Benchmarking the User Experience of Different AI Talking Head Generation

Samaneh Taghizadeh

Thesis submitted for the degree of
Master in Applied Computer and Information Technology (ACIT)
30 credits

Department of Computer Science
Faculty of Technology, Art and Design

OSLO METROPOLITAN UNIVERSITY

Autumn 2023

Benchmarking the User Experience of Different AI Talking Head Generation

Samaneh Taghizadeh

Benchmarking the User Experience of Different AI Talking Head Generation

http://www.oslomet.no/

Printed: Oslo Metropolitan University

Abstract

Interviews of abused and maltreated children are an important and often the primary source of information in court trials, but the children's testimonies are too often invalidated because they are collected in an unreliable and incorrect way. Thus, performing such interviews must follow strict protocols, stating the questions in an open and non-leading way, i.e., it is a sensitive and challenging task for professional practitioners. While utilizing real cases of child abuse for training therapists might be harmful and sophisticated, effective practice and training is essential to improve their skills. With advancements in machine learning and generative tools in recent years, their great potential can be exploited for educational purposes. We are researching an AI-based avatar to mimic the behavior of maltreated children to be used for interview training in such scenarios, here with a focus on the visual part of the avatar. In this respect, a talking-head generation is an AI-based tool that can be utilized for training professionals to deal with child abuse cases. By using such tools, challenging and sensitive situations can be simulated. In this study, we will examine the potential of different generative models for talking head avatars in the context of child abuse. We created different videos using current state-of-the-art models such as MakeItTalk, First Order motion, and Talking-Face PC-AVS with different styles. These styles include Cartoony, Painted, and Original. By running a user study, the effectiveness in terms of realism and preferences has been examined. The results show that the Talking-Face PC-AVS model creates better user experiences. Furthermore, according to the user study, we did not observe significant differences among the different styles of the videos. To improve the talking head avatars in child abuse applications, choosing the appropriate model should be considered as the first priority. We expect that more sophisticated models regardless of the style, can contribute to educating experts for the child abuse interviews remarkably.

Acknowledgments

I would like to express my deepest appreciation to my main supervisor, Pål Halvorsen, and Saeed Shafiei Sabet and Pegah Salehi as my co-supervisors for their guidance and support during the entire course of this thesis.

Also, I am grateful to the whole family of SimulaMet, for giving me the opportunity to participate in the talking-head avatars project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Child abuse is a serious problem that has far-reaching consequences for the physical, emotional, and psychological well-being of children. According to the World Health Organization (WHO)[1], child abuse refers to all forms of physical or emotional maltreatment, sexual abuse, neglect or negligent treatment, and exploitation that result in actual or potential harm to the child's health, development or dignity (World Health Organization, 2021). Child abuse can occur in any setting, including the home, school, and community. It is estimated that millions of children worldwide are affected by abuse each year, with many cases going unreported.

In society, it is crucial to prioritize the well-being of children and ensure they receive the greatest care, to prevent any form of mistreatment. By taking proactive measures to protect children, we can avoid the potential fundamental problems that may arise as consequences. This perspective emphasizes the significance of creating a safe and nurturing environment for children, highlighting the importance of collective efforts to safeguard their rights and promote their welfare.

According to WHO, child maltreatment results in the annual death of at least 850 children under the age of 15. Approximately 71% of homicide deaths occur in low-income and middle-income countries, where rates are 2.4 times higher compared to high-income states. Boys make up 60% of the victims in these cases. Homicide rates are higher among children under 4 years old compared to older children aged 5-9 and 10-14. Eastern Europe experienced a peak in child homicide rates during the economic and political transition, and while the rates have declined, they remain higher in this region (Sethi et al., 2013).

[1] https://www.who.int/

The effects of child abuse are devastating and can last a lifetime. Children who experience abuse are more likely to suffer from a range of emotional, behavioral, and physical problems. They may experience anxiety, depression, post-traumatic stress disorder (PTSD), and other mental health disorders. Physically, children who experience abuse may suffer from injuries, poor growth, and developmental delays. In addition, they may be at a higher risk of developing chronic health conditions such as heart disease, diabetes, and obesity (Norman et al., 2012).

The negative effects of child abuse are not limited to the individual child. They can also impact the child's family, friends, and community. For instance, child abuse can lead to increased healthcare costs, decreased work productivity, and increased crime rates. As a result, it is important for healthcare providers, psychologists, and social workers to be aware of the effects of child abuse and to provide appropriate care and support to affected children.

The prevention of child abuse is a complex and multifaceted issue that requires the involvement of multiple sectors, including healthcare, education, law enforcement, and social services. Prevention efforts should focus on addressing the underlying causes of abuse, such as poverty, social inequality, and family dysfunction. In addition, it is important to provide support and resources to families and caregivers to help them provide a safe and nurturing environment for their children (Lambie, 2005).

So far, we can state that child abuse is a serious problem that has far-reaching consequences for the physical, emotional, and psychological well-being of children. There should be serious plans for solving this important problem. Healthcare providers, psychologists, and social workers have an important role to play in preventing and addressing the effects of child abuse. By working together, we can help to ensure that all children have the opportunity to live healthy and fulfilling lives. In this case, it is crucial to communicate with the victims and inquire about the details of the abuse. Professionals need to ask different questions to find out various aspects of that misbehavior and gather all facts in this regard. Being aware of these details can assist society in finding effective solutions in this regard.

Given the sensitivity and importance of the topic of child abuse, it is vital to consider the challenges associated with working with children, particularly those who have been affected by abuse. Such children may be more sensitive than their peers. One of the critical steps in addressing child abuse is the ability to communicate effectively with the affected children and conduct interviews to gather details about the maltreatment. This procedure holds significant importance for the healthcare system's enhancement of more effective resolutions to this issue. However, prevailing literature underscores diverse factors that create hindrances for children when it

comes to openly sharing their abuse encounters with professionals. Thus, it becomes vital for professionals involved in handling child abuse cases to comprehend the essential needs of children and skillfully attend to those demands during the disclosure phase (Ettinger, 2022). Nevertheless, conducting interviews with child abuse victims presents distinct challenges. These youngsters often encounter difficulties in placing trust in others and might have undergone trauma inflicted by close family members, thereby adding complexity to the interview process.

To effectively interview this population, professionals need extensive knowledge and expertise, along with significant practical experience. The ability to behave appropriately and establish trust with child abuse victims comes from repeated practice, honing their skills over time. Therefore, both theoretical knowledge and practical experience are essential in improving the quality of interviews with child abuse victims. underscoring the need for professionals who work with them to be well-prepared.

To this end, extensive training and practice are necessary before professionals begin working with children. This is a standard process in many other professions, such as piloting. Pilots do not fly airplanes during their initial training but instead progress through various levels and simulations before being granted permission to fly. In the child abuse topic, simulation is extensively different. Professionals working in the field of child abuse face significant challenges when it comes to practice. Each child is unique and has their own set of rights that must be protected, and even the slightest misstep or inexperience when interacting with a child can have severe and lasting consequences for their well-being.

It is therefore imperative to develop a platform that enables professionals to practice their skills without harming any children. The development of such a platform is an important need of society, given the critical importance of effective interventions for child abuse victims.

Another use of these types of platforms is to conduct interviews with children by themselves, provided that they can interact with them in real-time and ask questions related to the child's previous answers. In fact, this has lots of benefits. Some of the researchers found that children reported feeling more comfortable and less anxious when interviewed by a virtual agent compared to a human interviewer and could reveal their secrets easier. The study also found that the virtual agent was perceived as more empathetic and supportive (Hamzelou, 2017). Therefore, for both of these purposes and even more tasks, the presence of intelligent avatars appears to be essential. One important consideration is to ensure that the interface is trustworthy and friendly so that children can communicate with it easily. It can be helpful for

trainers to interact with the interface as a child as well.

In conclusion, it is clear that there is a need for tools to train professionals who work with child abuse victims, including social workers and law enforcement officers. Virtual agents provide a valuable means for professionals to practice interviewing techniques and improve their communication skills with child abuse victims in a safe and controlled environment. Actually, studies have shown that computer-based interactive learning can further improve investigative interviewing skills (Powell et al., 2016).

The collaborative research project undertaken by Department of Holistic Systems in SimulaMet[2], in conjunction with Faculty of Social Sciences in OsloMet, presents a cutting-edge endeavor aimed at transforming the landscape of child welfare and law enforcement practices (Baugerud et al., 2021). By pioneering a novel approach to conducting interviews with maltreated children, this project merges the realms of computer science and social sciences to yield groundbreaking outcomes. At its core, the project seeks to elevate the quality of investigative interviews through the creation of interactive, lifelike child avatars. Drawing upon the synergies of artificial intelligence, computer vision, and natural language processing, these avatars are poised to become pivotal components of a comprehensive interview-training program. The project's focus on empirical training techniques, informed by the analysis of past investigative interviews, aims to equip professionals with advanced skills, enabling them to navigate the intricate task of conversing effectively with maltreated children (Hassan, Salehi, Riegler et al., 2022; Hassan, Salehi, Røed et al., 2022).

This project not only tackles a vital societal challenge but also highlights the potential of technology-driven solutions to enhance the well-being of vulnerable individuals. By bridging the divide between technology and social welfare, it offers an inspiring paradigm that aligns with the broader endeavor of leveraging technological innovations to address urgent social concerns. The project's interdisciplinary essence, integrating the technical capabilities of computer science with the social sciences insights into human behavior, showcases the potency of collaboration in crafting meaningful solutions. The project's focus on augmenting communication and investigative skills seamlessly resonates with the broader goal of utilizing technology for positive societal transformation (Salehi, Hassan, Lammerse et al., 2022a).

[2] https://www.simulamet.no/simulamet-projects

1.2 Problem Statement

As previously mentioned, adequate training for professionals and psychologists in conducting interviews with children is of paramount importance in their work with young individuals. To create an effective platform for this purpose, practical tools are necessary. Fortunately, there are various technological tools that can be used to create such systems. Talking head models can serve as an excellent resource in this regard.

Talking head models, also known as virtual agents, are computer-generated characters that simulate human-like conversation and facial expressions (Htike, 2017). These models have been increasingly used in various applications, including child abuse projects. During these interviews, professionals have the opportunity to practice their interviewing skills using talking head avatars. This approach ensures that trainers can thoroughly train without any concerns about potentially harming children while asking necessary questions. Additionally, the virtual agent can be programmed to provide information and support to the child, such as resources for counseling or legal assistance (Salehi, Hassan, Lammerse et al., 2022b).

There are several talking head models (e.g MakeItTalk[3](Y. Zhou et al., 2020), PC-AVS model [4](H. Zhou et al., 2021), FOMM model [5](Siarohin et al., 2019b), OneShot[6](T.-C. Wang et al., 2021), FACIAL [7](C. Zhang et al., 2021b) and ...) used to create avatars, each with its own strengths and weaknesses. Some of these models are older and represent the first generation of this technology, generating avatars using only a single picture and audio. Although these models can be relatively simple to use, some of them are more complex and require a large number of inputs, resulting in more realistic avatars.

In addition to models, the choice of materials used to create avatars can also significantly impact their overall appearance. There are various options available, including using real photographs and recordings to create an avatar, opting for cartoony or stylized images, or using images with filters that give them a painted look. Similarly, there is the choice between using human-made sounds or robotic-generated ones for the audio component of the avatar. The inquiries investigated in this study revolved around users' interactions with talking head avatars, exploring their experiences during communication and their emotional responses to witnessing the avatar's speech. The study delved into aspects such as users' preferences, and their perceptions of the avatar's realism. This study examined features related to both the

[3]https://github.com/yzhou359/MakeItTalk
[4]https://github.com/Hangz-nju-cuhk/Talking-Face_PC-AVS
[5]https://github.com/AliaksandrSiarohin/first-order-model/tree/master
[6]https://github.com/zhanglonghao1992/One-Shot_Free-View_Neural_Talking_Head_Synthesis
[7]https://github.com/zhangchenxu528/FACIAL

visual aspects of avatars and their realism while speaking. From the models presented earlier, a subset was chosen for a user study aimed at addressing the following research inquiries:

- **RQ1:** Which models can generate avatars that generally give a better feeling to the audience?
- **RQ2:** What distinctions emerge in user experiences when interacting with real, cartoony, and painted avatars? How do these differences influence viewers' perceptions and emotional responses toward these avatars?

We believed that addressing these questions could help fill a minor gap within our scope of work concerning generated talking head avatars. The findings from this study have the potential to guide developers in refining various aspects of this field, particularly in areas aligned with our study's focus, such as exploring different styles and models.

1.3 Scope and Limitations

For answering those research questions, some models have been listed, and an attempt was made to select the ones that had an important role in the talking head models field. After creating a short list, the goal was to run those models and generate different avatars. All models were required to use the same picture and audio in order to produce avatars with the same features but different qualities due to their diverse algorithms. The key aspect to highlight here is the critical importance of our focus. Our objective revolved around comprehending users' experiences with avatars created by these models, rather than benchmarking their performance based on various metrics. Thus, our scope was intentionally narrowed down to achieve this specific goal.

The study's focus was on addressing the issue of child abuse, prompting the use of avatars featuring children's images. Initially, the text concerning child abuse was selected, but considering its potential impact on participants' emotions, the concept shifted to something general.

To ensure a comprehensive analysis, the investigation aimed to encompass outcomes for both boys and girls, resulting in the selection of two distinct characters. We decided to explore three distinct styles: a real image, a cartoonish style with large eyes, and a painted look with soft colors. This choice was made after considering various options, such as caricatured designs and diverse painting techniques. We also created videos using those other styles, but significant differences were not observed. While these other styles seemed intriguing, we believed that focusing on these three specific styles

would offer a more pronounced contrast in participants' emotional responses. This approach allowed us to gain valuable insights into how participants perceived each style and how their reactions differed among the chosen avatars.

The applicability of the results and conclusions from this study might not extend to other domains within this field. The findings within our narrow scope provide the basis for our reliance. One of the study's limitations arises from this constraint. For instance, results could differ if avatars featured adult individuals or if alternative models with superior performance in distinct aspects were chosen. We opted to minimize the number of videos in the study to expedite participant responses in the questionnaires, a choice that stemmed from the desire for efficient engagement. However, it's important to note that this decision also posed a limitation to our study. This pragmatic approach allowed us to streamline the study process, enabling participants to share their insights comfortably within a reasonable time frame.

1.4 Research Methods

A research method is a structured approach used to collect, analyze, or experiment with data in order to address research questions or test hypotheses and achieve a specific research objective. These methods can be qualitative, quantitative, or a combination of both, depending on the type of data and analysis techniques employed. In this study, our primary objective was to identify optimal models for the execution and generation of videos. Our exploration focused on utilizing two distinct characters across three different styles. This comprehensive approach resulted in the generation of a total of 18 videos, each spanning approximately 10 seconds. We included an anchor video to establish a baseline. To ascertain the effectiveness of our approach, we conducted a user study, framing specific questions designed to align with our research objectives. This step was pivotal in collecting the necessary insights to address our research inquiries effectively. Upon the completion of the user study, our attention shifted to the analysis phase. The data amassed from the user study was meticulously processed and organized, leveraging PowerBI for creating informative box plots. Moreover, the analytical journey delved deeper as we employed SPSS software to conduct a repeated measures ANOVA analysis (Girden, 1992). The results of this multifaceted analysis, to be expounded upon in subsequent sections, provided us with invaluable insights into the scope of our objectives and the trajectory of our work. This study's unique blend of empirical exploration, user engagement, and advanced statistical analysis serves as the foundation for substantiating our research methodologies and yielding profound insights into the optimization of video generation models.

1.5 Ethical Considerations

In our context, ethical issues have various aspects. As artificial intelligence (AI) continues to play an increasingly significant role in various fields, the ethical implications of its use are becoming more crucial. Deepfakes, a highly sensitive issue, possess the potential for both malicious and socially harmful purposes. Consequently, steps are being taken to create ethical guidelines and regulations that prohibit the unethical use of this technology (Meskys et al., 2020). Talking heads are a similar technology to deepfakes, with some differences, but they present similar ethical challenges. Like deepfakes, talking heads also can be used to create fake news or defame individuals. This can have serious consequences, such as damage to reputations or even public safety.

Another ethical consideration in machine learning systems is the importance of promoting fairness and accountability(Veale & Binns, 2017). There is a possibility of bias and discrimination in the data used for training machine learning models. In case the training data is biased, it can result in the generation of videos that are also biased, which can perpetuate unfair stereotypes and discriminatory practices.

Furthermore, it is essential to ensure transparency and obtain informed consent when creating and utilizing these models. Individuals whose images are used should be fully informed about the usage and have the option to withdraw their consent. In our specific case of creating avatars using children's pictures, videos, and voices, this issue becomes even more sensitive, and we must seek their permission at every step. It is crucial to prioritize and address this matter appropriately. To address this issue in the study, a GAN-generated image was utilized to create one of the avatars. The image was obtained using a website that generates pictures of individuals who do not actually exist ('This Person Does Not Exist', n.d.). For the female avatar, an image previously used in another study was utilized, and permission was obtained for its use. The audio utilized in the study was generated from robotic sound (Labs, 2023).

Another aspect of ethical considerations in this study was related to questionnaires. User studies needed to be conducted to evaluate the avatars created by different models. Therefore, a data collection process was undertaken. All questionnaires were conducted anonymously during the study to ensure that participants were certain that their ideas would remain anonymous.

1.6 Main Contributions

The core inquiries of this study were geared towards gaining a comprehensive understanding of participants' perceptions and satisfaction, with a focus on assessing their overall experiences, the realism of speech and movements, and their comfort levels during interactions with avatars. In the later phase, we aimed to discern participants' preferences among avatars with cartoony, painted, or real picture styles. It's noteworthy that the study also involved comparing the outcomes of different talking head models using consistent audio, images, and videos.

To achieve our research objectives, a user study was conducted where participants first watched videos made using a model, showing one character in a certain style. After that, they answered six questions that ranged from bad to excellent. Participants watched 19 videos in total, but the first one was excluded from our analysis since it served as an anchor. The user study was carried out using Microsoft Forms and was distributed to individuals through various platforms to collect responses.

The main focus was to understand how comfortable participants felt when they used these avatars. Additionally, we asked more detailed questions about how well the avatars' lip movements matched and how natural their head movements looked. We also wanted to know how good the avatars were overall. We looked at different kinds of avatars, like boys and girls with cartoony, painted, and real pictures. We compared and studied them closely.

In the upcoming paragraphs, a closer look will be taken at the research questions and provide a simple breakdown of the findings we have uncovered in response to each one.

- *"Which models can create avatars that make people feel better?"*.

This question was initially raised in the problem statement section. Out of the three models which are tested, the one that lets us control the avatar's poses stood out. The features considered in the judgement of these models based on six important things: how good the whole experience felt, how the avatar's head moved, whether its lip movements matched the speech, how well it talked, the video quality, and how comfortable people were with it. Interestingly, the model with pose control scored better in all six aspects.

- *"What distinctions emerge in user experiences when interacting with real, cartoony, and painted avatars? How do these differences influence viewers' perceptions and emotional responses toward these avatars?"*.

Throughout the user study, a notable observation emerged, indicating that the

various styles employed did not exert a significantly influential effect on the outcomes. Notably, among all the features assessed, videos adopting the Original style predominantly claimed the top positions. There were instances where participants displayed a comparable preference for the cartoony style, although this distinction did not yield a noteworthy variance in the overall assessment.

1.7 Thesis Outline

This thesis is structured across six comprehensive chapters. The initial two chapters establish the context and essential background, while the following trio of chapters delve even deeper into the study's core. As we reach the concluding chapter, we encounter insightful summaries, conclusions, and a glimpse into potential avenues for future research. To achieve a more comprehensive understanding, we will now delve into a detailed breakdown of the content within each chapter.

- Chapter 1: **Introduction**

Crucial information has been incorporated into the initial chapter. The motivation behind the whole study, including the significance of the child abuse topic and essential statistics, has been explained. Furthermore, the problem statement and all the limitations encountered during this study have been detailed. In this context, ethical concerns, which encompass various aspects briefly discussed in this chapter as well, hold a significant place.

