesting that they pulled the user away from the immersive emotional experience.

Fourthly, 4.4.3 Audio-Visual Coherence

We found the biggest performance gap in coherence, which is the credibility of the link between what we see and what we hear. The Traditional soundscape had a very low standard deviation (SD=0.49) and an extraordinarily high score (M=4.63), suggesting that participants strongly agreed that the sounds were an accurate representation of the on-screen events. Expert sound design is characterised by this. However, the Pre-trained AI (M=3.82) and the Fine-tuned AI (M=3.90) both achieved substantially lower scores. Players' faith in the game world depends on accurate temporal synchronisation and contextual matching, two areas where the automated pipeline falls short. This discovery highlights a crucial flaw in the pipeline.

4.4.4 Fourth Construct: General Excellence

This last structure was like a verdict on the whole soundtrack. Consistent with the general trend, the Traditional condition (M=4.27) was considered to provide the best overall experience. Both the pre-trained and fine-tuned AI settings were categorised as "mediocre" or "average" and awarded nearly comparable ratings (M=3.40 and M=3.53, respectively). This indicates that participants did not completely dismiss the AI soundscapes as lacking in quality; rather, they clearly ranked them lower than the human-designed alternative. Similarly, the last survey question, which asked if they would like a whole game with this music, reiterated that the conventional method was significantly better.

In conclusion, the quantitative evidence proves time and time again that the classic soundscape is the best option. Although they served their purpose, the AI-generated versions were lacking in audiovisual coherence and didn't shine in any particular area.

I am the one who is being referred to here.

4.5 Player Perspective in Qualitative Analysis

To understand the "why" behind the numbers, go no further than the qualitative information gleaned from the free-form survey questions. We can determine the relative merits of each condition by examining the expressions used by the eleven individuals who took part in the study. This research revealed four main topics.

Idea 1: The Quality Spectrum: "Polished" vs. "Underwater"

The basic audio quality and fidelity was the most talked about subject. When compared to the artificially made soundscapes, there was a huge chasm in perception.

Audiences were quite complimentary of the expert execution of the Traditional (Scene 2) score. People who took part in the study used terms like

- "Highest quality audio, perfect synchronisation, multiple layers of audio, overall enjoyable to listen to."

- "Much more polished."

- "Realistic, pleasant to the ear, rewarding." This language shows that you value the design's clarity, craftsmanship, and intentionality. In addition to their practical use, the sounds were also thought to be visually beautiful.

In contrast, the Pre-trained AI (Scene 1) received a lot of flak for being unreliable. Emotional feedback was given:

- "The audio quality was poor and it sounded like you were underwater."

- "Tense, engaging, low quality."

- "The enemy beeping. It felt unpolished." The player's subjective experience is fully supported by the low SRMR scores from the objective analysis, since the descriptions of being "underwater" and "muffled" are in perfect alignment with them.

Although it brought additional balancing issues, the fine-tuned AI (Scene 3) was considered as a more faithful version than the pre-trained one. Users had varying opinions on the matter, with one saying it was "more balanced," another saying it was "the hit sound effect felt out of place and too loud," and a third saying it had "too much audio details that reduces immersiveness." It seems that using higher-quality source sounds alone isn't enough; the skill of mixing and balancing is just as important.

Principle No. 2: Harmony and the "Uncanny Valley" of Auditory Experience

Audiovisual coherence was crucial, as indicated by the quantitative results. Many were impressed by how well the Traditional soundscape worked together:

- "Sound was synchronised with my action inputs."

- "The sound matched the most with the game."

On the other hand, participants found it annoying when the AI settings frequently entered a "uncanny valley" where the sounds were near but not quite accurate. For both AI situations, this was observed:

For the pre-trained sequence, "some sounds where not on point with the action."

- Regarding the fine-tuned scene: "The sound wasn't synchronised unlike scene 2." Some players were able to perceive the illusion broken by this millisecond-long desynchronisation, which led to lower immersion and quality scores.