- Chapter 2: **Background and Related Works**

In this chapter, various aspects of related research are explored, as evident from its title. The initial section is devoted to prior studies on child abuse, while the subsequent parts delve into technical facets such as machine learning methods and the creation of talking head avatars. The discussion also covers topics like deepfakes and face-swapping. The uncanny valley concept is also discussed in this chapter.

- chapter 3: **Research Methodology**

The Research Methodology chapter delves into the models utilized for video production in the user study. An elucidation of the assorted tools employed across the study's duration is presented in the initial segment of this chapter. Additionally, a discussion is offered regarding the selection rationale for these models from an array of alternatives, as outlined at the outset. A summary table encapsulating the chosen models is also provided, offering a concise overview of their key attributes.

- chapter 4: **Experiment**

In the Experiment chapter, we delve into the details of all the materials we used in the user study. This includes the information we collected from the participants, as well as their background details. We thoroughly examine the questionnaires that were given to the participants during the study. Following that, we illustrate the process of cleaning and preparing the data to obtain the final results and visual representations. This chapter outlines the systematic approach we followed to carry out the study.

- chapter 5: **Outcome and Findings**

This chapter showcases numerous plots and elucidates the data derived from participants. The results are categorized into three sections: the main effect of the model, the main effect of the character, and the main effects of the style. The latter part of the chapter delves deeper into the second segment of the questionnaire, which centers around the uncanny valley concept. Detailed insights obtained from the questionnaires are presented.

- chapter 6: **Discussion and Limitations**

This chapter focuses on a comprehensive discussion of various findings and their alignment with our initial hypotheses. One prominent observation was the disparity between our anticipation of the significant importance of the uncanny valley concept and the lack of confirmation from our user study results. Furthermore, we delve into the challenges encountered while executing models and generating videos. The essence of this chapter lies in the comparative analysis of our results with our own assumptions.

- chapter 7: **Conclusions and Future works**

This chapter presents a comprehensive overview of the entire study along with the conclusions drawn from the analysis of the conducted user study. While we initially intended to conduct a second phase of the user study, various limitations hindered its execution. As part of future work, we propose exploring the extensive potential sub-topics within this domain, considering the scope for further investigation and development.

Chapter 2

Background and Related Works

In order to cover the interdisciplinary nature of the topic, the related work chapter will survey different areas. One crucial aspect of this research concerns child abuse and maltreatment. Thus, firstly, an overview of relevant papers and statistics on this issue will be presented. Then, a review of the academia on various machine learning models and their applications will be provided, including the differences between audio-driven and video-driven models, the progress of different models, and the emergence of complex yet highly effective talking head models. The chapter will explore the combination of these two topics, which involves using talking head avatars to develop a tool for taking action against child abuse.

2.1 Child Abuse

To recognize the importance of child abuse issues, statistics were examined. According to a report on preventing child maltreatment published by WHO (World Health Organization), the severity and duration of maltreatment vary. At its worst, the premature death of 852 children under the age of 15 is caused every year. However, this represents only the tip of the iceberg, as its non-fatal forms are much more common, resulting in serious and far-reaching health and social consequences (Sethi et al., 2013). Community surveys provide a better understanding of the scale of the problem, with a prevalence of 9.6% for childhood sexual abuse in Europe (girls 13.4%, boys 5.7%), 22.9% for physical abuse, and 29.1% for emotional abuse. Global estimates suggest that the prevalence of physical neglect is 16.3% and emotional neglect is 18.4%. Projections indicate that approximately 18 million (range 18 million to 55 million) children in the Region have experienced some form of maltreatment, based on a conservative estimate that at least 10% of children suffer from maltreatment. It is essential for vital registration and official statistics to be improved at the country level in order

to evaluate and monitor the scale of the issue, particularly in measuring trends in the most severe cases. Improved record-keeping of children, supplemented by regular surveys, is considered crucial by concerned professionals to detect the much larger proportion of maltreatment in the community that goes unnoticed by child protection agencies.

However, it should be noted that these statistics pertain to countries where people have access to information. It is important to acknowledge that there are numerous instances of maltreatment occurring in many countries, which often go unreported due to various challenges and barriers. Maltreatment in the community is often chronic in nature, rather than acute. Supportive interventions for familial malfunction and parenting assistance are required by most families, rather than retribution and blame. A child's healthy development depends on safe, stable, and nurturing relationships with parents and other caregivers. Severe and recurrent maltreatment may result in toxic stress, impact brain development in childhood, and cause cognitive impairment and the adoption of health-risk behaviors, with adverse mental and physical health outcomes. Post-traumatic stress disorder has been reported in as many as one-quarter of abused children, and child maltreatment may be responsible for nearly a quarter of the burden of mental illness, particularly when combined with other adverse childhood experiences (Sethi et al., 2013).

A deeper investigation into the child abuse statistics and impact is out of the scope of this study. A brief glance at the current reports suggests how important it is to improve therapeutic and clinical practices to resolve child abuse-related issues in society. As emphasized earlier in Chapter 1, the goal of using new technologies like AI-based tools in this context is to provide more effective ways for training professionals. Going deep into machine learning techniques, we can find extensive possibilities and potential for these applications.

2.2 Machine Learning

Machine learning today is connected to what people think of as artificial intelligence. It is a large field within information technology, neurology, artificial intelligence, and other fields, where the end goal is to build a model that is a representation of large datasets. Customized algorithms are applied to datasets to allow computers to learn the desired outcome. Today's society has learned about machine learning through the appearance of a rather complex NLP model called ChatGPT (Abdullah et al., 2022), which seemingly changed the public opinion on what AI is and how it has the potential to automate complex tasks to make life easier. The following section briefly addresses different machine learning categories in general and specific techniques that

are relevant to this study.

2.2.1 Supervised Learning

Supervised learning is identified by the usage of annotated training data, where the "supervisor" provides guidance to the learning system regarding the labeling of training examples. The labeling typically consists of class labels in classification problems. Models are induced from these training data using supervised learning algorithms, which can then be applied to classify unlabeled data. The process of supervised learning involves creating a map between a set of input variables X and an output variable Y, which can be utilized to predict the outputs for new data. This technique holds a crucial position in machine learning and is of utmost significance in the processing of multimedia data (Cunningham et al., 2008).

2.2.2 Unsupervised Learning

Unsupervised learning is a type of machine learning in which the machine receives input data without any supervised target outputs or rewards from the environment. Despite the lack of feedback, unsupervised learning can be based on the idea of building representations of the input that can be useful for making decisions, predicting future inputs, and communicating with other machines. Unsupervised learning is essentially concerned with identifying patterns in data that go beyond pure unstructured noise. Clustering and dimensional reduction are two simple examples of unsupervised learning (Dike et al., 2018).

2.2.3 Reinforcement Learning

Reinforcement learning is a type of machine learning that involves a machine interacting with its environment through actions, which result in rewards or punishments. The aim is for the machine to learn to take actions that maximize future rewards or minimize punishments over its lifespan. Reinforcement learning is closely related to decision theory and control theory, which deal with similar problems, and the solutions to these problems are often formally equivalent, although different aspects are emphasized (H.-n. Wang et al., 2020).

2.3 Generative Models

Generative models belong to a class of machine learning models with the objective of learning and replicating the underlying distribution of a given dataset. Their

main function is to produce new data points resembling the training data they were provided. These models find extensive applications in image synthesis, natural language processing, and data augmentation. Figure 2.1 illustrates the timeline of progress in this field (Toshpulatov et al., 2023). In the subsequent paragraphs, we will delve into some of the methods in greater detail.

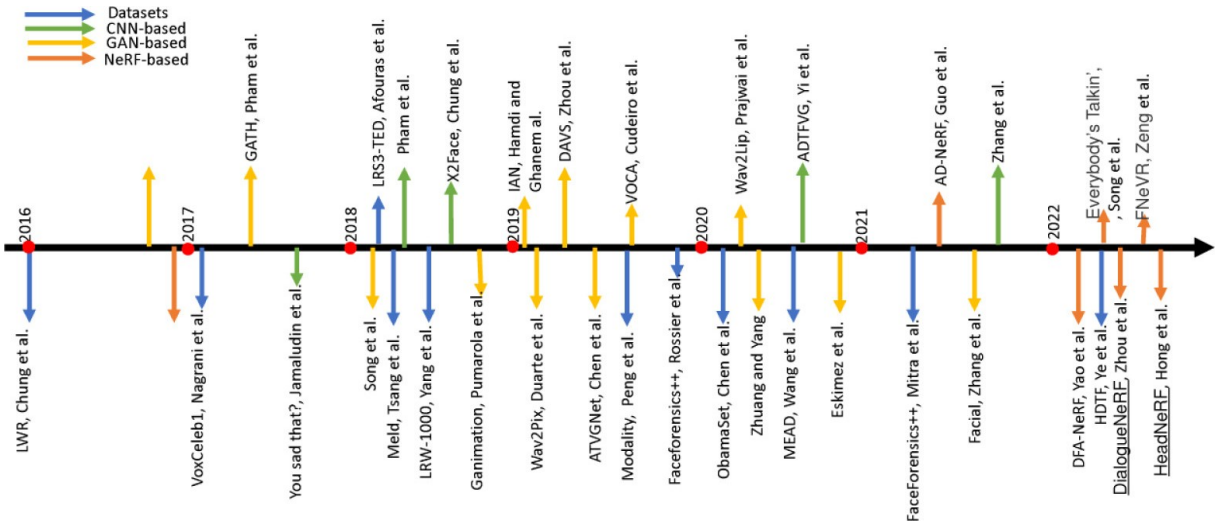


Figure 2.1: Chronological milestones on Talking face generation, including CNN, GANs, and NeRF methods (Toshpulatov et al., 2023)

2.3.1 Convolutional Neural Networks (CNN)

CNN is a type of neural network used for image synthesis and talking face generation in videos. They work on the principle of signal values arranged in a regular grid, and the interactions between these values are limited to a local neighborhood. CNNs have been successful in creating realistic images and synchronizing the movement of a talking face with the audio. This involves predicting sequential pixels to generate images and using Conditional PixelCNN to condition the model on specific vectors (Toshpulatov et al., 2023).

CNN offers promising solutions for generating talking faces in videos. The Cascaded Refinement Network and X2Face are two notable methods that demonstrate the capabilities of CNNs in producing realistic images and controlling facial expressions and poses. These advancements open up new possibilities for video face editing and puppeteering, with potential applications in various fields of computer vision and multimedia.

2.3.2 Variational Autoencoder (VAE)

Variational Autoencoders and their diverse adaptations have found extensive application across various domains, including dialogue generation, image synthesis, and learning disentangled representations(Shao et al., 2020). VAE is a popular generative modeling approach extensively used for unsupervised representation learning. It consists of two interconnected components - the encoder (recognition model) and the decoder (generative model) - working in tandem to produce meaningful latent representations and approximate posterior distributions (Kingma & Welling, 2019). In a VAE, the input data is compressed into a compact representation called the latent space by the encoder. This condensed representation contains essential information about the input. Then, the decoder reconstructs the original data from the latent representation. The VAE's ability to disentangle different factors of variation makes it useful for various applications. One fascinating application is talking head generation, where VAEs can create realistic video sequences of talking faces with control over facial expressions and speech. This makes VAEs a powerful tool in the fields of artificial intelligence and computer graphics.

2.3.3 Generative Adversarial Network (GAN)

Generative adversarial networks are a type of deep neural network architecture designed for unsupervised machine learning tasks in various domains like computer vision, natural language processing, and medical image analysis. The core idea of GANs is to have multiple neural networks compete against each other, leading to optimization and improvement (Toshpulatov et al., 2023). Essentially, GANs act as generative models, learning the underlying distribution of data classes. The pioneering work on GANs was done by Goodfellow et al. in 2014, where they used multilayer perceptrons (MLPs) to model image representations in a latent vector space (Goodfellow et al., 2014). They are inspired by game theory.

The fundamental structure of GANs consists of two sub-networks, namely the generator network and the discriminator network. The generator's primary role is to generate synthetic data samples, attempting to produce realistic data instances that resemble the real data distribution. On the other hand, the discriminator network functions as a critic, distinguishing between the generated samples and real data. As the training process iterates, the generator refines its ability to produce more convincing data, while the discriminator improves its capacity to accurately distinguish between real and fake data. Eventually, the GAN reaches a balanced state where the generator creates highly plausible data samples, and the discriminator struggles to discern between real and generated data, yielding an effective generative

model (Kammoun et al., 2022). The figure presents an overview of the GAN model, showcasing selected images from a project available on GitHub (Perarnau et al., 2016).

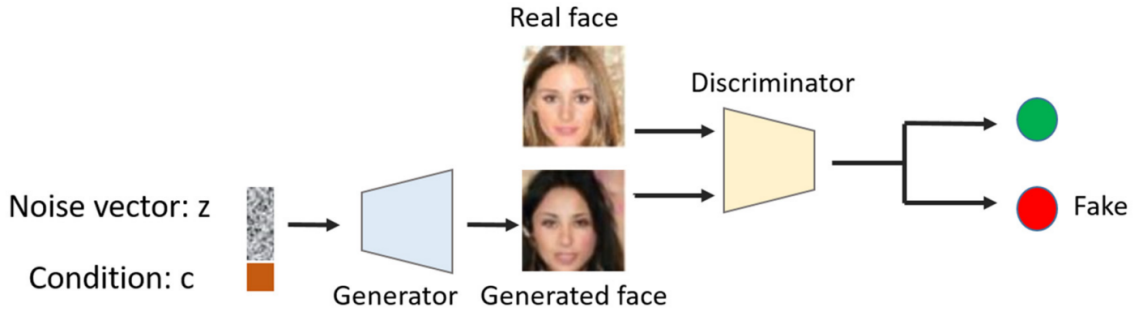


Figure 2.2: Baseline GANs architecture for face generation(Kammoun et al., 2022)

The versatility of GANs and their ability to create realistic data have led to a myriad of applications in the fields of artificial intelligence and computer science. They have been successfully applied to tasks like image generation, style transfer, super-resolution, and data augmentation, as well as text-to-image synthesis and generating realistic human faces. GANs have revolutionized the field of generative modeling and continue to drive innovation in various domains, offering exciting opportunities for advancing artificial intelligence research and applications.

2.3.4 Neural Radiance Field (NeRF)

NeRF is an innovative approach for synthesizing photorealistic views of complex scenes using neural networks. Instead of using traditional 3D mesh or voxel representations, NeRF represents scenes as continuous functions, allowing it to handle intricate geometry and appearance more effectively. The core idea involves optimizing a deep neural network, specifically a fully connected multilayer perceptron (MLP), to map 5D coordinates (spatial location and viewing direction) to volume density and view-dependent RGB color. By marching camera rays through the scene and using volume rendering techniques, NeRF can render novel views of the scene from different viewpoints (Mildenhall et al., 2021).

One of the key advantages of NeRF is its ability to generate high-resolution, realistic images without relying on discretized voxel grids, overcoming the storage limitations associated with traditional methods. This continuous representation also makes the optimization process more efficient, enabling NeRF to achieve state-of-the-art results in view synthesis tasks. Additionally, NeRF can handle real-world scenes with complex geometry and appearance, providing impressive results for various applications, such as view synthesis and virtual reality. Overall, NeRF offers a powerful and versatile

solution for generating lifelike views of scenes, elevating the field of computer vision and graphics to new heights.

2.4 Face Generation

Speech-driven talking face generation is an intriguing field that focuses on creating lifelike facial expressions synchronized with speech. Recently, researchers introduced a novel approach to rendering visual emotional expression in speech-driven talking face generation. Their system takes a speech utterance, a single face image, and an emotion label as input and generates a talking face video that authentically expresses the specified emotion (Eskimez et al., 2021).

One of the key contributions of this approach is its ability to independently control visual emotion expression, making it highly flexible and suitable for diverse applications. By leveraging a neural network, emotional talking faces can be directly generated from input speech and emotion labels, outperforming existing methods in terms of image quality, synchronization, and emotion expression. Recognizing the significance of emotion in speech communication, the proposed method directly conditions the talking face generation on an independent emotion variable, enabling more personalized and expressive results. Unlike existing techniques that estimate emotions from speech or map speech features to facial movements, this approach delivers a more refined emotional representation. Generating faces from single wild images presents a challenging task, which can be addressed through photographs or user interface modeling. Before delving into state-of-the-art methods, a classic face-generation approach called the 3D morphable model representation (3DMM) is explored. The 3DMM provides a continuous parameterization of 3D texture and shape interpretations within a specific category, mapping low-dimensional parameters to high-dimensional ones of textured 3D models while incorporating statistical data using density functions (Toshpulatov et al., 2023).

The versatility of 3DMM extends to various domains, including human face recognition, face generation, animation, lip synchronization, experimental psychology, entertainment, and texture knowledge representation. To overcome the limitations of conventional 3D reconstruction techniques in reconstructing 3D models from paintings, researchers have combined 3DMM with other methods, resulting in significant advancements in 3D human face reconstruction and landmark localization. Furthermore, deep learning techniques, particularly GANs, and CNNs, have revolutionized face generation by synthesizing high-quality, realistic human face images. CNNs have been instrumental in creating facial expressions and recognizing facial characteristics. Recent studies have focused on enriching 3D human face geometry with photomet-

ric data using GANs, enabling the generation of high-resolution facial geometry and reflectance maps.

2.5 DeepFake / FaceSwap

Deepfake is a term derived from "deep learning" and "fake." It refers to a type of fake image and video generation technology based on artificial intelligence, particularly deep learning techniques like auto-encoders and GANs. Deepfakes have made significant progress in recent years, making it easier for people with little knowledge of video editing to create face synthesis videos. Face synthesis videos involve manipulating facial expressions in videos to create realistic and convincing results. Traditionally, face synthesis required specialized tools and professional video editing skills. However, with the advancements in deep learning, especially GANs and auto-encoders, open-source algorithms and tools like DeepFakes have emerged, making face synthesis more accessible to a broader audience (T. Zhang et al., 2020).

There are three main types of face synthesis:

2.5.1 Face Reenactment

This involves transferring the facial expression of one face (source face) to another face (target face). The target face provides the mouth expression for the source face. Face-Reenactment is a cutting-edge technology that allows for the alteration of facial expressions in videos, opening up possibilities for synthesizing speeches for public figures like politicians. In recent years, significant progress has been made in this field. For instance, in 2014, Justus et al. introduced the groundbreaking Face2Face method (Thies et al., 2016). This technique parameterizes various facial aspects such as posture, illumination, and expression, enabling the transfer of expressions by retrieving similar mouth features from a database and blending them seamlessly with the rest of the face.

2.5.2 Face Swap

In face swap, the facial expression of the source actor is kept the same, but the face is replaced with that of the target person. The target face provides the identity for the source face. In the process of face swapping, only specific parts of the source face, such as the mouth, nose, or eyes, can be exchanged with their corresponding features from the target face. This swapped face is then treated as augmented training data for the source face. For each source face image, a similar-looking target face is obtained using k-nearest neighbors (KNN) with an averaged feature map as input for the face part swapping operation.

2.5.3 Face Generation

This type of face synthesis involves generating a talking face video from a single face image and a sequence of audio. It uses neural networks to generate facial expressions based on the input audio. Researchers have worked on improving these face synthesis techniques by disentangling face attributes and identities, using face segmentation, and introducing temporal coherence to create more realistic results. While deepfake technology has various applications, it has also raised concerns about its potential misuse, especially in creating deceptive and misleading content. Efforts are being made to develop countermeasures and identify deepfake videos to address these challenges.

2.6 Talking-Head Generation

Talking-head generation is a computer vision task that involves creating realistic video footage of a person speaking or delivering a message. The field has undergone significant advancements thanks to the use of GANs, which have enabled the generation of high-quality video content. To evaluate the effectiveness of talking-head generation models, researchers have developed well-defined standards and evaluation metrics. These metrics include emotional expression, lip synchronization, and blink motion, and are based on human perceptual judgment. Overall, the field of talking-head generation continues to evolve, with researchers developing new techniques and models to improve its performance.In particular, the development of talking-head generation models has led to the creation of child avatars that can be used in various applications, such as education and entertainment (Chen et al., 2020). Talking head models aim to generate realistic videos of a person's head movements and facial expressions based on a given audio input. These models are useful in applications such as video conferencing, gaming, and virtual reality. There are two main approaches to building talking head models: audio-driven and video-driven.