Third Theme: Ambience's Double-Sided Sword

Disagreement arose over the scenes' ambient, or background, sound. Ambience for the Traditional scenario was understated and mostly unnoticed, which is a good indicator of good sound design because it adds to the atmosphere without being overly noticeable.

The AI scenes' ambient tracks, on the other hand, were both noticeable and divisive. One user praised the Pre-trained scene's "the background humming. It puts you in a different world," while another criticised it, calling it "constant very annoying background noise." The Fine-tuned version had the same problem, with one participant calling it "mildly annoying ambience." A player who liked the Traditional scene said it was because it was "preferrable over the obnoxious ambience noises in 1 and 3." This illustrates a major challenge for AI generation: making atmospheric tracks that don't get repetitive or annoying over time.

Fourth Theme: Strong Overall Preference

Participants were asked to identify the scene with the most engaging sound design overall in the final poll question. The findings provided a unified conclusion based on all the input that had come before:

Seven out of eleven participants (63.6%) were from Scene 2 (Traditional).

- On Stage 3 (Fine-tuned), two out of eleven individuals (18.2%)

- One out of eleven participants (9.1%) in Scene 1 (Pre-trained)

- One participant out of eleven (9.1%) expressed no preference.

The clear winner was the Traditional soundscape. Participant responses echoed the aforementioned themes; the "vote" winner elaborated on their selection by saying it had the "highest quality audio, perfect synchronisation, multiple layers of audio, overall enjoyable to listen to." The fact that the baseline Pre-trained AI was chosen by just one person highlights its overall inadequate performance on this project. Although it is not a full solution, the slight advantage of the Fine-tuned version over the Pre-trained one implies that domain-specific data can be useful.

I am the one who is being referred to here.

4.6 Conclusion: Drawing Connections

In order to answer the fundamental research questions and make sense of the study's results, the researchers used a triangulation of qualitative, quantitative, and objective data. There appears to be a distinct hierarchy of efficacy among the evaluated sound design approaches, as the results from all three data sources are extremely congruent. This section provides a summary of the findings and discusses their wider significance for AI-powered sound design.

1. Distinct Disparity: AI vs. Human Craftsmanship The main objective of the study was to evaluate the performance of an AI pipeline that was run automatically vs a more conventional approach that relies on human experts to create sound designs. The results show that human sound designers still have it beat when it comes to quality, consistency, and polish when compared to existing generative AI processes. This finding was not nuanced; rather, it was the overarching trend in all metrics.

At launch, the objective measures revealed a technological shortcoming, revealing that AI-generated noises were both noisier and less distinct. Consequently, player perception of this technological gap translated into decreased immersion and quality ratings, as corroborated by the user ratings. By the end of the qualitative comments, the particular problems that people had with the traditional soundscape had been described as "muffled," "un-synchronized," and "annoying." The success of the traditional soundscape, according to the participants, lay not only in the high quality of the individual sounds but also in the "layers" and "balance" that it had. Rather than creating a sonic environment, the present AI pipeline functions as a generator of assets. Unlike human designers, it doesn't take into account the big picture when planning the experience's pacing, emotional arc, and audio hierarchy.

2. Whether or not fine-tuning on a domain-specific dataset may substantially enhance the outcomes was the subject of the second study question. According to the data, its benefits are minimal and unpredictable. A minority of participants favoured the simulated fine-tuned condition over the pre-trained baseline, but it lagged behind the conventional design and had worse emotional involvement scores.

In order to grasp this subtlety, the qualitative data is crucial. Condition C made use of high-quality, stylistically-appropriate sounds thanks to the "Wizard of Oz" method. Nevertheless, additional issues arose due to the pipeline's inadequate handling of these noises, including the inability to appropriately adjust their loudness, timing, and relationship to other sounds. Even a well-selected sound effect can come off as more abrupt than one of lower quality but well executed if played at an inappropriate volume or at the wrong millisecond. Due to its ability to reframe the problem, this discovery is crucial. Developing smarter methods to integrate and blend those sounds into the gaming environment is the key to better artificial intelligence sound design, not merely better generative models. A collection of high-quality sounds is not enough, according to the research; the real skill is in figuring out how to put them to use.