2.6.1 Audio-driven Talking-Head Generation

Audio-driven models use only the audio input to generate the talking head video. These models are trained to learn the relationship between speech and facial movements from a large dataset of paired audio and video recordings. The advantage of audio-driven models is that they can generate videos even if the input video is not available or is of low quality. A study by Suwajanakorn et al. (2017) proposed an audio-driven talking head model that uses a deep learning architecture called GANs to generate realistic videos of a person speaking based on the audio input. The model

was trained on a dataset of more than 17 hours of video recordings of a person speaking in different poses, lighting conditions, and backgrounds. The study showed that the audio-driven model could generate realistic videos that closely matched the lip movements and facial expressions in the audio input (Suwajanakorn et al., 2017).

2.6.2 Video-driven Talking-Head Generation

Video-driven Talking-Head Generation is a technique used in the field of computer graphics and artificial intelligence to create realistic and expressive talking heads from a given video input. The goal is to synthesize a video of a person speaking with natural facial expressions and lip movements, based on the input video of the person. The technique typically involves using deep learning models, such as GANs and Recurrent Neural Networks (RNNs), to analyze the facial movements in the input video and then generate corresponding realistic facial animations for the talking head. This technology has various applications, including video editing, animation, virtual reality, and creating lifelike avatars in video games and virtual environments. While video-driven models have shown to be more accurate and realistic than audio-driven models, they require more data and computational resources to train. Additionally, they may not perform well when the input video is of low quality or unavailable.

Both audio-driven and video-driven models have their advantages and disadvantages in talking head models. Audio-driven models are useful when the input video is not available or is of low quality, while video-driven models can generate more accurate and realistic videos but require more data and computational resources. Further research is needed to improve the performance of both types of models and to explore new approaches for generating talking head videos.

2.7 Uncanny Valley

One of the main issues with talking head avatars is the uncanny valley. As the avatar becomes more human-like, users may start to feel uncomfortable or even repulsed by it. This can be due to subtle imperfections in the avatar's appearance or behavior that create a sense of unease in the user (Mori et al., 2012).

To avoid the uncanny valley, designers of talking head avatars must carefully balance realism with stylization. They may also need to adjust the avatar's facial expressions and movements to make them more natural and engaging. Additionally, user testing can help identify any aspects of the avatar that may trigger the uncanny valley response. The graph 2.3 presented in (Igaue & Hayashi, 2023) clearly illustrates the threshold of the uncanny valley.

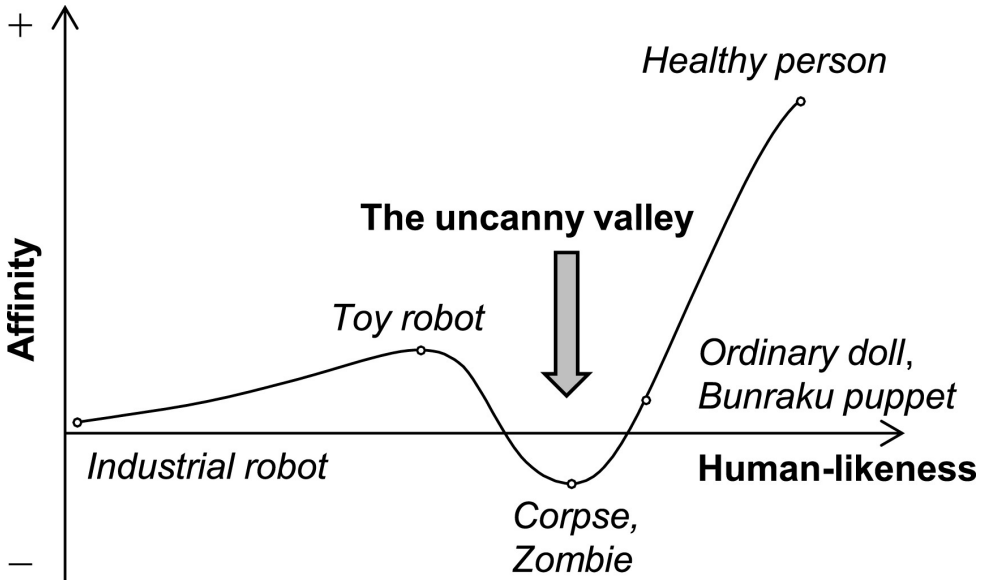


Figure 2.3: Graph of Uncanny Valley Threshold represented with (Igaue & Hayashi, 2023)

Insights into the mechanisms underlying the uncanny valley and potential design principles for bridging the gap between human likeness and comfort level are presented in MacDorman et al.'s paper (MacDorman et al., 2009). These principles are also relevant to talking head avatars. Adjusting the texture and level of detail in the avatar's skin can have an impact on its perceived eeriness and human likeness. Additionally, unconventional facial proportions may be more unsettling in photorealistic faces, and a discrepancy in the size and texture of the eyes and face can contribute to an eerie appearance (MacDorman et al., 2009).

In conclusion, the design of talking head avatars can be affected by the uncanny valley phenomenon. Realism and stylization must be balanced by designers, who must carefully adjust facial expressions and movements to prevent the uncanny valley response in users from being triggered. The principles outlined in (MacDorman et al., 2009) can guide the design of talking head avatars that are both engaging and realistic. In our user study, a cartoony and painted style was chosen to examine whether it elicited a better response from users in terms of the uncanny valley phenomenon.

In summary, we examine some of the mentioned methods in this section and try to propose a benchmark for the talking head generation methods based on multiple factors in the context of realistic facial expressions. The more sophisticated models are supposed to provide better results, while the uncanny valley phenomenon suggests there should be a trade-off between realism and stylization.

Chapter 3

Research Methodology

This chapter is dedicated to elucidating the methodologies and models employed in crafting the videos for our user study. Diverse tools were utilized for this purpose, and these tools and the underlying technology will be expounded upon in the subsequent section. Furthermore, insight will be provided into the rationale underpinning the selection of the three models amidst a multitude of alternatives.

Within the realm of tools, a distinction was made between those used for executing model-related code and those employed for in-depth result analysis. The comprehensive elucidation of these tools will be undertaken within this section, affording a comprehensive understanding of the procedural aspects adopted.

3.1 Tools and Technologies

Various tools are employed in this thesis. Some of these tools are used during the execution of the three models, while others are utilized for analyzing the conducted user study. In this section, we will provide a brief explanation of some of these tools.

3.1.1 Google Colaboratory

Google Colab is a cloud-based platform provided by Google for running and sharing Jupyter notebooks. It offers a free environment that allows users to write and execute Python code, as well as access to powerful GPU and TPU resources. Google Colab provides pre-installed libraries and supports collaboration, enabling multiple users to work on the same notebook simultaneously. It is widely used by data scientists, researchers, and students for data analysis, machine learning, and collaborative coding projects. Since it runs entirely on Google's servers, users can access Colab notebooks from any device with an internet connection without the need for local installations

or configurations. Due to the use of various models and the requirement to execute them, it becomes crucial to ensure compatibility between different Python libraries. To achieve this goal, a tool was chosen that could install all the necessary libraries from the beginning. This approach facilitated seamless integration and ensured the smooth execution of the models. However, a drawback of this approach was that uploaded files, such as pictures, audio, and videos, were not automatically saved. As a result, users had to re-upload these files each time they started running the code after a certain period. This inconvenience could be time-consuming and interrupt the workflow during repeated executions.

3.1.2 Python

The majority of the work in executing models in this thesis relies on Python 3, the latest version of the Python programming language, which was released in 2008. Python is a popular and versatile programming language known for its simplicity and readability. It is widely used in various domains, including web development, data analysis, artificial intelligence, and automation. Python's straightforward syntax and extensive libraries make it accessible for beginners and efficient for experienced developers. Its versatility and strong community support have contributed to its widespread adoption across different industries, making Python a valuable tool for tackling a wide range of programming tasks.

3.1.3 Python Libraries

In this subsection, some of the important Python libraries used during our work will be mentioned. First, **FFmpeg-python** is a Python library that conveniently interfaces with the FFmpeg multimedia framework. FFmpeg, a powerful and versatile command-line tool for handling audio and video data, is utilized by ffmpeg-python through a wrapper around its command-line commands. This enables Python developers to harness the capabilities of FFmpeg within their Python scripts and applications, facilitating various multimedia processing tasks such as video and audio conversion, editing, and manipulation directly from Python code. The library offers a user-friendly and efficient way to interact with FFmpeg functionality, simplifying the process of working with multimedia data in Python projects. We used this library in all three models. Next, **Scikit-learn**, is a popular and widely used machine learning library for Python. It provides a comprehensive set of tools for various machine-learning tasks, including classification, regression, clustering, dimensionality reduction, and more. Scikit-learn is built on top of other Python libraries such as NumPy and SciPy, making it easy to integrate into existing Python data analysis workflows. Scikit-

learn also provides features for data preprocessing, model evaluation, and model selection, making it a complete package for machine learning projects. It is widely used in academia and industry for various applications, and its active community ensures that it stays up-to-date with the latest advancements in the field of machine learning. This library has been used in the implementation of our models. The **Librosa** library was utilized in both the MakeItTlk model and the PC-AVS model in this study. It serves as a Python library designed specifically for audio and music signal processing, offering a user-friendly interface to handle audio data and extract relevant information from audio files. Librosa enables users to analyze, manipulate, and extract features from audio signals, making it a versatile tool for various audio-related tasks. Its widespread use in the field of audio processing and music research can be attributed to its simplicity, efficiency, and extensive documentation. For developers and researchers working on audio-related projects, Librosa proves to be an invaluable resource, facilitating effective audio data processing and analysis within their Python applications. **PyTorch**, used in the FOMM model motion model and PC-AVS model, stands as a prominent open-source machine learning library extensively utilized in the artificial intelligence domain. Created by Facebook's AI Research Lab (FAIR), PyTorch offers a versatile and efficient framework for constructing and training deep learning models. Its unique "define-by-run" approach, featuring a dynamic computational graph, empowers developers to construct models on-the-fly while executing the code, leading to simplified debugging, experimentation, and implementation of intricate architectures.

3.1.4 Analysis Tools

In statistical analysis, SPSS (Statistical Package for the Social Sciences) and Power BI have been used as essential tools for data analysis. SPSS is a widely recognized software package that offers a comprehensive set of tools for statistical analysis and data management. It caters to the needs of researchers and analysts, enabling them to perform a wide range of statistical tests, from basic descriptive statistics to advanced inferential analysis. With its user-friendly interface, SPSS simplifies data manipulation, visualization, and reporting, making it a preferred choice for researchers in social sciences, psychology, and other fields.

Power BI also is a powerful business intelligence tool developed by Microsoft. While it is primarily used for data visualization and reporting, it also integrates statistical functionalities to analyze and interpret data effectively. Power BI offers a variety of visualization options, including charts, graphs, and dashboards, enabling users to gain insights from complex datasets quickly. Additionally, it allows users to create interactive reports and share them with stakeholders, facilitating better decision-

making processes.

3.2 Models

In order to evaluate users' experiences with various talking head avatars, it was necessary to select specific models to execute and generate results for user feedback. Multiple models were considered, encompassing both older and more recent iterations. One of the criteria used for selecting models was to ensure that they had a functional GitHub page with open-source codes that could be executed with Google Colab. In the initial stage of the user study, our focus was on audio-driven models, which are well-known in the field. We aimed to compare the results of two of these models with a more recent model.

Initially, we gathered a long list of talking head avatars spanning different years. From this list, we narrowed it down to nine models, including one called Facial [1](C. Zhang et al., 2021a). This model uses implicit attribute learning to create realistic dynamic talking faces. It combines techniques like face reconstruction, audio feature extraction, and face rendering to generate lifelike animations for various purposes. We attempted to use this model, but it required a specific dataset for training. However, finding or creating the needed dataset was time-consuming and unfeasible, leading us to exclude this model from our selection.

The other Model in our shortlist was the "LiveSpeechPortraits[2]" model that is an implementation of a real-time photorealistic talking-head animation system (Lu et al., 2021). This system generates personalized talking-head animations driven solely by audio signals at a frame rate of over 30 fps. It consists of three stages: deep audio feature extraction, learning facial dynamics and motions from the audio features, and synthesizing photorealistic renderings with explicit control of head poses using image-to-image translation networks. The model showcases the ability to create high-fidelity personalized facial details and offers better performance compared to existing techniques, as demonstrated through qualitative and quantitative evaluations. This model offers a user-friendly web demo[3] that's quite convenient to use. However, its limitation became apparent in our case. The demo only allowed selections from their predefined characters and audio options, lacking the option to utilize custom audio or video inputs. This misalignment with our project's objectives led us to remove this model from consideration. The subsequent option explored was the "Wav2Lip[4]"

[1] https://github.com/zhangchenxu528/FACIAL
[2] https://github.com/YuanxunLu/LiveSpeechPortraits
[3] https://replicate.com/yuanxunlu/livespeechportraits
[4] https://github.com/Rudrabha/Wav2Lip

model, recognized for accurately syncing lip movements with speech in videos. This model is versatile, accommodating various identities, languages, and even CGI faces. It draws from a research paper presented at ACM Multimedia 2020 (Prajwal et al., 2020). The Wav2Lip model effectively generates lip-sync animations, aligning lip movements with different audio inputs to produce remarkable results. However, this model's applicability was limited by our dataset, which solely contained pictures of characters without corresponding videos. Consequently, due to this constraint, we couldn't utilize the Wav2Lip model for the purposes of this study.

There are two other models named "One-Shot[5]" and "PIRenderer[6]". The One-Shot Free-View Neural Talking Head Synthesis(T.-C. Wang et al., 2021) has a goal to create dynamic talking-head videos for video conferencing. It generates talking-head animations from a single image, allowing viewing from different angles. The repository includes training, demo capabilities, and pre-trained models for various resolutions. It focuses on "one-shot free-view" synthesis, using rotation matrices and post-processing like face restoration for improved results. The "PIRenderer" model, from the ICCV2021 paper(Ren et al., 2021), introduces a unique method for creating controllable portrait images using Semantic Neural Rendering. By untangling 3DMM parameters, the model enables intuitive control over facial motion synthesis, useful for portrait editing, alignment, imitation, and audio-driven reenactment. The model's architecture intelligently manipulates facial motions based on fully untangled 3DMM parameters, offering a powerful tool for creating realistic and manageable facial animations and advancing portrait image synthesis. However, both models lack a Google Colab demo that was promised to be available soon but hasn't been published yet due to system limitations and time constraints. As a result, they were removed from the current list and considered for future work due to challenges in executing the code and the time required. Table 3.1 shows the summary of all unused models and the reason of exclusion.

[5] https://github.com/zhanglonghao1992/One-Shot_Free-View_Neural_Talking_Head_Synthesis
[6] https://github.com/RenYurui/PIRender

Model Name	Type	Summary of Model	Reason for Exclusion
Facial	Dynamic Talking	Utilizes implicit attribute learning for realistic talking faces.	Specific dataset needed for training impractical to obtain.
LiveSpeechPortraits	Real-time Animation	Generates real-time personalized talking-head animations based on audio signals	Web demo limitations only predefined choices allowed.
Wav2Lip	Lip-Sync Animation	Accurately syncs lip movements with speech in videos.	Dataset lacked character videos for lip-syncing.
One-Shot	Talking Head Synthesis	Creates dynamic talking-head videos from single images.	Lack of Google Colab demo unavailable due to constraints.
PIRenderer	Controllable Portrait	Generates controllable portrait images using Semantic Neural Rendering.	Absence of Google Colab demo unavailable due to constraints.

Table 3.1: Excluded Models from the short list

Finally, for this study, three models were selected: **MakeItTalk Model**, **FOMM model**, and **PC-AVS model**. Each model will be briefly outlined, with a little more explanation provided for each:

3.2.1 MakeItTalk Model

MakeItTalk is a model that generates realistic and expressive talking head videos from audio input. The model consists of a motion generation module, a synthesis module, and a fine-tuning module. The motion generation module uses a 3D mesh to generate facial animations that match the audio input. The synthesis module renders the animations into photorealistic videos, and the fine-tuning module further improves the quality of the videos by learning from a small number of real-world video frames. The resulting model can be used to synthesize high-quality videos of a person speaking, even if there is no video of that person actually speaking the words in question (Y. Zhou et al., 2020).

To achieve this, the researchers developed a novel architecture for the deep learning model, which combines a motion encoder that extracts facial movements from a video, a speech encoder that converts audio into a latent representation, and a decoder that generates the final video. The model is trained using a loss function that encourages it to generate videos that match both the facial movements and the audio.

The researchers demonstrate the effectiveness of their approach through a series of experiments, including a comparison to other state-of-the-art methods for generating talking head videos. They also show how their model can be used for a variety of applications, including voice conversion and caricature generation. The model's overview is depicted in Figure 3.1.

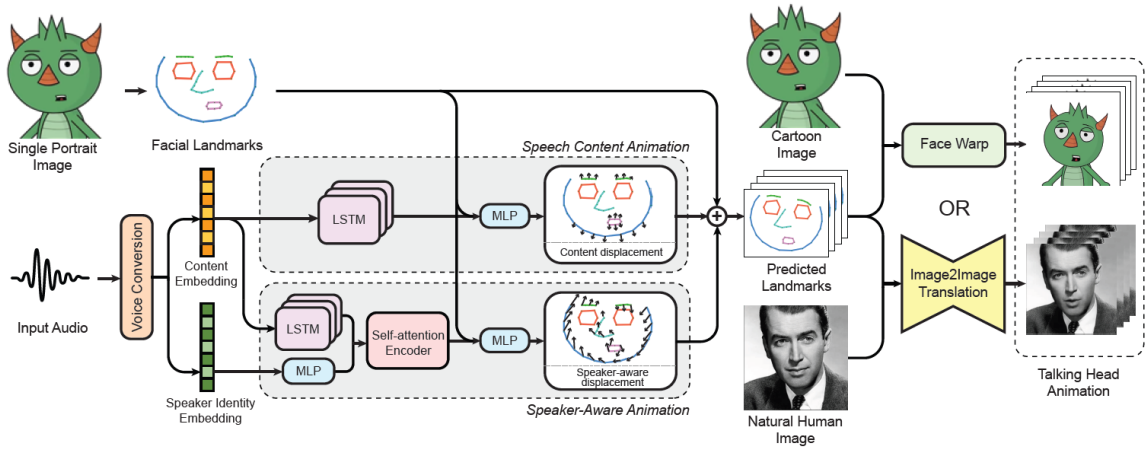


Figure 3.1: The overview of the MakeItTalk model has been published in their paper (Y. Zhou et al., 2020)

The authors of the paper also propose a technique for voice conversion, which allows the generated videos to have the voice of a different person than the one in the reference images. This is achieved by training a separate neural network to convert the voice of the input audio to the desired target voice.

They also have faced some limitations in the current approach and suggested avenues for future research, such as incorporating more complex facial expressions and gestures and improving the model's ability to handle variations in lighting and pose. Overall, this method presents a promising avenue for generating high-quality talking head videos with a wide range of applications.

To utilize the MakeItTalk model, we must have an audio file in the *.wav format and a single avatar picture with a resolution of 256*256 pixels. The resulting video shows the main output in the center, the landmark video on the left, and the picture on the right. The final step involves cropping the video to retain only the main portion. This can be done either by modifying the code or by using video editing software, with the latter option resulting in higher video quality. After completing these steps, the final videos are produced. The MakeItTalk model can be found in this GitHub directory[7]

3.2.2 First Order motion model

The FOMM model is a novel approach for image animation based on key points and local affine transformations. The method models the motion between two frames as a set of keypoint displacements and local affine transformations, which are efficiently computed using a FOMM model Taylor expansion approximation. The motion

[7] https://github.com/yzhou359/MakeItTalk.git

representation is then combined with the appearance of the source image using a generator network to generate an animated sequence (Siarohin et al., 2019a).