3. Future of Game Audio Implications This research in no way downplays the possibilities of artificial intelligence in the realm of sound creation. Instead, it shows where it stands in relation to its potential and where it has to improve in order to be a more practical benchmark. Despite

its limitations, the LLaVA-to-AudioLDM pipeline was able to automate the conceptual connection between images and audio. It was the implementation, not the idea, that was flawed.

Rather than an AI-powered "sound designer in a box," the findings point to AI-powered tools that augment human designers as the most viable future direction. To avoid using the same sounds over and over again, an AI may create a large database of possible versions. It might be utilised to automate the arduous process of putting footsteps, freeing up the designer to concentrate on creating more dramatic and heroic noises. Although it was a critical methodological fix for this study, the "human-in-the-loop" approach may perhaps be the best workflow for curating LLaVA prompts. A human's depth, taste, and holistic vision are necessary to turn a collection of sounds into an engaging, immersive experience, whereas AI can handle the breadth of content development.

Discussion (Expanded Version) Chapter 5

5.1 Concise Overview of Major Results

With the help of cutting-edge AI models, this study aimed to test a new automated process that could create soundscapes for video games. A solid mixed-methods study yielded results that tell a coherent story. A large performance disparity between the AI pipeline's output and the soundscapes created using conventional, human-led approaches is revealed by the primary findings. From rigorous technical measures of clarity and fidelity to subjective player evaluations of immersion, coherence, and general quality, the Traditional soundscape (Condition B) was clearly better in every respect. Reproducing human artistry in a creative arena as intricate as sound design is incredibly challenging, as this ground-breaking discovery highlights.

Condition C, the Simulated Fine-tuned model, was the artificial intelligence condition that outperformed the baseline in a very little and inconsistent way. An already-trained model (Condition A). Even though the fine-tuned version used higher-quality source sounds, it nonetheless presented new problems with mixing and balance that participants found distracting, in contrast to the pre-trained model's severe criticisms of low fidelity and inconsistent synchronisation. This evidence points to the fact that the key obstacles for AI in this field are not limited to just creating assets, but rather to the more intricate and creatively difficult tasks of integrating them into a coherent whole.

5.2 Discussion of Results in Light of Original Research Questions

Chapter 1's primary research topics are directly and thoroughly addressed by the findings of this study. Taking a realistic view of the state of the art in generative AI, this analysis of the results shows both the promising areas and the major limitations of this technology when applied to the intricate and imaginative work of sound design.

5.2.1 Differentiating Between AI-Generated and Human-Crafted Audio: The Quality Gap

In the first study, researchers wanted to see how artificial intelligence (AI) soundscapes stacked up against more conventional, human-designed soundscapes in terms of effectiveness and quality.

The results are clear: the impact of the human-designed soundscape was far greater. That "quality gap" showed up in a few important places. The artificial intelligence-generated assets had lower SRMR scores, which proved that they were noisier, more distorted, and most importantly, less clear. The player's experience was greatly affected by this technical shortfall, which was not just a theoretical metric. According to the qualitative comments, the pre-trained AI audio had a "underwater" or "muffled" sound, which is a very accurate interpretation of the low SRMR statistic. This proves that players are able to detect these technological issues.