The method uses a set of key points, which are defined as distinctive locations on the image surface, to represent the motion between two frames. The key point displacements are computed by comparing the locations of the key points in the two frames. The local affine transformations are used to describe the non-rigid deformation of the image surface between the frames.

To model occlusions in the image, the method uses a binary mask that indicates which parts of the image should be inpainted. The mask is generated by computing the foreground and background regions based on the difference between the source and driving frames. The method then applies the mask to the appearance of the source image to generate an inpainted image that is used by the generator network.

The generator network consists of a series of convolutional layers that encode the appearance of the source image and the motion representation of the driving video. The encoded features are then passed through a series of deconvolutional layers to generate the animated sequence. A comprehensive understanding of their model can be gleaned from the detailed overview presented in figure 3.2.

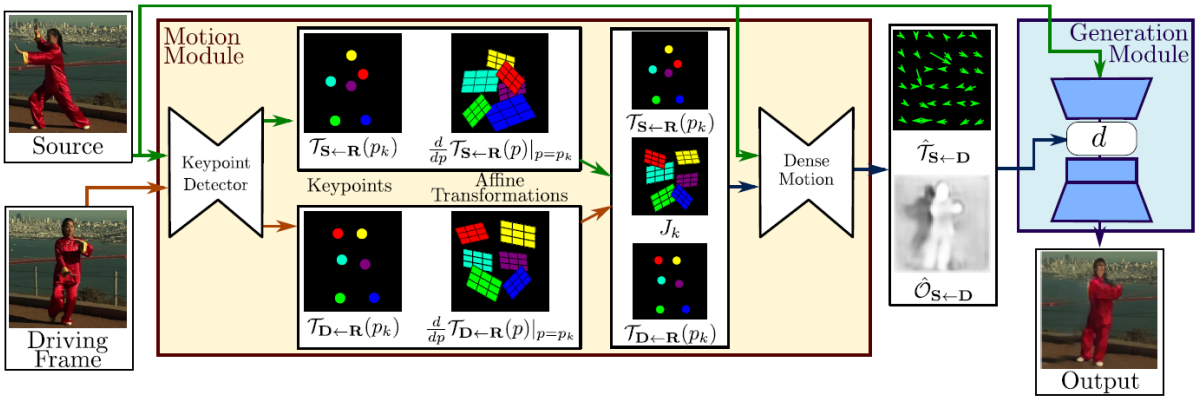


Figure 3.2: The overview of the FOMM model has been published in their paper (Siarohin et al., 2019b)

The method was evaluated both quantitatively and qualitatively on several benchmark datasets and showed superior performance compared to state-of-the-art methods. The method demonstrated particularly impressive results on datasets with highly non-rigid objects, such as the Tai-Chi-HD dataset, indicating the effectiveness of the local affine transformations in modeling complex motion.

To generate a talking head avatar using the FOMM model, we require an image with dimensions of 256*256 pixels and a video in *.mp4 format. The video should contain

the desired audio that will be incorporated into the final result. Additionally, the video should not exceed in size. The model utilizes the audio and head movement from the video to generate a synchronized and lip-synced final video on a single image. To achieve this, we recorded a video with exaggerated lip movements and added a robotic audio track to create the desired final version. The Link for checking more details about this model is in the footnote [8] .

3.2.3 PC-AVS model

The Pose-Controllable Audio-Visual System (PC-AVS) is a framework for generating talking faces with accurate lip-sync and free pose control. This model is more recent compared to the other two. The model consists of three modules: the Audio-Visual Encoder, the Pose Decoder, and the Image Generator. The Audio-Visual Encoder extracts features from the input audio and video, which are then modularized into latent identity, speech content, and pose space. The Pose Decoder learns to decode the pose information from a reference video and generates a pose code, which is combined with the modularized features to generate the talking face image using the Image Generator (H. Zhou et al., 2021). The Overview of the model has been extensively detailed in their published paper, as depicted in figure 3.3.

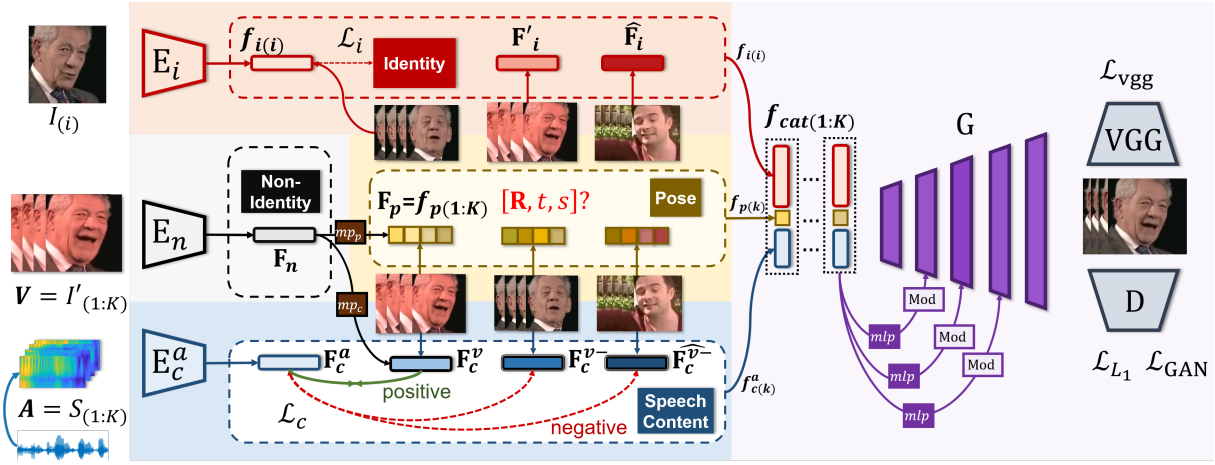


Figure 3.3: The overview of the PC-AVS model has been published in their paper (H. Zhou et al., 2021)

The PC-AVS model uses a novel complementary learning procedure that optimizes both lip-sync accuracy and pose control. It also introduces a contrastive loss function that improves the synchronization between the audio and video. The model achieves high performance on several metrics, including lip-sync accuracy, head movement

[8] https://github.com/AliaksandrSiarohin/first-order-model.git

naturalness, and video realness. It also demonstrates great robustness under extreme conditions, such as large poses and viewpoints.

Technical issues addressed in the paper include the design and training of the three modules, the complementary learning procedure, the contrastive loss function, and the evaluation metrics used to assess the performance of the model. Ablation studies were conducted to evaluate the impact of different aspects of the model, such as the pose code length and the design of the generator. Overall, the PC-AVS model represents a significant advancement in the field of audio-visual synthesis and has the potential for various applications, such as virtual assistants, video conferencing, and entertainment.

To use the PC-AVS model, certain inputs are required which are specified in a script called "prepare-testing-files.py". This script provides great flexibility in formulating metadata through the use of several flags. The first flag, "src-pose-path", is used to indicate the driving pose source path, which can be either a *.mp4 file or a folder containing frames in the form of "0*.jpg" starting from 0. The second flag, "src-audio-path", is used to specify the audio source path, which can be a *.mp3 audio file or a *.mp4 video file. If a video file is given, the frames are automatically saved in a folder, and the "src-mouth-frame-path" flag is disabled. The third flag, "src-mouth-frame-path", is used when the audio path is not a video path, and it provides the folder containing the video frames that are synced with the source audio. The fourth flag, "src-input-path", is the path to the input reference image, which is converted to frames when it is a video file. Lastly, the "csv-path" flag specifies the path to the metadata csv file to be saved. This CSV file can be manually modified or additional lines can be added to it following the rules defined in "prepare-testing-files.py" or the data loader "data/voxtest-dataset.py". The "misc" folder contains several demo choices, including ones used in the provided video. Users are free to rearrange these files across folders and are welcome to record their own audio files. We have used the "src-pose-path" from the main code but included our own picture and video with the voice which we wanted to have on a result video. Here [9] is the GitHub URL to check more details about this model.

In summary, the method used in this study includes different parts such as implementing the models to generate different videos with varying characteristics and running a user study to evaluate the impact of different cases. In Table 3.2, an overview of the used models is provided. This table presents important details such as the types of models, their GitHub locations, required inputs, and notable features. This streamlined presentation allows readers to easily grasp the models' characteristics. The user study is explained in detail in the next chapter.

[9] https://github.com/Hangz-nju-cuhk/Talking-Face_PC-AVS.git

Model Name	Type	GitHub	Required Inputs	Notable Features
MakeItTalk	Audio-Driven	Link	Audio file (in *.wav format) Avatar picture (256x256 pixels)	Realistic facial animations
FOMM	Image Animation	Link	Image (256x256 pixels) Video (in *.mp4 format)	Effective for highly non-rigid objects
PC-AVS	Audio-Driven	Link	Driving pose video (in *.mp4 or image frames) Audio source (in *.mp3 or *.mp4 format) Input reference image	Lip-sync accuracy, free pose control

Table 3.2: Summary of the used Model

Chapter 4

Experiment

The Experiment chapter introduces a comprehensive user study focused on gauging how people perceive talking head avatars. By interacting with 19 animated avatar videos, participants assessed qualities such as how real and high-quality the avatars appeared. The subsequent sections provide a breakdown of the materials employed in creating this study. Furthermore, the chapter offers a detailed explanation of the two questionnaires used. These sections offer insights into the specific questions, their objectives, and the reasons behind their inclusion in shaping the study. Additionally, it's worth noting that the study includes demographic information about the participants, providing a comprehensive view of their backgrounds.

4.1 User Study

The research study conducted for this Master's thesis aimed to evaluate the performance of communication networks using a passive test paradigm based on ITU-T Rec. P.809 (Schmidt et al., 2018). This approach involves collecting data about network performance without active engagement with the network by monitoring network traffic and measuring network parameters such as packet loss, delay, and congestion (Barman et al., 2023). To collect data for the study, a crowdsourcing approach was employed to gather input and feedback from a diverse group of individuals. The questionnaire was distributed randomly through various online platforms including Discord, Telegram, and group emails. Additionally, targeted emails were sent to research groups in other universities to solicit participation in the study. This approach proved to be an effective and cost-efficient method for collecting a large amount of data and feedback from participants.

For this user study, a total of 37 individuals participated, each of whom was selected at random to complete the questionnaire. The questionnaire was distributed through

various platforms and research groups as well as among individuals who were believed to be interested in participating and providing feedback. It took almost 10 days to reach out to a good number of responses. However, there was a problem with the questionnaire, as shuffling the questions was necessary to avoid bias in the answers. Using Google Forms was not an option for this purpose, so we had to search for different free platforms for collecting data. We found that Microsoft Forms was a very good tool for gathering data. But we still encountered an issue as we could only lock one period for one questionnaire and shuffle the rest. We needed to have some initial questions at the beginning, followed by a set of shuffled questions, and then another section for our overall assessment, which also needed to be shuffled. As a result, we had to create two separate questionnaires.

To be able to match the two questionnaires and identify which person answered which questionnaire, we included two unprompted questions to track the twin questionnaires. We did not ask any personal questions, but for some analyses, we needed to know which questionnaire belonged to which person. Overall, the combination of crowdsourcing and passive monitoring of network traffic provided valuable insights and feedback from a diverse group of individuals, enabling us to evaluate the performance of communication networks effectively. The questionnaires could be found in A and B.

4.1.1 Demographic of participants

As part of our study, participants were provided with optional questions that aimed to gather information about their gender, age, profession, and experience with talking head user studies. While participants were not required to provide personal information, these questions were included to gain a better understanding of the characteristics of the respondents and their opinions. Out of the total participants, 34 individuals voluntarily provided their gender, while 35 participants provided their age. To facilitate the understanding and visualization of the study's findings, all diagrams related to the research will be provided below.

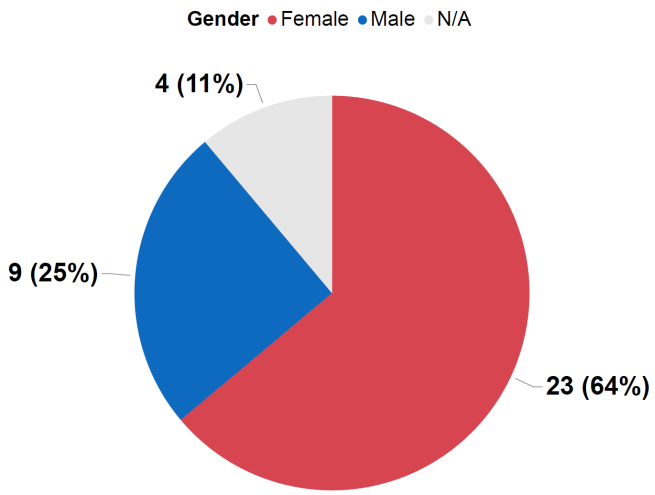


Figure 4.1: Gender Distribution

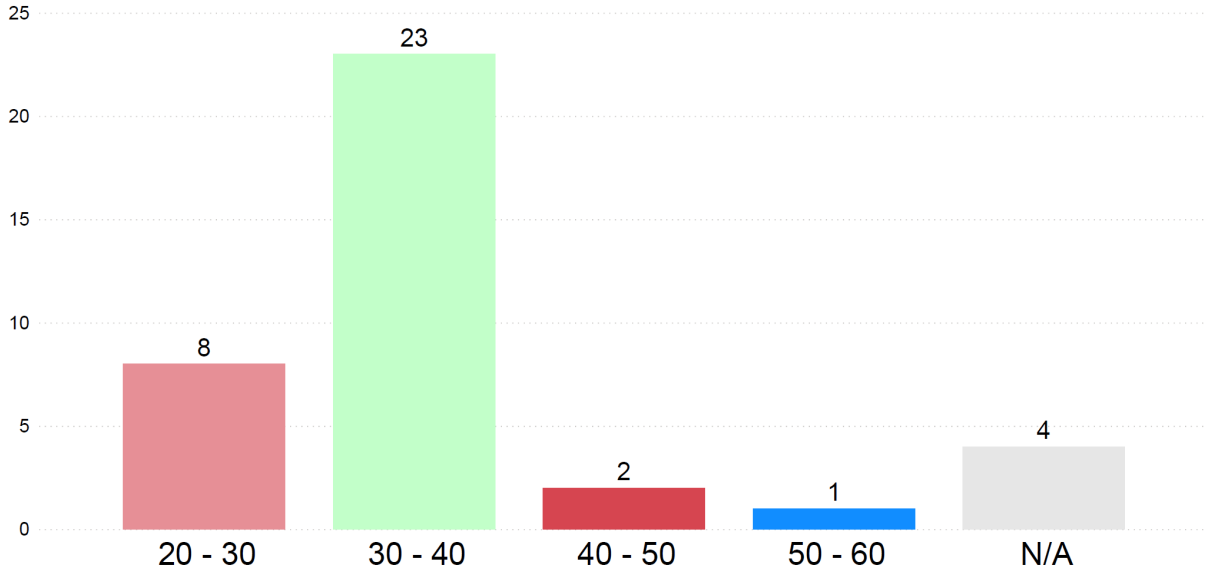


Figure 4.2: The number of participants in each Age Range

The question about the profession was answered by 30 individuals. The participants in this user study belong to diverse professional backgrounds that can be categorized into three main categories: medical and health professions, science and engineering professions, and business and administration professions. The first category includes professions related to healthcare, such as doctors and pharmacists, as well as general interns. The second category includes professions related to science and engineering, such as water hydraulic engineers, Ph.D. cancer biologists, and software developers. The third category includes professions related to business and administration, such as

digital marketing managers, research project coordinators, and innovation managers. It is worth noting that some professions may belong to multiple categories. A chart 4.3 can be drawn to visually represent the distribution of the participants' professions across the three categories.

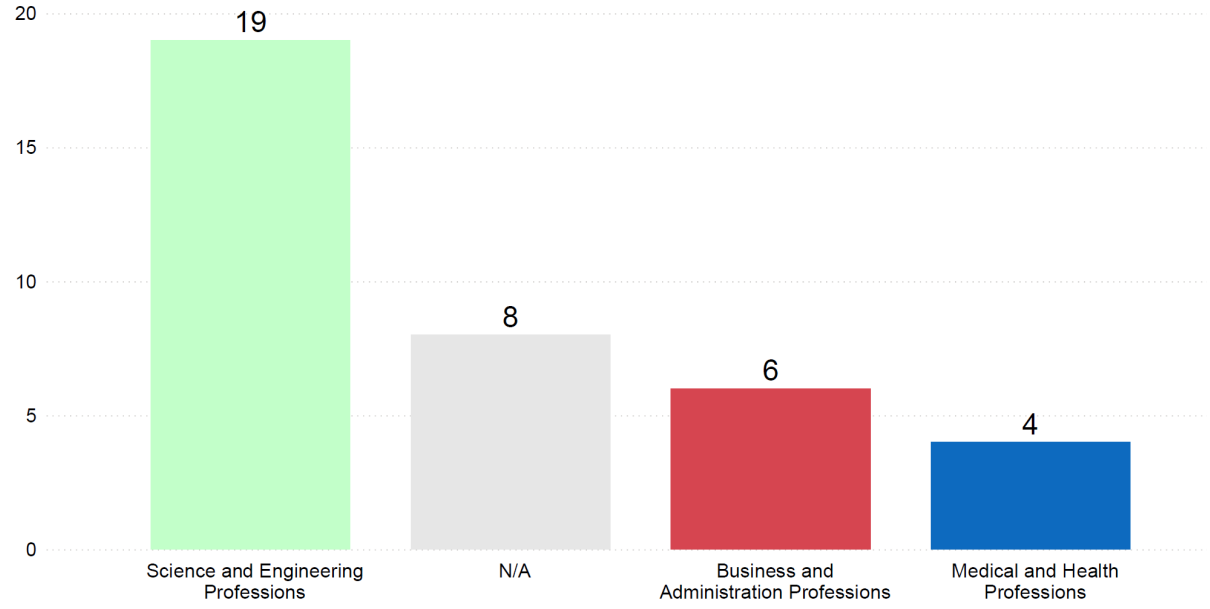


Figure 4.3: Profssion Distribution

4.1.2 Questionnaire

To provide transparency and aid in comprehending the research methodology, the complete questionnaire will be included in the appendix of this study. The questionnaire consisted of specific questions that aimed to gather relevant information for the study.

Main Questionnaire

Pretest questions were administered to gather information about the answerers' details and to check the appropriateness of sound on their devices. Two random questions were also included to determine which questionnaire belonged to each participant as explained before.

The main study involved participants watching videos and rating their experience based on a series of questions. These questions were inspired by the work of (Wilson et al., 2018), which proposed a questionnaire for the assessment of realism in stimuli. Some questions were also inspired by another questionnaire conducted in (Salehi,

Hassan, Shafiee Sabet et al., 2022). Questions about the quality and naturalness of head movement were added to the questionnaire to assess this aspect as well.

A comprehensive summary of the questions presented in the main-test part of the questionnaires, along with the corresponding codes assigned to each question, is provided by Table 4.1.

Labels	Codes
How was your overall experience with the avatar?	**OverallQoE**
How accurately did the lips move in sync with the audio?	**LipSync**
How do you rate the naturalness of head movement?	**HeadMove**
How realistic the avatar was talking?	**Talking**
How do you rate the overall quality of this video?	**VideoQuality**
How comfortable do you feel conversing with this avatar?	**Comfortably**

Table 4.1: Codes Represented by Each Question

The individual questions and their respective purposes in the main test part will be presented below. This will assist in better understanding the methodology and provide a comprehensive overview of the study's objectives.

First question: ***How was your overall experience with the avatar?***

This is a broad and general question that seeks to gather feedback on the user's experience with the talking head avatar. This question aims to capture the user's overall impression of the avatar's effectiveness as a means of communication and whether it was engaging and useful.

Second question: ***How accurately did the lips move in sync with the audio?***

This question aims to evaluate the lip sync accuracy in the videos, which is a crucial aspect of communication between the user and the talking head avatar. The purpose of including this question is to compare the users' overall sentiment with the quality of the lip sync and examine their relationship.

Third question: ***How do you rate the naturalness of head movement?***

One important aspect of the user's experience with the avatar is the naturalness of head movement.

When the avatar responds naturally to the user's head movements, it can enhance the user's sense of presence in the virtual environment. This sense of presence can lead to a more immersive and engaging experience. Furthermore, the natural head movement can also improve communication between the user and the avatar. By mimicking natural human behavior, the avatar can convey subtle cues and emotions that are not easily conveyed through text or voice communication.

In addition, the natural head movement can provide the user with more interaction and feedback based on the avatar's behavior. For example, the avatar may nod its head in agreement or shake its head in disagreement, providing the user with instant feedback. This feedback can enhance the user's sense of control over the avatar and the virtual environment, leading to a more satisfying experience.

Fourth question: ***How realistic the avatar was talking?***

For talking head avatars to effectively communicate with users, it is crucial that they are capable of producing natural-sounding speech and establishing a strong connection with the user. Given that users often rely on visual cues, such as the avatar's mouth movements, to determine whether the avatar is speaking accurately, it is important to design avatars that can produce realistic mouth movements and convey emotions through facial expressions.

Fifth question: ***How do you rate the overall quality of this video?***

While previous models have addressed numerous features of these avatars, such as lip sync and head movement, one crucial aspect that must not be overlooked is the quality of the video. High-quality video is essential in facilitating accurate communication between the user and the avatar, as it enables the user to discern details such as facial expressions and gestures more clearly. In addition, better video quality can enhance the user's overall experience and increase engagement with the avatar. Therefore, designers of talking head avatars must prioritize the quality of the video to ensure that users can interact with the avatars effectively and enjoyably.

Sixth question: ***How comfortable do you feel conversing with this avatar?***

In communication between users and talking head avatars, one of the most crucial

aspects is to ensure that users feel comfortable interacting with the avatars. This is especially critical in fields such as child abuse, where talking head avatars can be used to assist professionals in communicating with children. When users feel comfortable, professionals can ask concise and effective questions, and children are more likely to answer honestly and provide more detailed information. Thus, designers of talking head avatars must prioritize user comfort to enhance the effectiveness and accuracy of communication between the user and the avatar in sensitive fields such as child abuse.

Post-Test Questionnaire

After watching each video, participants completed a main questionnaire. In addition, a post-test questionnaire was administered after participants finished the test and watched all the videos. The post-test questionnaires were administered to obtain a general overview of users' answers with regard to different avatar styles, including cartoony, painted, and original picture avatars. All the questions from the post-test part can be found in Table 4.2.

Questions	Labels
Q1	Generally, which avatars have the most realistic appearance for you?
Q2	Generally, which avatars do you like the most?
Q3	For a conversation (talking to a computer), which avatars do you prefer to use?
Q4	Considering the avatars' talking, which was the most realistic?
Q5	Generally, which avatars were the most believable for you?

Table 4.2: Post-test Questions

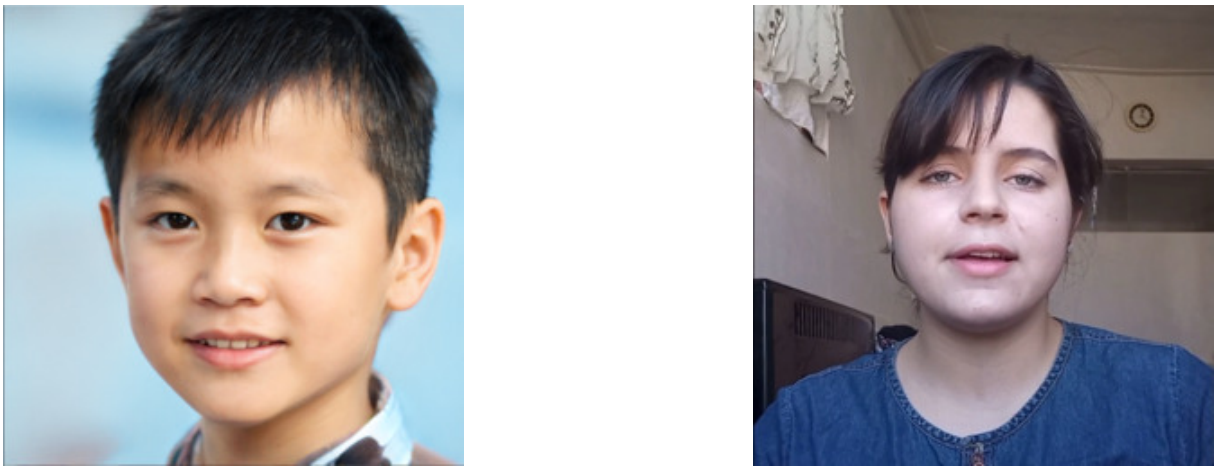
In questionnaire research, the Likert scale is commonly employed to obtain participants' preferences or level of agreement with a given statement or set of statements. The Likert scale is a non-comparative scaling technique that employs an ordinal scale and is unidimensional, measuring a single trait. The scale usually includes a range from "Strongly Disagree" to "Strongly Agree," with "Neither Agree nor Disagree" in the middle, although some practitioners advocate for the use of 7 or 9-point scales to provide greater granularity (Bertram, 2007).

For our user study, we utilized a 5-point Likert scale, with scores ranging from 1 to 5. The words "bad," "poor," "fair," "good," and "excellent" were used to represent scores 1 to 5 in our analysis. The use of the Likert scale enables us to quantitatively measure participants' responses and facilitates data analysis. The Likert scale was initially developed by Dr. Rensis Likert, a sociologist at the University of Michigan, in the 1930s

as a scientific method of measuring psychological attitudes. Since then, the Likert scale has been widely used in various fields, including social sciences, marketing research, and psychology. Its ability to capture participants' attitudes in a quantitative manner has made it a popular tool in many research settings.

4.1.3 Test Materials

In this study, different aspects of image generation were tested and examined by selecting two pictures of characters. Donya, a real girl who had participated in another study, was used, and we had permission to use her pictures. Kian's picture was generated with StyleGAN (Karras et al., 2020) from the website click here. Figure 4.4 shows the Original pictures in this study.



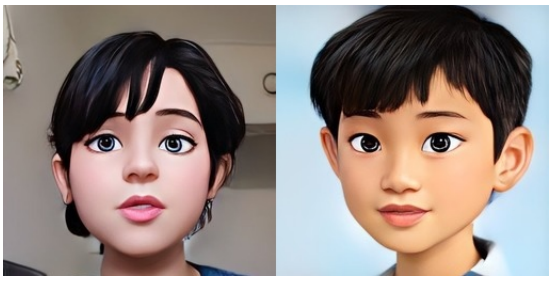
(a) Kian

(b) Donya

Figure 4.4: Illustration of the two distinct Characters evaluated in the user study.

Three different styles were generated for each character. The objective was to find a code that could generate cartoonish and painted versions of the pictures. Useful code was discovered on Github to generate results, but the quality was considered inadequate. An attempt was made to apply Wang et al.'s proposed model for generating animated images to our pictures (X. Wang & Yu, 2020). While the results obtained were satisfactory, the significant difference between the original and final videos was not easily discernible for those particular images. Therefore, they were not utilized as the final videos. However, some codes were not able to generate high-quality images (Yang et al., 2022), so a website was used to generate different styles for the characters[1]. Below in figure 4.5, the cartoony, painted, and original styles used in the user study can be found.

[1] https://toonme.com/



(a) Cartoonish

(b) Painted

(c) Original

Figure 4.5: Illustration of the three distinct styles evaluated in the user study.

While the voices were not a primary focus of the study, we used a robotic voice generated with the help of a Website to discuss topics related to children and schools. The original version of the text focused on child abuse in schools. However, after the production of the video, it was discovered that the content had the potential to cause distress among participants. We aimed to create a positive feeling and avoid any references to child abuse that could negatively affect participants. Furthermore, the use of children's voices in the video was deemed ethically inappropriate, resulting in a lack of credibility. In addition, the text had to be read in a brief span of 10 seconds. These issues were resolved by modifying the text and replacing the robotic child sound. For the anchor condition video, the frequency of the audio was altered to create an unsettling effect, while still utilizing the same audio. The final text was as follows: ***"Today was a fantastic day at school. My teachers were supportive and encouraging, and my classmates were friendly. I was introduced to a lot of new concepts that really sparked my curiosity."***. Used audio both for Donya and Kian can be found below:

Kian's Sound

Figure 4.6: Kian's sound.

Donya's Sound

Figure 4.7: Donya's sound.

Items		Acronyms
Models	MakeItTalk	**M**
	PC-AVS (Pose Controllable)	**P**
	FOMM (First Order)	**F**
Character	Donya (girl)	**D**
	Kian (boy)	**K**
Style	Original	**O**
	Cartoony	**C**
	Painted	**P**

Table 4.3: Encoding Scheme.

Finally, 19 videos (1 anchor condition + 18 main videos) were generated with all the pictures using three different models. Table 4.3 contains the necessary abbreviations for each character, model, and style, which are required to tag the videos appropriately for analysis. Having these tags is crucial to be able to conduct a thorough analysis of the videos.

Table 4.4 is created to summarize all video combinations, and each video was assigned a specific code to facilitate analysis.

All the videos were uploaded to YouTube to provide links for the questionnaire. This is the YouTube channel link and all the videos can be found Here[2]. Screenshots of some of the videos can be seen in Figure 4.8.

[2]https://www.youtube.com/channel/UCsbhzj0XIHtkP3mXfr6ovcg

Model	Gender	Style	Video Name
M	D	P	**MDP**
M	D	C	**MDC**
M	D	O	**MDO**
P	D	P	**PDP**
P	D	C	**PDC**
P	D	O	**PDO**
F	D	P	**FDP**
F	D	C	**FDC**
F	D	O	**FDO**
M	K	P	**MKP**
M	K	C	**MKC**
M	K	O	**MKO**
P	K	P	**PKP**
P	K	C	**PKC**
P	K	O	**PKO**
F	K	P	**FKP**
F	K	C	**FKC**
F	K	O	**FKO**

Table 4.4: All possible combinations of video names

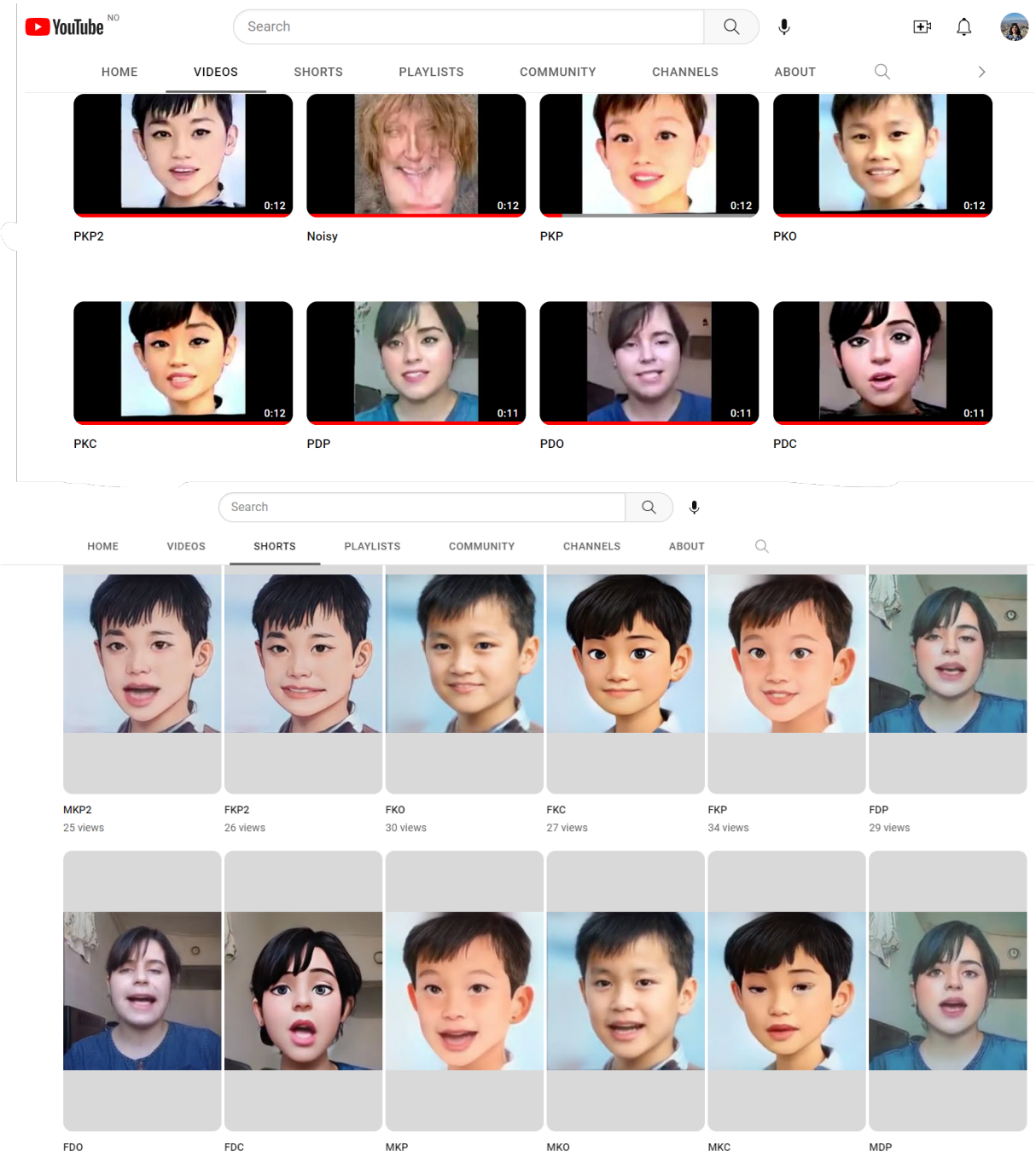


Figure 4.8: Screenshots of some of the videos on YouTube.

In addition to the 18 videos, an anchor condition, which was the first video, was added. An anchor condition is a technique used in research to ensure the accuracy and reliability of collected data. It involves introducing a reference point or a known answer that participants can use as a basis for their responses. This helps to minimize the influence of irrelevant factors, such as random or uninformative numbers, that may bias the responses of the participants (Chapman & Johnson, 1994).

In the user study conducted, an anchor condition was employed to validate the

responses collected from participants in a video questionnaire. The questionnaire aimed to assess the participants' perceptions of a video with poor sound quality.

By employing an anchor condition, we were able to enhance the validity and reliability of the collected data, which is essential for drawing meaningful conclusions from research studies. By introducing a reference point or a known answer, researchers can minimize the influence of irrelevant factors and ensure that participants provide honest and accurate responses.

4.1.4 Data cleaning

The pre-test questionnaire included a query that required the participants to transcribe the numbers they had heard from a YouTube video. This question was intended to assess the participants' ability to perceive the audio of the videos accurately. It was found that one of the participants had provided an incorrect response, and therefore, their data were excluded from the subsequent analysis.

Given the administration of two questionnaires, it was crucial that participants respond to both sets of queries. However, one participant failed to answer the second questionnaire. To account for this, two random questions were employed to retrieve the participant's response, which was then removed from both questionnaires. These pre-analysis procedures aimed to ensure a more precise and pristine dataset for further data analysis and discussion.

In summary, all the details about the questionnaires, both before and after participants answered them, have been presented in this chapter. The organization of the questionnaires and their materials was discussed, followed by a comprehensive elaboration of the questions and their individual purposes. The demographics of the participants and the methods employed to collect the answers are subsequently outlined.

Chapter 5

Outcome and Findings

The data gathered from users was analyzed using SPSS (Statistical Package for the Social Sciences), a powerful tool that enables the illustration and tabulation of data, and provides a more comprehensive understanding of it.[1]

SPSS is a software package widely used in social science research and statistical analysis. It allows researchers to manipulate and analyze large datasets with a wide variety of statistical techniques, including descriptive statistics, correlation analysis, regression analysis, factor analysis, and many others.Therefore, tables were generated through SPSS to discuss the results, which will be provided below. In this study, the GLM Repeated Measure test in SPSS was selected to examine the data, as it offers enhanced insights based on the specified input parameters. Repeated measures analysis in General Linear Model within SPSS involves the examination of data where the same subjects are measured under multiple conditions. This method takes into consideration the correlated nature of the repeated measurements within each subject, enabling the exploration of within-subject effects and interactions between factors. Repeated measures analysis in SPSS GLM is particularly valuable for studying the impact of different interventions, assessing variations across related conditions, or analyzing the effects of manipulated variables on the same subjects. It allows for the investigation of main effects and interactions between factors, providing a comprehensive understanding of how variables influence each other within the same subjects across various conditions.

Initially, different parameters were defined in the SPSS to conduct the study. Six main questions were asked about each video, each targeting an important feature of the talking head avatars. The overall quality of experience (**OverallQoE**) was measured, as well as the comfort level of users while watching the videos, which was

[1] https://www.ibm.com/products/spss-statistics

summarized as **Comfortable**. The remaining features that were assessed were **lipSync**, **Talking**, **HeadMovement**, and **videoQuality**.Table 4.1 contains the questions related to each feature, providing a comprehensive overview of the questions asked during the study. Following this, models, styles, and characters were defined to enable the analysis of data in relation to each of these parameters. This step was necessary to effectively analyze and interpret the data collected and to gain a more comprehensive understanding of the users' experiences with the talking head avatars. During the user study, participants were provided with choices ranging from "Bad" to "Excellent." However, to facilitate result analysis, it was necessary to convert these qualitative responses into numeric values. Accordingly, we assigned a value of 1 to "Bad," 2 to "Poor," 3 to "Fair," 4 to "Good," and 5 to "Excellent." By doing so, we were able to observe and interpret the results using numerical representations.

Additionally, all box plots in the following pages were created using the PowerBI software [2], known for its strong data visualization capabilities. The software's powerful dashboard allows for displaying data details effectively. The data was gathered from Microsoft Forms, provided in Excel format, and then imported into PowerBI. By transforming the data, various plots were generated using the PowerBI software. These box plots display the range of participants' insights, and the black dot inside each box represents the average answer provided by the users.

In our study, three main factors were focused on: models, characters, and styles, as mentioned earlier. For each factor, we had six different features to look at. The results obtained for each of these factors in relation to the other variables will be presented in the following pages. The objective is to identify and analyze the main effects of these factors within the context of our research.