Based on these findings, it seems that the present AI pipeline is good at generating assets but not so great at creating comprehensive soundscapes. While it can respond to text with individual sounds, it doesn't have the creative sensibility or contextual understanding to make such sounds cohesive. Audio is not only placed; it is layered, balanced, timed down to the millisecond, and designed to contribute to a cohesive emotional arc. There was no indication that the AI could handle such complex tasks in its present form. This led to a situation that sounds like an audio "uncanny valley," where the noises were close to what was happening in the context but didn't quite seem right. A little out of sync or the wrong timbre can be more distracting than an entirely wrong sound because it highlights the artificiality of the sound and breaks the player's immersion. Because of issues with timing and texture quality, AI-generated soundscapes sometimes fell into this valley, giving participants a disjointed and unpolished experience.

Section 5.2.2: The Discreet Function of Modification

What is the impact of fine-tuning a generative audio model using a domain-specific dataset on the quality and contextual relevance of the generated soundscape? This was the second research question aimed at answering.

Though more subtle, the findings are just as illuminating in this case. Despite high hopes, the simulated fine-tuning procedure failed to deliver the desired dramatic improvement in performance. Only a minority of users favoured it beyond the pre-trained baseline; however, it lagged behind the conventional design and received worse marks for emotional involvement.

The problem is reframed, which is why this discovery is essential. In order to guarantee that the fine-tuned condition made use of high-quality sounds that were stylistically appropriate and a perfect match for the prompts, the "Wizard of Oz" methodology was employed. As a result, the assets themselves were of sufficient quality, and the problem was in the execution. This is corroborated by the qualitative data, as participants stated that the fine-tuned version included sounds that were "too loud," "distracting," or had "too much audio details." What this means is that it takes more than just a sound library of flawless sounds to make a fantastic soundscape. Because it lacks the human touch to apply the art of mixing, timing, and balancing, the automated pipeline is unable to create the desired result.

An awareness of psychoacoustics and attentive listening are essential components of the "art of the mix," a pivotal phase in sound design. The designer takes human factors into account by carving out frequency space to avoid sounds obscuring one another, controlling loudness through dynamic range compression, and establishing a sense of space through reverberation. They regularly and intuitively decide which noises should be front and centre to direct the player's focus and which should be in the background. Instead of considering each sound as an individual occurrence and arranging them in sequence, the AI pipeline ignored all of these capabilities. A lack of depth and concentration in the otherwise well-designed soundscape was the consequence of this chaotic and disorganised sound environment.

Link to Previous Works 5.3

Research on generative AI in the arts has been going on for some time, and this dissertation adds to and draws from that conversation. Both the possibilities lauded in recent research and the worries raised about the limitations of present models are validated by the outcomes.

The achievement of producing visually-driven audio that is appropriate to the given situation is in line with the encouraging outcomes demonstrated in research on models such as AudioLDM 2 and LLaVA [You can reference these studies in your literature review, for instance, Liu et al., 2024]. As a domain-specific, practical use of these technologies, this study verifies their basic capabilities for cross-modal generation. The results of this study offer a "snapshot in time" of the capabilities of these robust, general-purpose models in the context of a real-world process.

Still, research has shown that generative models struggle with things like long-term coherence, subtlety, and contextual knowledge, and this disparity in quality between AI-generated and human-designed soundscapes is consistent with those findings [You can cite relevant papers here]. A model can make what sounds like "footsteps on gravel" for three seconds, but it doesn't know how that clip should fit in with the other ten minutes of the game. Present models do not have this larger aesthetic and temporal awareness, according to this study's results on distracting ambience and poor coherence.

In addition, the study's finding that sound implementation and mixing are just as important as asset quality lends credence to established ideas in game audio design. Professionals and academics in the field have long stressed that a soundscape is not just the components but an orchestrated whole that serves to direct the player's focus and elicit feelings. The automated pipeline's inability to achieve the same level of orchestration shows how important human-centric design principles are, even with advanced generative tools.

5.4 Game Audio Implications from the Study

What this study finds about the future of game development and the sound designer's position is really practical and important. Based on the findings, it seems that the story of AI displacing human creativity is unrealistic and started too soon. An alternative, more practical and fruitful view is that of AI as a potent instrument that supplements, not replaces, human creativity.