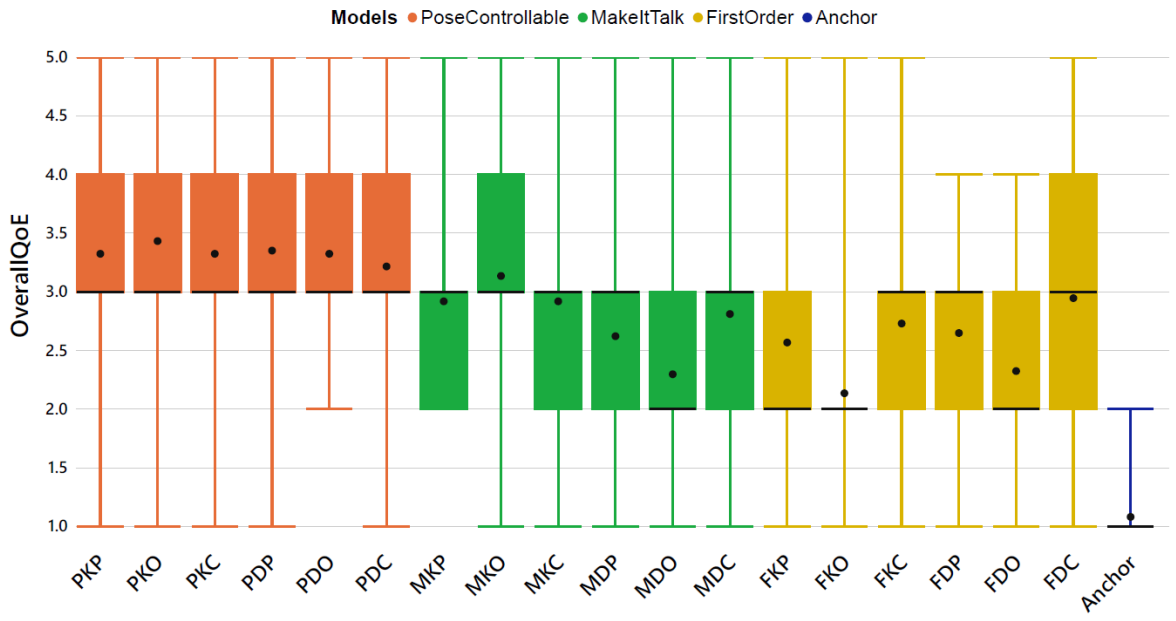
5.1 The main effect of the models

The box plots depicting the results of each video in terms of their models will be examined first. To maintain consistency in all the box plots, the FOMM models were represented by the color yellow, the MakeItTalk models were represented by green, and the PC-AVS models were represented by orange. This color scheme was used across all charts to ensure uniformity and ease of comparison. The users' satisfaction level for each video was rated from 1 to 5, with 1 indicating bad satisfaction and 5 indicating excellent satisfaction. In box plot 5.1, Part (a), it is evident that the PC-AVS model had better output than the other two models in term of OverallQoE feature. The MakeItTalk model followed closely behind, while the FOMM model had the lowest

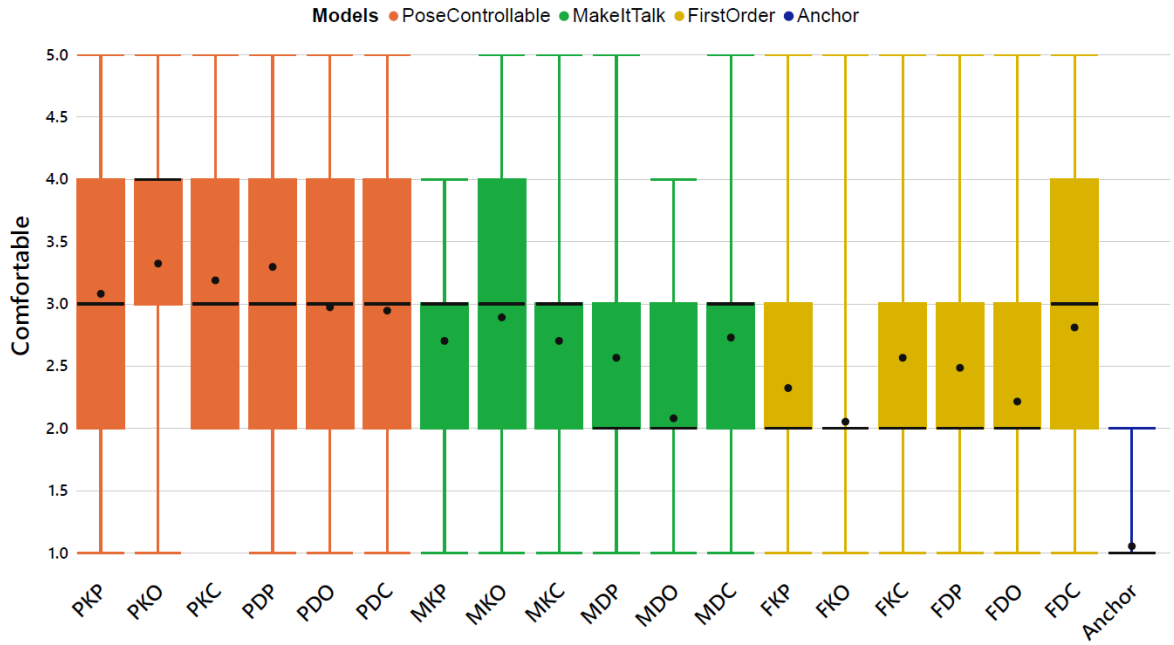
[2] https://powerbi.microsoft.com/en-us/

user satisfaction rating. It is important to highlight that the PKO video emerged as the winner in terms of the best overall Quality of Experience (QoE), as observed in the box plot. On average, it received a favorable rating of 3.43 from participants. This video was generated using the PC-AVS model and featured the character Kian (boy) with the Original style. Other videos created using the PC-AVS model showed minimal differences compared to the winning PKO video. The MKO video also performed well, obtaining a respectable score of 3.14. Similarly, the FDC video, produced by the FOMM model, achieved a decent rating of 2.95. However, the FKO video, which was a combination of the FOMM model, Kian character, and Original style, received the lowest score of 2.14. Furthermore, the PKO video received the highest comfortability rating of 3.32, closely followed by the second video, PDP (PC-AVS, Donya, Painted style), with only a slight difference between their comfortability scores. This can be seen in part (b) of the box plot (Figure 5.1). Once again, the FKO video had the lowest comfortability performance among all the videos, garnering a score of 2.05. The analysis conducted using SPSS corroborates the patterns observed in the Power BI box plots.

A repeated measures ANOVA was conducted to evaluate the effect of Model on OverallQoE scores. Mauchly's test indicated that the assumption of sphericity was violated, $\chi^2(2) = 4.418$, $p = 0.110$, so degrees of freedom were corrected using Huynh-Feldt estimates ($\varepsilon = 0.938$). There was a significant main effect of Model, $F(1.875, 67.508) = 57.677$, $p < 0.001$, partial $\eta^2 = 0.616$. Post hoc tests showed OverallQoE scores were significantly higher at PC-AVS compared to MakeItTalk ($p < 0.001$) and FOMM ($p < 0.001$). Scores at MakeItTalk and FOMM did not significantly differ. In terms of the Comortability feature, Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2) = 3.490$, $p = 0.175$, and therefore degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = 0.960$). The effect of Model on Comfortable scores was significant at the .05 level, $F(1.919, 69.091) = 38.038$, $p < 0.001$, partial $\eta^2 = 0.514$. Post-hoc pairwise comparisons with a LSD adjustment indicated that there was a significant difference between the Comfortable scores at MakeItTalk and FOMM ($p = 0.008$), Comfortable scores were significantly lower at MakeItTalk than at FOMM ($p = 0.008$), and Comfortable scores were significantly higher at PC-AVS than at FOMM ($p < 0.001$).



(a) OverallQoE



(b) Comfortable

Figure 5.1: Plots of Questionnaire-Based Features - OverallQoE and Comfortability

Regarding LipSync, box plot 5.2, part (a), illustrates that the participants perceived three videos to be almost identical. These videos were generated using the PC-AVS model and featured Kian, regardless of the style used. Their score was 3.68 which is a good score. However, the FKO video, which was generated using the FOMM model and featured the Original Kian style, had very poor performance in this aspect.

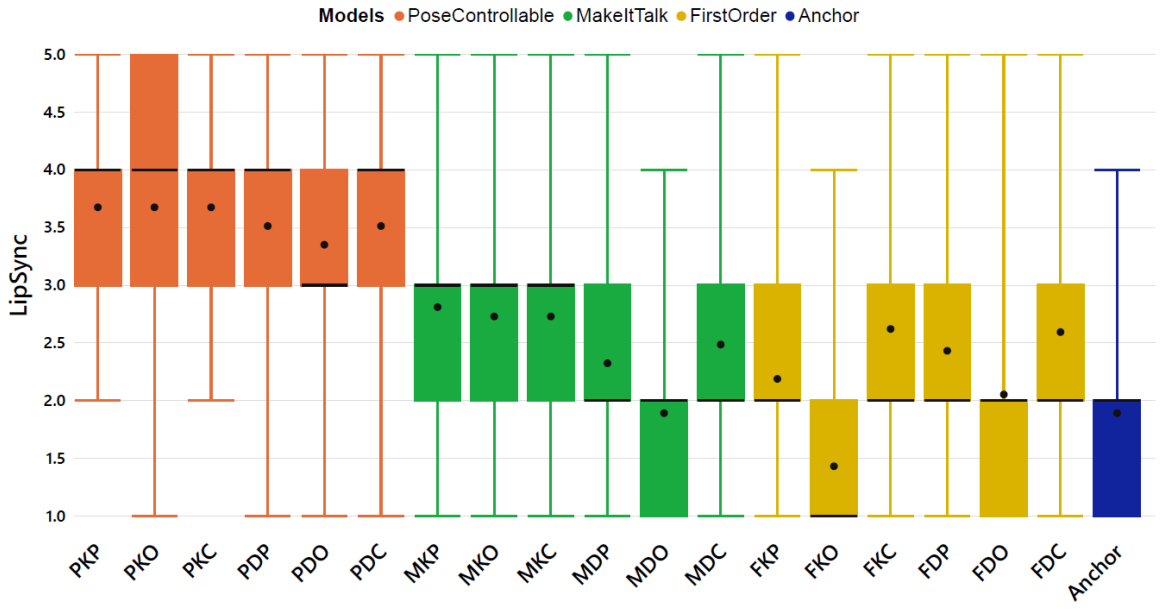
In the analysis conducted on the LipSync feature data through SPSS, the assumption of sphericity was found to be violated by Mauchly's test ($\chi^2(2) = 6.274$, $p = 0.043$), prompting us to adjust degrees of freedom using Huynh-Feldt estimates ($\varepsilon = 0.898$). A significant main impact stemming from Model emerged, with an F(1.796, 64.653) value of 73.959, $p < 0.001$, and a noteworthy partial η^2 effect size of .746. Follow-up assessments unveiled that LipSync scores were significantly higher in the PC-AVS compared to both MakeItTalk ($p < 0.001$) and FOMM ($p < 0.001$). Moreover, significant differences existed between ratings of MakeItTalk and FOMM ($p < 0.001$).

In terms of HeadMovement, the participants perceived the PDP video to have the best performance with a score of 3.27. This video was generated using the PC-AVS model and featured Painted Kian. This information is illustrated in box plot 5.2, Part (b). MDO which is the video made by MakeItTalk model, Donya character, and Original style had a very poor performance with a score of 2.14. Additionally, A repeated-measures ANOVA was conducted with the aim of comprehending the influence of Model on HeadMove scores. In this regard, Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 2.856$, $p = 0.240$. There was a significant main effect of Model, $F(2, 35) = 19.562$, $p < 0.001$, partial $\eta^2 = 0.352$. Post hoc tests showed HeadMove scores were significantly higher at PC-AVS compared to MakeItTalk ($p < 0.001$) and FOMM ($p < 0.001$). scores at MakeItTalk and FOMM did not significantly differ. In terms of the Talking feature, the box plot analysis revealed that the PKO video had the most positive impact on the participants, obtaining an impressive score of 3.5. Additionally, all the other videos created by the PC-AVS model outperformed the videos produced by the two other models. MKO, the original Kian made by MakeItTalk model, and FDC video which we know is Cartoony Donya made by FOMM model also did a good job among two other models. For quality factor Talking, the repeated-measures ANOVA yielded nearly an identical results. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 0.860$, $p = 0.650$ that prompting us to adjust degrees of freedom using Huynh-Feldt estimates ($\varepsilon = 1.000$). There was a significant main effect of Model, $F(2, 35) = 48.506$, $p < 0.001$, partial $\eta^2 = 0.574$. Post hoc tests showed Talking scores were significantly higher at PC-AVS model compared to MakeItTalk model ($p < 0.001$) and FOMM ($p < 0.001$). Scores at MakeItTalk model and FOMM did not significantly differ. The last feature to be evaluated in this study is the quality of the videos. It was observed that PKO

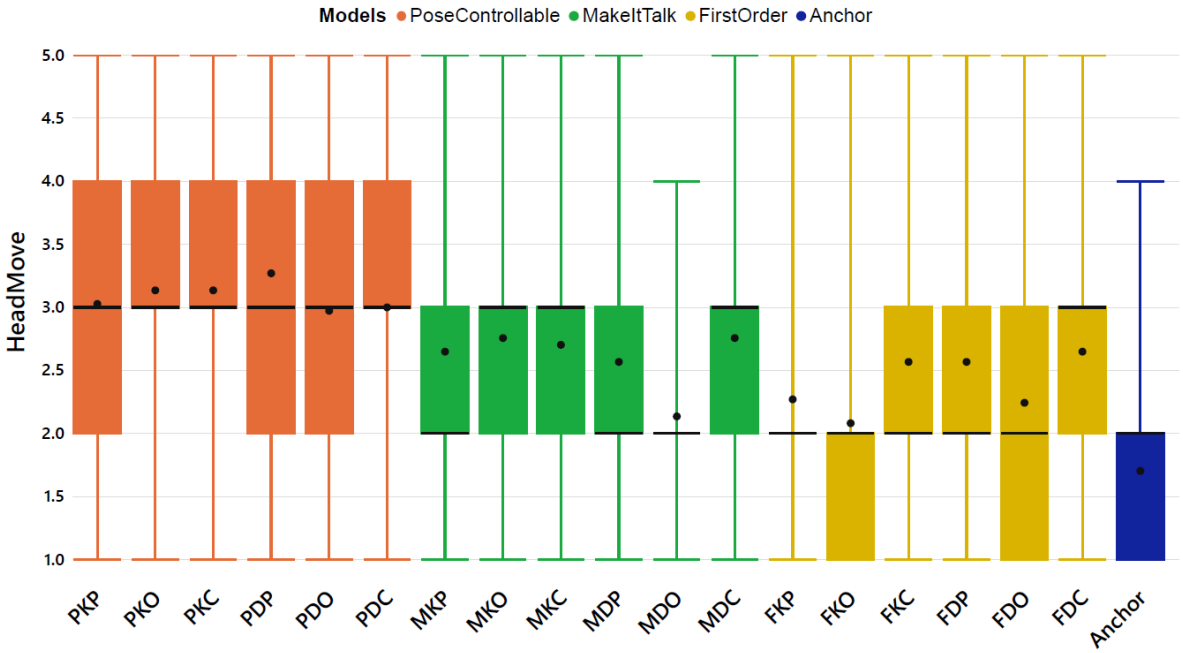
received the highest quality rating from the participants, with an average score of 3.43. Other videos generated by the PC-AVS model also received positive feedback from the participants, indicating good satisfaction. This can be seen in the box plot depicted in Figure 5.3, with Parts (a) and (b) showing the distribution of quality scores. By looking closely at analysis coming from a repeated-measures ANOVA to evaluate the effect of Model on VideoQuality scores, Mauchly's test indicated that the assumption of sphericity was violated, $\chi^2(2) = 5.826$, $p = 0.054$, so degrees of freedom were corrected using Huynh-Feldt estimates ($\varepsilon = 0.907$). There was a significant main effect of Model, $F(1.814, 65.308) = 60.269$, $p < 0.001$, partial $\eta^2 = 0.626$. Post hoc tests showed VideoQuality scores were significantly higher at PC-AVS model compared to MakeItTalk ($p < .001$) and FOMM ($p < .001$). Scores at MakeItTalk model and FOMM model did not significantly differ. In summary, for an overarching perspective on the videos' performance concerning user experience, it is beneficial to refer to Table 5.1 which presents the mean and standard deviation of data across all six features.

	MakeItTalk Model		**FOMM Model**		**PC-AVS Model**	
Features	Mean	StandardDeviation	Mean	StandardDeviation	Mean	StandardDeviation
LipSync	1.89	0.81	2.49	1.02	2.32	0.94
Comfortable	2.61	0.13	2.41	0.12	3.14	0.15
Talking	2.61	0.13	2.37	0.12	3.22	0.14
HeadMove	2.60	0.13	2.40	0.14	3.09	0.14
VideoQuality	2.66	0.13	2.44	0.12	3.27	0.14
OverallQoE	2.78	0.13	2.56	0.12	3.33	0.13