It appears from this research that the concept of a "sound designer in a box" that can take a game build as input and produce a whole soundscape is still somewhat far off. The problems with quality, consistency, and mixing that the pipeline has shown show how important it is to have human designers with holistic, intuitive, and aesthetic judgements.

A "human-in-the-loop" workflow may be in the cards for the future, notwithstanding the pipeline's limited results so far. Full automation isn't the best use of this technology right now; rather, it's for the reduction of boring, time-consuming jobs. As an illustration:

The usage of a model such as AudioLDM 2 would allow a designer to easily solve the issue of repetitiveness by creating dozens of variants of a single footstep or weapon sound.

The designer would thus have more time to concentrate on the more dramatic, "heroic" sounds that make up the player's experience, rather than the tedious job of placing and timing thousands of innocuous noises (such as footfall and vegetation rustles).

To facilitate rapid prototyping, this pipeline could be utilised to rapidly create a "first pass" soundscape for a prototype. This would give the development team an idea of the game's audio identity prior to typically hiring a sound designer. The early phases of a project sometimes do not have the funding for a specialised sound designer, therefore this is especially helpful for independent developers and smaller firms.

Beyond that, this study paves the way for AI to be used in procedural and dynamic sound design. The capabilities of such a pipeline for real-time production are very remarkable, even though this study only utilised a linear video clip. Game engines with AI might create custom sound effects in real time for procedural events like physics-based building collapses, which don't have any pre-existing animations. In contrast to static, pre-authored audio elements, this would enable a fully dynamic and infinitely changing soundscape.

With this AI model as a creative collaborator, the sound designer is given more agency, free from mundane tasks, and given more creative freedom, all while maintaining complete creative control over the end product.

5.5 Restriction of Study

For a fair assessment of the findings and to direct future research, it is necessary to note the study's limitations.

Due to the small sample size, the statistical power of the quantitative findings is limited, despite the rich and insightful qualitative data collected from 11 individuals. Though the seen patterns are significant indicators, it would be statistically risky to extrapolate them to the whole gaming population. Additionally, all of the people that made up the demographic were avid gamers. Whose perspectives have not been investigated are those of non-gamers and casual gamers, who might have diverse expectations and degrees of critical listening.

Prototype Game: A single, simple 2D platformer was used to conduct the experiment. The results might not be directly transferable to different types of games, but this was a conscious decision made to avoid bias. Consider a competitive first-person shooter: the AI pipeline had a hard time keeping up with the gameplay requirements of extremely clear audio and accurate spatialization. An artistic level well beyond the system's shown capabilities is required for narrative-heavy horror games, which rely on nuanced atmospheric sounds and strategically timed jump scares.

For reasons of practicality, we had to mimic the fine-tuned state by employing a "Wizard of Oz" approach. However, this method does not mimic the actions of a fully tuned generative model, even though it was successful in testing the idea of using domain-specific sounds. In reality, a fine-tuned model may either bring new artefacts or have problems like "mode collapse," in which it becomes overfit to the training data and can only generate a small subset of audible tones. It is important to note that the results for this condition should be seen as a hypothetical ideal situation for asset selection, rather than a direct assessment of the generative output of a fine-tuned model.

Recommendations for Future Studies 5.6

Naturally, numerous interesting directions for future research can be inferred from the results and caveats of this dissertation.

First and foremost, we need to find a way to fine-tune AudioLDM 2 so that we can do a replication of this study. An improved evaluation of domain adaptation's merits would result from this, as would knowledge of the output's unique traits and any artefacts it may produce.

Improving AI for Audio Implementation and Mixing: Moving forward, studies ought to centre on developing AI systems that can intelligently add audio, rather than only making audio assets. Some possible areas of investigation here are AI-powered mixing agents with the ability to operate DAWs or audio middleware for game engines (such as FMOD or Wwise). This kind of agent might be programmed to make mixing decisions according to broad objectives like "make this scene feel more tense" or "ensure the dialogue is clear above the explosion."

Using a similar AI pipeline to analyse other types of games might be a valuable exercise. An in-depth analysis of the system's strengths and drawbacks could be achieved by comparing its performance in first-person shooters with more chaotic and dense soundscapes and in walking simulators with more subtle and atmospheric music.

Embedding this process inside a game engine, such as Unity or Unreal Engine, enabling dynamic generation is an important next step beyond linear video. Among the most intriguing potential uses of this technology in the future, real-time, dynamic sound production for procedural events, might be explored with this capability.

Additional Research with More Participants: To confirm the quantitative results of this study and to investigate the impact of age, culture, and gaming experience on the perception of AI-generated sound, additional research with a bigger and more varied sample of participants would be helpful.

Chapter 6, Conclusion

Section 6.1 Restatement of the Research ProblemA foundational aspect of contemporary video game production is the creation of evocative and emotionally engaging soundscapes. Though they get the job done, the conventional approaches to sound design are extremely time-consuming, resource-intensive, and expert-level labour-intensive. The ever-expanding and ever-changing worlds of modern games are too much for this method to handle. The fast development of generative AI, on the other hand, holds revolutionary promise as it introduces a fresh approach to content production. The necessity to investigate the feasibility of a completely automated AI pipeline for creating video game sound in a rigorous and practical manner inspired this dissertation, which aims to move beyond theoretical hype in this area. At issue was whether or not the most cutting-edge models could, in the opinion of seasoned players, provide sound that satisfies the rigorous technical and aesthetic requirements of a classically constructed soundscape.

6.2 Findings and Methods Synopsis

A new two-stage generating pipeline was developed, implemented, and assessed in this work to tackle this issue. The procedure started with silent gaming footage being analysed by a vision-language model called LLaVA. The model then generated a time-stamped sequence of descriptive text prompts. After that, AudioLDM 2, a text-to-audio latent diffusion model, was used to generate the matching ambiences and sound effects. One version of the game prototype had a conventionally designed soundscape (Condition B), another had audio from the baseline pre-trained AI (Condition A), and a third had audio representing a virtual fine-tuned AI (Condition C). Eleven participants evaluated all three versions in a comparative mixed-methods experiment. Both quantitative measures of audio quality and qualitative comments from users on immersion, coherence, engagement, and overall quality lent credence to the assessment.

The study's conclusions were clear and supported by the various types of evidence. There was a glaring "quality gap" between the AI-generated and human-designed audio. In addition to being demonstrably more technically accurate, the classic soundscape also received unanimously better ratings across all subjective dimensions. Participants commonly described the output of the pre-trained AI soundscape as "muffled," "un-synchronized," and of "low quality," which was a direct correlation with its poor objective clarity ratings and rendered it the worst performer. Although conceptually superior, the simulated fine-tuned model provided inconsistent and small benefits. Despite having access to high-quality, domain-specific sounds, the results demonstrated that the AI pipeline's lack of skill in blending and implementing them introduced additional problems that were equally harmful to the user experience as the baseline model's poor fidelity.

6.3 The Research's Contribution and Its Implications

Specifically, this dissertation contributes to the field of game design by laying out an evidence-based standard for a full visual-to-audio generative pipeline. This paper gives a realistic and grounded evaluation of the capacity of the technology by carefully recording the accomplishments and, more importantly, the shortcomings of this strategy. It shifts the focus of generative models' theoretical promise to their practical use in a challenging creative context.

For the industry's trajectory in the gaming industry, these results have major ramifications. There is no immediate threat to the human sound designer's job security, according to the research. The best way ahead, instead, is for people to work together and use AI as a tool to help out, not as a standalone creator. According to the research, human intuition, contextual understanding, and taste are still necessary for the holistic orchestration of sound assets, which is the real art form of sound design. The research indicates that in the future, designers would use AI to automate repetitive processes like putting footsteps, develop variations of assets quickly, and prototype concepts while keeping complete creative control over the finished soundscape. Without compromising on aesthetic quality, this "human-in-the-loop" paradigm claims to increase productivity and innovation.