Table 5.1: Means and Standard Deviations of Features by Models

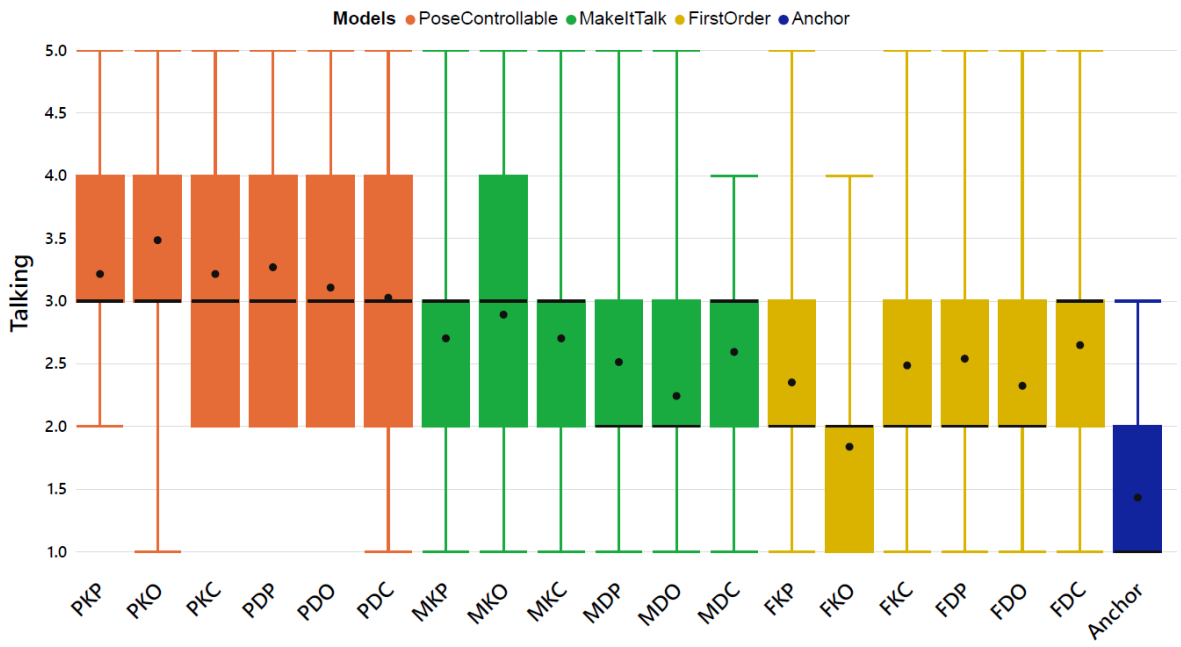


(a) LipSync

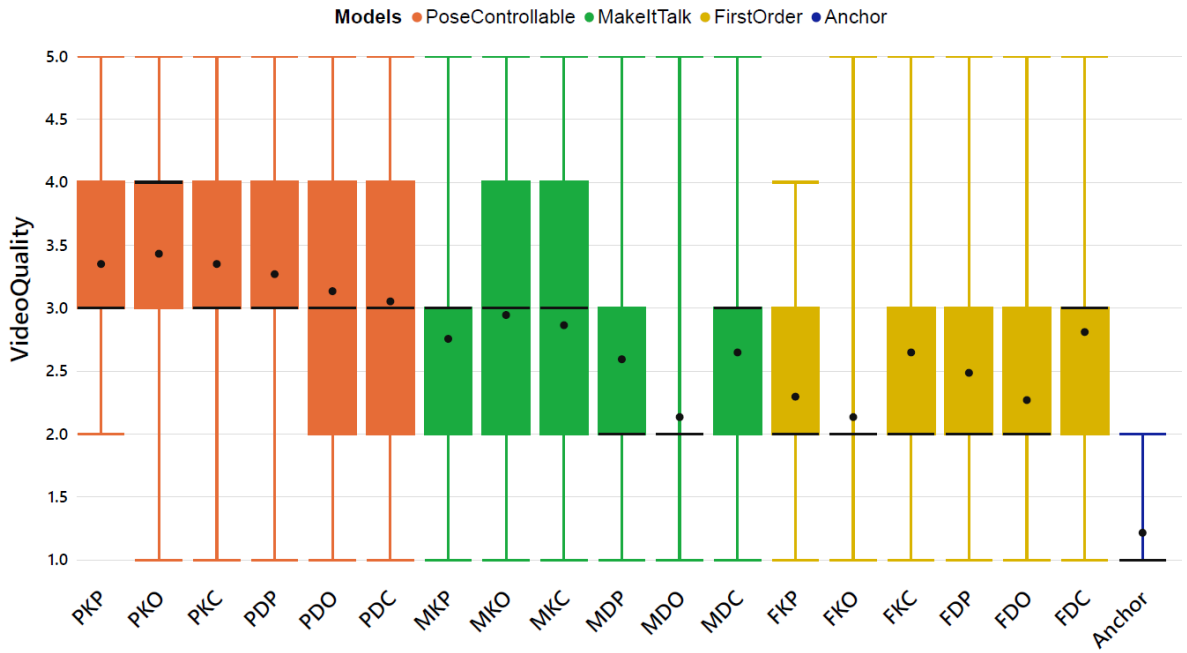


(b) Head Move

Figure 5.2: Plots of Questionnaire-Based Features - LipSync and HeadMove



(a) Talking



(b) Video Quality

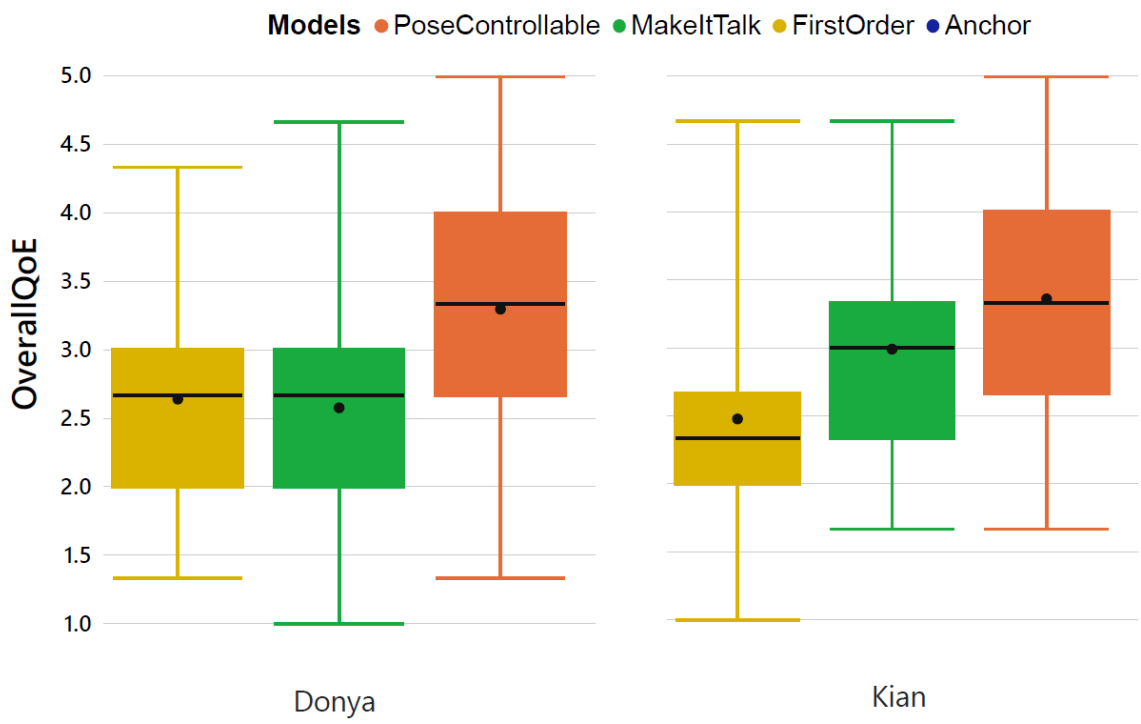
Figure 5.3: Plots of Questionnaire-Based Features - Talking and VideoQuality

5.2 The main effect of the characters

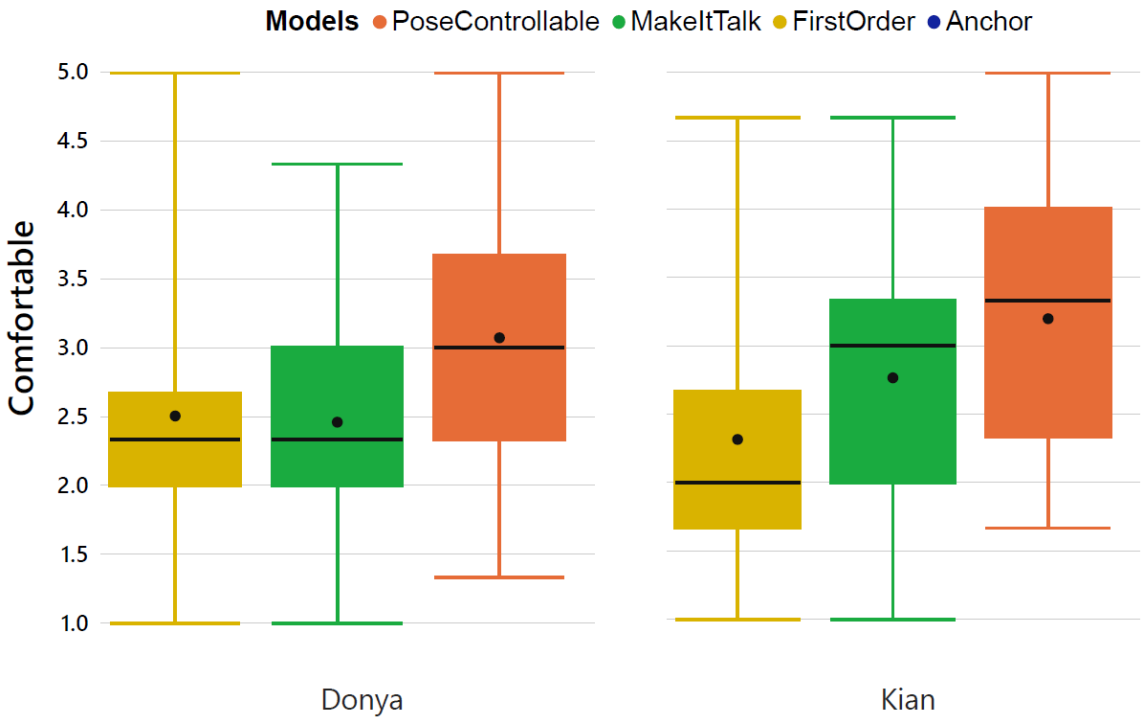
As mentioned before, two characters, Kian (boy) and Doya (girl), were used in the videos. The box Plot 5.4 (a) displays the OverallQoE of the participants towards the videos made with each character separately. It can be observed that there was not much difference between the two characters in terms of the final result.In the PC-AVS model, an average score of 3.36 was achieved by Kian, while Donya obtained a slightly lower average score of 3.30. In the MakeItTalk model, Kian exhibited a better performance compared to Donya. However, in the FOMM model, both Kian and Donya demonstrated identical performance levels.

Examining the results derived from the repeated measures ANOVA analysis, it also approves that the main effect of the Character variable did not yield statistical significance in the context of the OverallQoE feature ($F(1, 36) = 2.235$, $p = 0.144$, partial $\eta^2 = 0.058$). Regarding comfortability, individuals tend to experience slightly more comfort with Kian in the PC-AVS and MakeItTalk model, whereas Donya performed better in the FOMM model. These findings can be observed in the corresponding charts of our analysis 5.4 (b). However, in the repeated measure ANOVA these differences were shown to not be significantly different. The main effect of Character was not significant, $F(1, 36) = 0.225$, $p = 0.638$. This indicates there was no significant difference in comfortability scores between the two characters overall, when collapsing across model and style.

In terms of the lipsync feature, Kian exhibited better performance in both the PC-AVS and MakeItTalk models, with scores of 3.68 and 2.76, respectively. Donya also performed well, achieving a score of 3.46 in the PC-AVS model, but showed weaker performance in the MakeItTalk model. The lowest lipsync score was 2.08, which was attributed to Kian in the FOMM model. These findings can be observed 5.5 Box plot part (a). However, the repeated measure ANOVA showed the main effect of Character on quality factor LipSync was not significant, $F(1, 36) = 2.814$, $p = 0.102$, partial $\eta^2 = 0.073$. In the HeadMove feature, both Kian and Donya demonstrated nearly equal performance in the PC-AVS and FOMM models, achieving scores of 3.10 in PC-AVS. However, in the MakeItTalk model, Kian exhibited slightly better performance. These findings are illustrated in Box plot (b) of Figure 5.5. By looking at the values that come from repeated measure ANOVA, the main effect of Character was not significant in this feature as expected ($F(1, 36) = 0.035$, $p = 0.853$, partial $\eta^2 = 0.001$.).

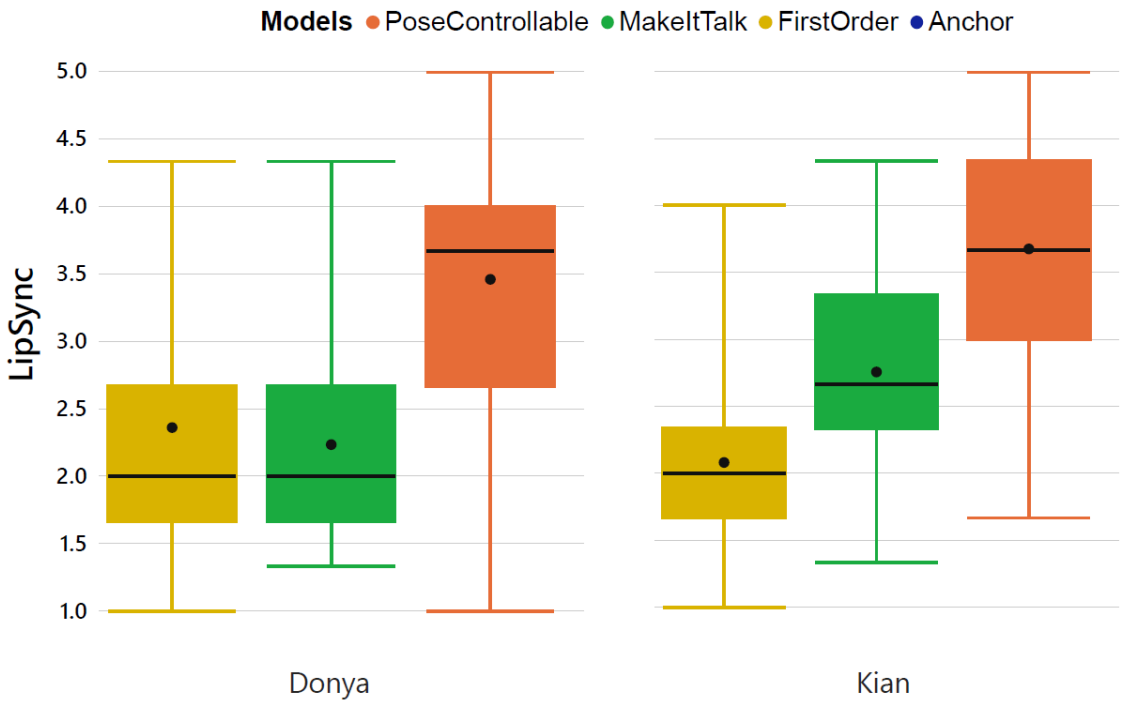


(a) OverallQoE

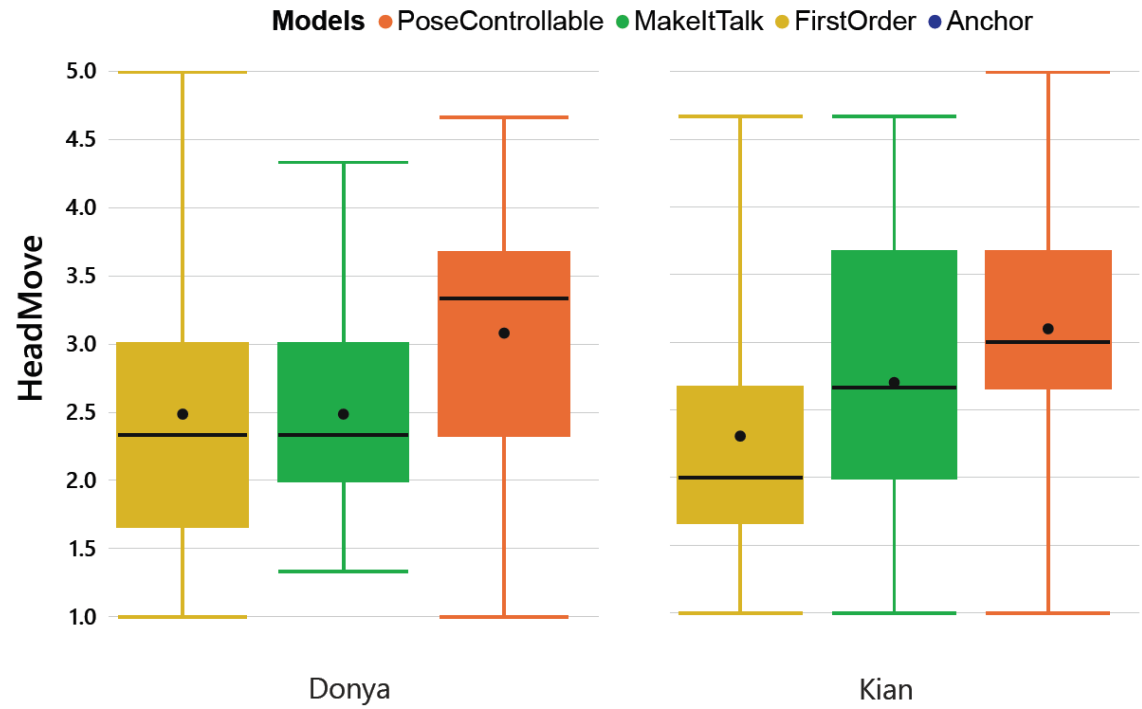


(b) Comfortable

Figure 5.4: The main effect of the characters - OverallQoE and Comfortability

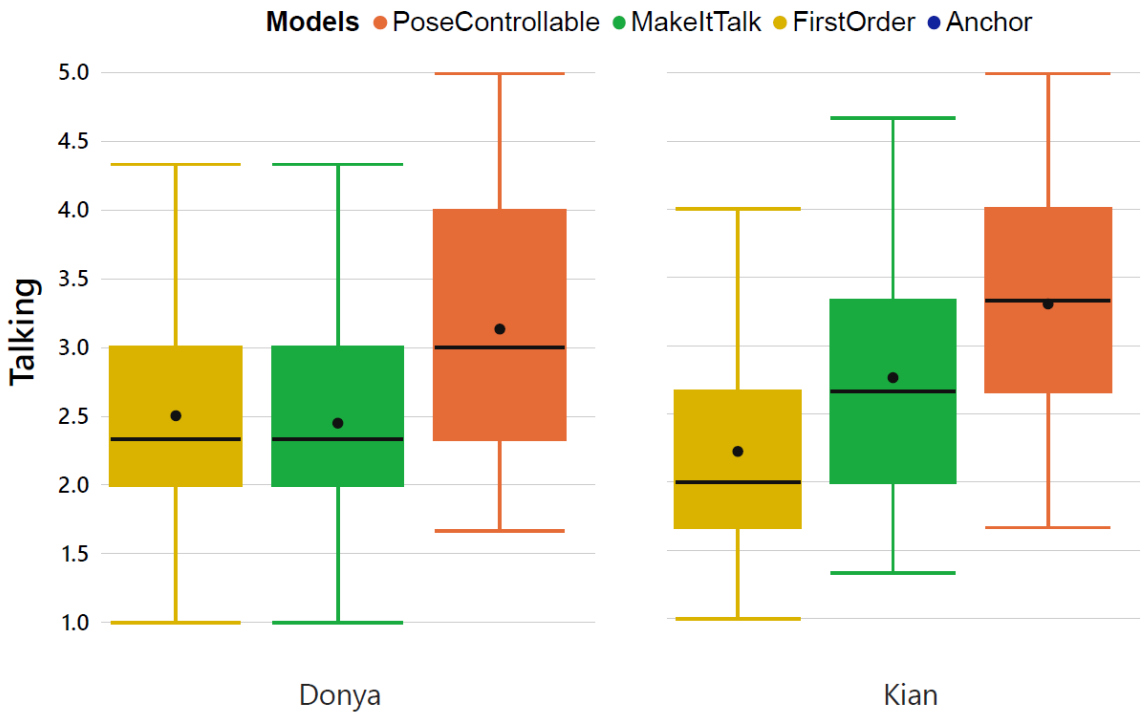


(a) LipSync

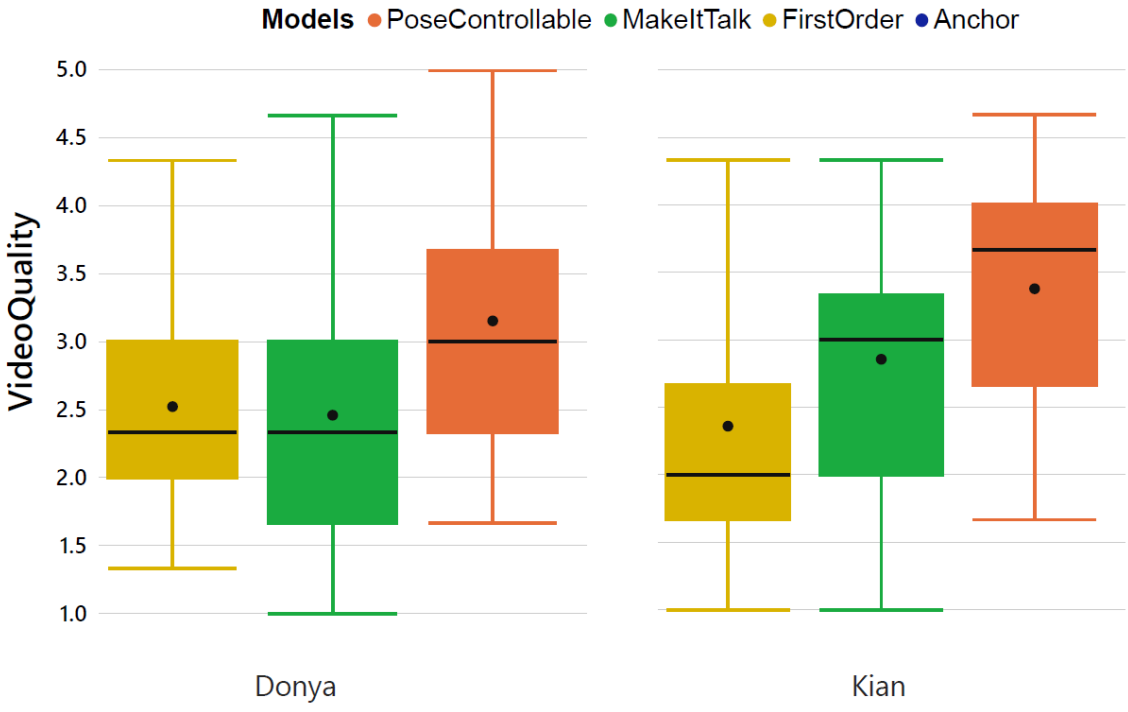


(b) Head Move

Figure 5.5: The main effect of the characters - LipSync and HeadMove



(a) Talking



(b) Video Quality

Figure 5.6: The main effect of the characters - Talking and VideoQuality

Regarding the talking feature, the box plots indicate that individuals tend to have a more comfortable experience with Kian when using the PC-AVS and MakeItTalk models. For example, Kian achieved a score of 3.31 for the PC-AVS model and 2.77 for the MakeItTalk model. On the other hand, Donya showed a better performance in the FOMM model with a score of 2.5, compared to Kian's score of 2.2. An examination of the data using repeated measures ANOVA also sheds light on the fact that the main effect of Character was not significant, F(1, 36) = 0.223, p = 0.639, partial η^2 = 0.006. The final feature we will explore is the video quality. The results for this aspect were consistent with the previous features, where the PC-AVS model emerged as the winner for both characters. Kian, in particular, performed exceptionally well with a score of 3.31. The MakeItTalk model also demonstrated good performance in terms of video quality, achieving a score of 2.77 for Kian. These findings are visually represented in the corresponding box plots (a) and (b) of our analysis 5.6. Actually, the findings obtained through the utilization of repeated measures ANOVA also demonstrate that the main effect of Character was not significant, F(1, 36) = 3.493, p = 0.070, partial η^2 = 0.088.