6.4 A Last Word on the Matter

Starting with an inquiry into automation, this thesis ends by reiterating the importance of teamwork. A fully autonomous AI "sound designer in a box" is still a ways off, but our research has shown us a more exciting and immediate reality: AI is becoming a strong collaborator in the creative process. Despite its shortcomings, the LLaVA-to-AudioLDM 2 pipeline is an impressive technical achievement that brings us one step closer to a world where innovative ideas encounter far less resistance when put into action. Instead than trying to supplant artists, we should focus on developing better tools that can grasp their intentions. In fact, the sound designer's work will be enhanced by the combination of human vision and artificial intelligence, allowing them to create tomorrow's worlds that are richer, more dynamic, and more immersive.

References

[1] J. Grasso, "Playing with Sound: A Theory of Interacting with Sound and Music in Video Games," *Society for American Music Bulletin*, vol. 43, no. 1, pp. 11-13, 2017.

[2] K. Collins, *Game Sound: An Introduction to the History, Theory, and Practice of Video Game Music and Sound Design*. Cambridge, MA: MIT Press, 2008.

[3] I. Davis, "Game Sound: An Introduction to the History, Theory, and Practice of Video Game Music and Sound Design," *Fontes Artis Musicae*, vol. 57, no. 2, pp. 226-227, 2010.

[4] K. Collins, *Playing with Sound: A Theory of Interacting with Sound and Music in Video Games*. Cambridge, MA: MIT Press, 2013.

[5] J. Janer, E. Gomez, A. Martorell, M. Miron, and B. de Wit, "Immersive Orchestras: Audio Processing for Orchestral Music VR Content," in *2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, Barcelona, Spain, 2016, pp. 1-2.

[6] A. G. Privitera, F. Fontana, and M. Geronazzo, "The role of audio in immersive storytelling: A systematic review in cultural heritage," *Multimedia Tools and Applications*, 2024.

[7] F. Menzer, "Preliminary Study on Integrating 3D Audio with 2D Game Engines for Immersive Storytelling," in *2024 2nd International Conference on Sustaining Heritage: Embracing Technological Advancements (ICSH)*, 2024, pp. 46-50.

[8] M. Hsu and M. Cheng, "Immersion experiences and behavioural patterns in game-based learning," *British Journal of Educational Technology*, vol. 52, no. 5, pp. 1981-1999, 2021.

[9] Z. Zhang, J. Zhang, X. Zhang, and W. Mai, "A comprehensive overview of Generative AI (GAI): Technologies, applications, and challenges," *Neurocomputing*, vol. 632, p. 129645, 2025.

[10] A. Einbond et al., "Embodying Spatial Sound Synthesis with AI in Two Compositions for Instruments and 3-D Electronics," *Computer Music Journal*, vol. 46, no. 4, pp. 43-61, 2022.

[11] B. Kapralos, "An Overview of a Course on Audio for Games and Interactive Media," in *2025 25th International Conference on Digital Signal Processing (DSP)*, 2025.

[12] H. Liu et al., "AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871-2883, 2024.

[13] Z. Chen et al., "Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 1753-1808, 2024.

[14] A. Hufschmitt, S. Cardon, J. Borelt, F. Wolf, and M. Addoum, "Adaptive Audio to Player Actions and Gameplay: a New Video Game," in *2024 IEEE Conference on Games (CoG)*, 2024.

[15] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26296-26306.

[16] C. Li et al., "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[17] D. Dai et al., "PA-LLaVA: A Large Language-Vision Assistant for Human Pathology Image Understanding," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2024.

[18] H. Elgendy and H. Cholakkal, "TX-LLaVA: Large Language and Vision Assistant for Temporal Changes in Chest X-Rays," in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 2025.

[19] T. Yıldırım, M. F. Amasyali, and A. C. Karaca, "LLaVa-RS: A Unified Model for Image and Change Captioning in Remote Sensing," in *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, 2025.

[20] K. Sharma et al., "LLaVA-PlantDiag: Integrating Large-scale Vision-Language Abilities for Conversational Plant Pathology Diagnosis," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024.

[21] D. Guan, "LLaVA-Oil Painting Appreciation: A Vision-Language Model for Enhancing Understanding of Oil Paintings Through AI-Driven Analysis and Conversation," in *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 2024.

[22] H. Wang and S. Benslimane, "Grounding Image Understanding to Oil and Gas Product Manuals: Refining LLaVA through Contextual Instruction Tuning," in *2025 Conference on Artificial Intelligence x Multimedia (AIxMM)*, 2025.

[23] Z. Ding and Q. Yang, "Enhance Image-to-Image Generation with LLaVA-generated Prompts," in *2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, 2024.

[24] J. Vora, A. Krishnan, N. Bouacida, and P. R. Shankar, "PTQ4ADM: Post-Training Quantization for Efficient Text Conditional Audio Diffusion Models," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[25] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-Augmented Text-to-Audio Generation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[26] T. Karchkhadze et al., "Latent CLAP Loss for Better Foley Sound Synthesis," in *EUSIPCO 2024*, 2024, pp. 351-355.

[27] J. Ren et al., "Generative Semantic Communication: Architectures, Technologies, and Applications," *Engineering*, 2025.

[28] *IEEE Standard for Game Voice Enhancement of Mobile Gaming*, IEEE Std 2861.4-2023, 2023.

[29] I. Khan, T. V. Nguyen, M. C. Gursesli, and R. Thawonmas, "Sonic Doom: Enhanced Sound Design and Accessibility in a First-Person Shooter Game," in *2025 IEEE Conference on Games (CoG)*, 2025.

[30] M. Beig, B. Kapralos, K. Collins, and P. Mirza-Babei, "An introduction to spatial sound rendering in virtual environments and games," *Computer Games Journal*, vol. 8, no. 3-4, pp. 199-214, 2019.

[31] M. S. Seyfioglu et al., "Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[32] K. Chen et al., "MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[33] Q. Xu et al., "M3A: A multimodal misinformation dataset for media authenticity analysis," *Computer Vision and Image Understanding*, vol. 249, p. 104205, 2024.

[34] Y. Zhuang et al., "From hearing to seeing: Linking auditory and visual place perceptions with soundscape-to-image generative artificial intelligence," *Computers, Environment and Urban Systems*, vol. 110, p. 102122, 2024.

[35] Y. Guo, "LLaVA-NeXT-Med: Medical Multimodal Large Language Model," in *2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC)*, 2025.

[36] H. Yamanishi, L. Xiao, and T. Yamasaki, "LLaVA-Tour: A Large Multimodal Model for Japanese Tourist Spot Prediction and Review Generation," in *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2024.

[37] Y. Yuan et al., "Leveraging Pre-trained AudioLDM for Sound Generation: A Benchmark Study," 2023.

[38] H. Liu et al., "Audiosr: Versatile Audio Super-Resolution at Scale," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1076-1080.

[39] Y. Liang and M. Li, "Vivid Background Audio Generation based on Large Language Models and AudioLDM," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024.

[40] R. Zhang and Y. Zhao, "Cluster-LLaVA: Mixture of Experts Multimodal LLMs Can Also Be Small," in *2025 10th International Conference on Cloud Computing and Big Data Analytics (ICCCBD)*, 2025.

[41] Z. Han and J. Hao, "Blind-Assisted Question Answering Optimization Model Based on LLaVA," in *2025 6th International Conference on Computer Engineering and Application (ICCEA)*, 2025.

[42] Y. Jiang et al., "Expressive Text-to-Speech with Contextual Background for ICAGC 2024," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024.