In summary, the table 5.2 shows the means and standard deviations for the 6 evaluated features across the two characters. For most features, the means are fairly similar between the two characters. The biggest difference is for OverallQoE, where Kian has a higher mean rating (2.93) compared to Donya (2.85). This indicates participants rated the overall quality of experience for Kian as mildly better than for Donya. For other features like LipSync, Comfortable, Talking, HeadMove, and VideoQuality the means are very close between the two characters, with at most a 0.13 difference in means. The standard deviations are also comparable between the two characters, showing a similar amount of variance in scores for both Donya and Kian.

	Donya		**Kian**	
Features	Mean	StandardDeviation	Mean	StandardDeviation
LipSync	2.70	0.81	2.82	1.05
Comfortable	2.71	0.13	2.73	0.13
Talking	2.72	0.12	2.75	0.12
HeadMove	2.70	0.13	2.69	0.12
VideoQuality	2.75	0.12	2.83	0.13
OverallQoE	2.85	0.12	2.93	0.12

Table 5.2: Means and Standard Deviations of Features by Characters

5.3 The main effect of the Styles

In this user study, three different styles were used including Original, Painted, and Cartoony as mentioned before. In the feature of overallQoE, it can be observed that the PC-AVS model exhibited consistent performance across all three styles. The slight variation in performance was observed in the original style with a score of 3.38, but the difference is negligible. Similarly, the MakeItTalk model demonstrated uniform performance across all styles, while the FOMM model showed lower performance in the original style with a score of 2.23.

The outcomes emerging from the analysis conducted via repeated measures ANOVA revealed that there was a significant main effect of Style on quality factor overallQoE, $F(2, 35) = 5.368$, $p = 0.010$, partial $\eta^2 = 0.130$. Regarding the comfortability factor, the results align with the findings from previous features, which are depicted in the box plots shown in figure 5.7 parts (a) and (b). There was a significant main effect for Style, $F(2, 35) = 4.068$, $p = 0.026$, partial $\eta^2 = 0.189$ regarding this feature.

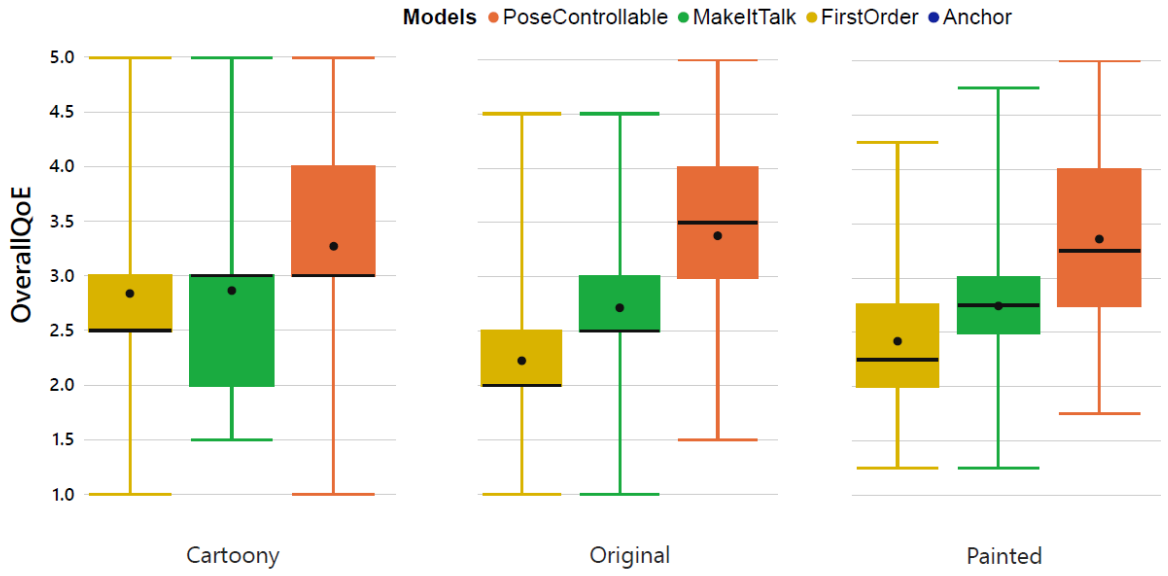
In the lipsync factor, all three styles exhibited similar performance, achieving an average score of approximately 3.5 in the PC-AVS model. However, in the MakeItTalk model, the cartoony style demonstrated slightly better performance with a score of 2.61. Interestingly, the FOMM model also achieved the same score of 2.61 in the cartoony style. However, the FOMM model performed poorly in the original style, obtaining a score of 1.74, which was one of the lowest performances among all features, styles, characters, and models. The repeated measure ANOVA shows a trend for the main effect of Style on lipsync, with $F(2, 35) = 3.218$, $p = 0.052$, and a partial η^2 of 0.155. It's evident that the p-value closely approaches the significance threshold.

In the HeadMove feature, the cartoony style exhibited slightly better performance across all models, although the differences were very minimal. The scores for the three styles in the three models were very close to each other in this feature. For a better understanding of the results, all the details are presented in the figure 5.8 part (b). By looking at the repeated measure ANOVA analysis, there was no significant main effect for Style, $F(2, 35) = 2.215$, $p = 0.147$, partial $\eta^2 = .058$ in term of the HeadMove feature. In the next two features, which are talking and video quality, the PC-AVS model demonstrated better performance. In the video quality feature, the painted style obtained a slightly higher score compared to the other two styles, while in the talking feature, the original style outperformed the others. The MakeItTalk model exhibited consistent performance in both features and across all three styles. However, the FOMM model showed weak performance in the original style for both features. The evidence supporting this observation can be found in the box plots provided in

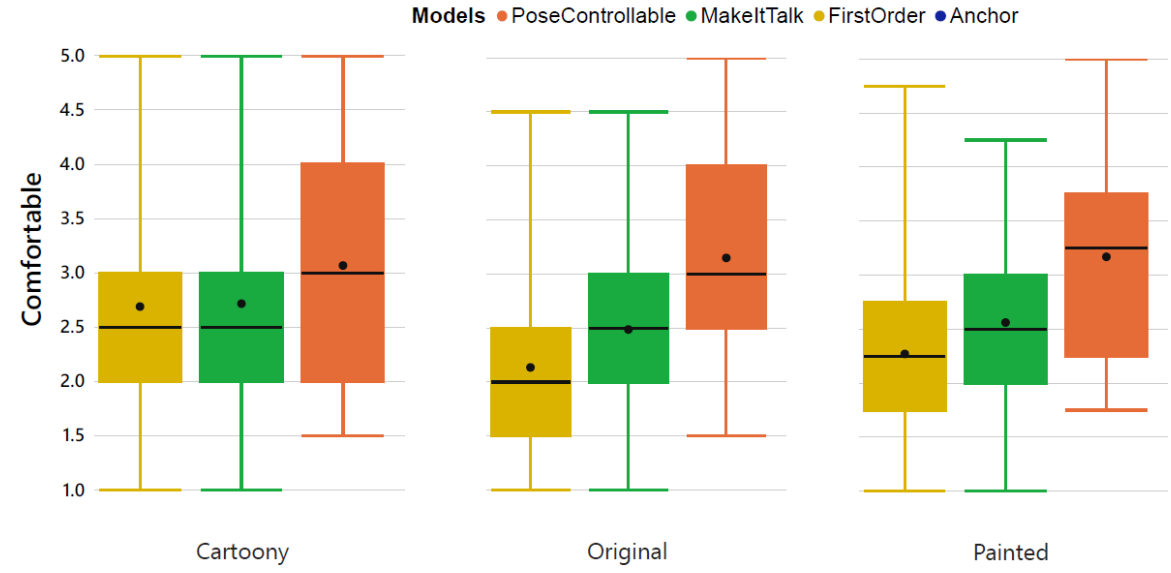
Figures 5.9 part (a) and (b). Looking at the results come from SPSS, no substantial primary impact was observed for Style, as indicated by the results of F(2, 35) = 2.473, p = 0.116, with a partial η^2 of 0.064, concerning the videoQuality feature. Similarly, for the talking feature, no significant primary effect for Style was apparent, with F(2, 35) = 2.714, p = 0.091, and a partial η^2 of 0.070. The following table 5.3 presents the means and standard deviations of various features across different styles: Cartoony, Painted, and Original. Each cell displays the mean and standard deviation values for a specific feature under each style.

	Cartoony		**Painted**		**Original**	
Features	Mean	StandardDeviation	Mean	StandardDeviation	Mean	StandardDeviation
LipSync	2.43	0.98	2.54	1.04	2.57	0.95
Comfortable	2.49	0.10	2.63	0.14	2.73	0.14
Talking	2.51	0.12	2.60	0.14	2.76	0.14
HeadMove	2.53	0.13	2.63	0.14	2.72	0.13
VideoQuality	2.51	0.12	2.63	0.13	2.76	0.13
OverallQoE	2.60	0.12	2.74	0.13	2.83	0.13

Table 5.3: Means and Standard Deviations of Features by Different Styles

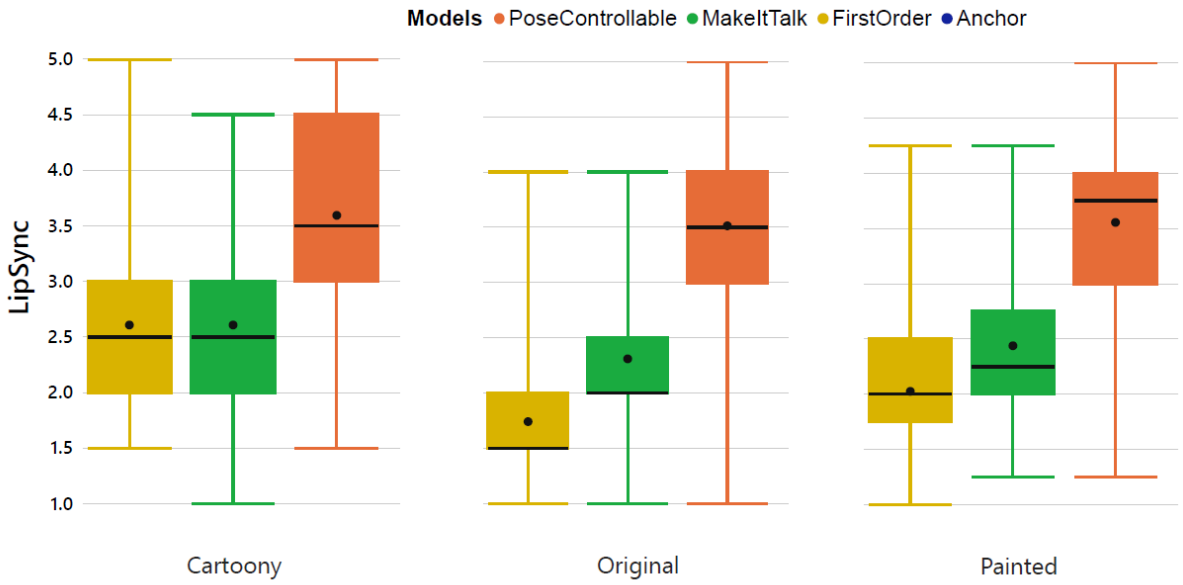


(a) OverallQoE

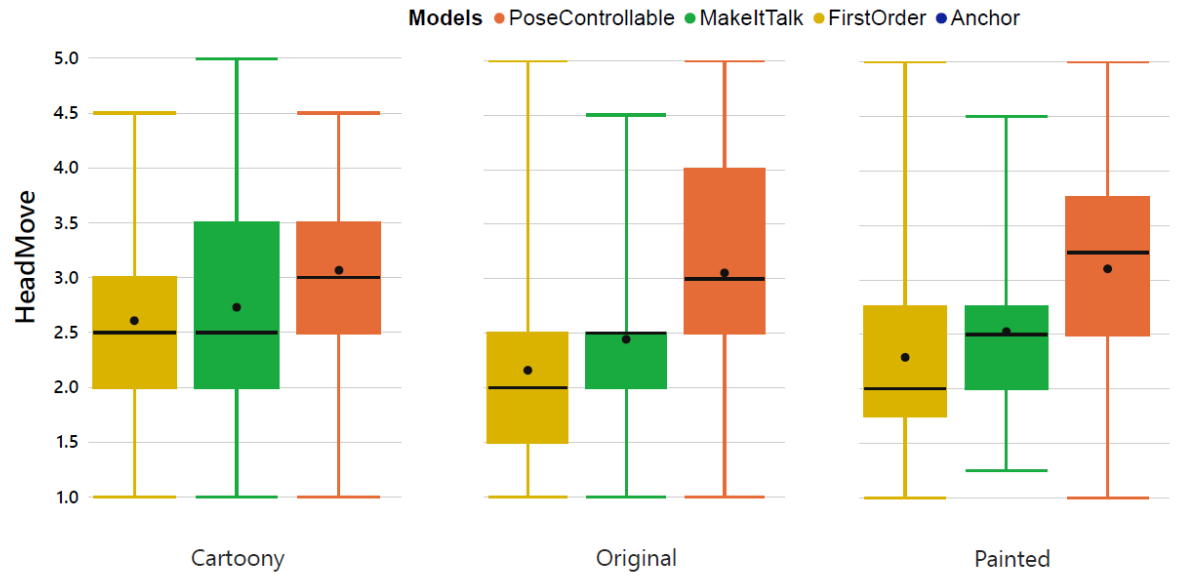


(b) Comfortable

Figure 5.7: The main effect of the Styles - OverallQoE and Comfortability

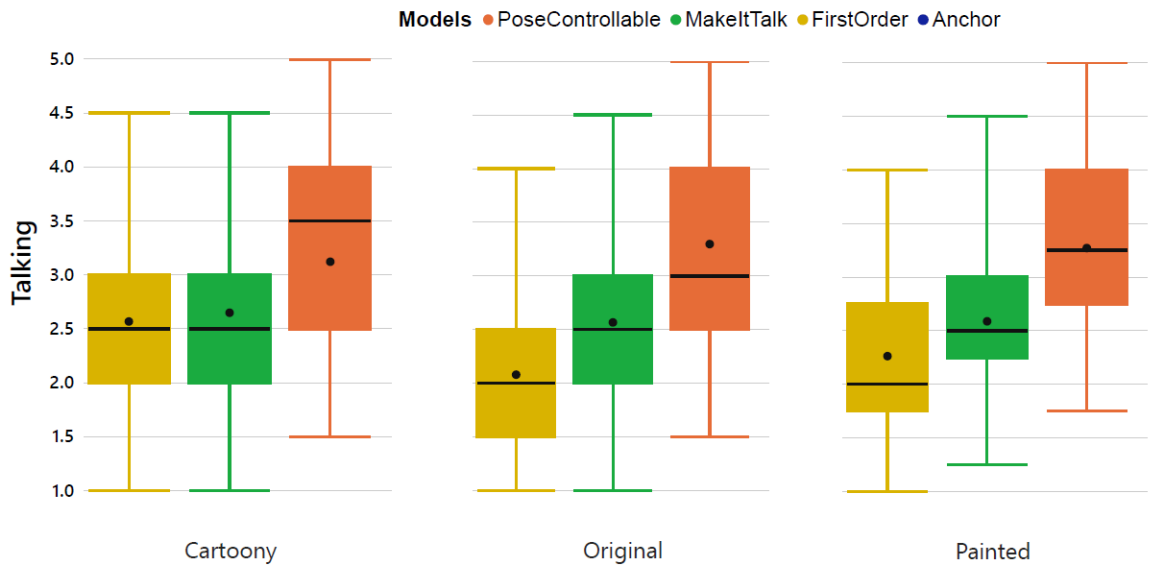


(a) LipSync

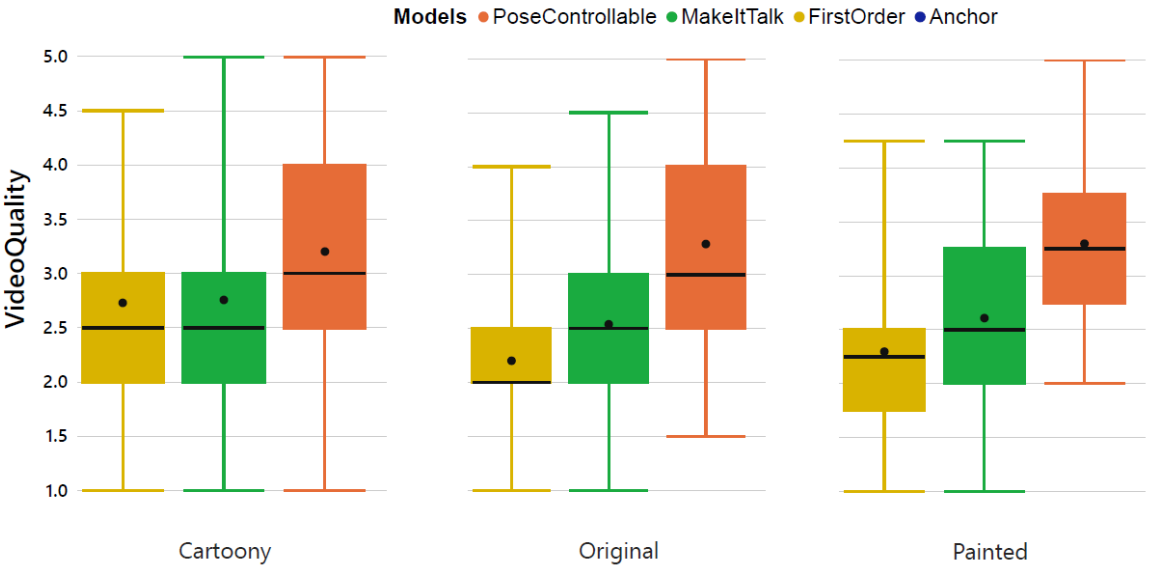


(b) Head Move

Figure 5.8: The main effect of the Styles - LipSync and HeadMove



(a) Talking



(b) Video Quality

Figure 5.9: The main effect of the Styles - Talking and VideoQuality

5.4 The interactions among Models, Characters, and Styles

Following an examination of the primary effects associated with individual models, characters, and styles, our attention turns towards comprehending the significance encompassing the intricate interplay between these variables. This comprehensive analysis is facilitated through the application of repeated measure ANOVA within the SPSS framework. The insights drawn from this analysis are a culmination of meticulous investigations performed for each distinctive attribute. Subsequently, an intricate breakdown of each feature ensues, commencing with the evaluation of "overallQoE." Within this context, it is imperative to highlight that the interaction between Model and Character yielded a notable degree of significance, with an F(2, 35) = 12.487, $p < 0.001$, along with a partial η^2 value of 0.258. Correspondingly, the interaction between Model and Style also exhibited significance, as indicated by an F0(4, 33) = 5.368, p = 0.002, accompanied by a partial η^2 value of 0.157. Furthermore, the interaction between Character and Style yielded a significant outcome, underscored by an F(2, 35) = 6.002, p = 0.006, and a partial η^2 value of 0.255. However, it's paramount to recognize that the pinnacle of significance was achieved within the three-way interaction encompassing Model, Character, and Style. This interaction was underscored by a compelling F(4, 33) = 4.810, p = 0.004, along with a substantial partial η^2 value of 0.368. It's intriguing to note that while the character factor alone did not exhibit a significant effect, its interactions with other components such as models or styles magnified its importance and led to notable impacts on users' experiences. This phenomenon underscores the intricate nature of user perception, where seemingly subtle interactions between variables can yield substantial and meaningful outcomes in shaping the overall user experience.

Analyzing the impact of head movement in our study, the interaction between Model and Character emerged as significant, yielding an F(2, 35) = 3.569, p = 0.040, with a partial η^2 value of 0.090. However, the interactions between Model and Style, as well as Character and Style, did not exhibit significance. Interestingly, the highest level of significance was achieved through the three-way interaction involving Model, Character, and Style. Here, the F-statistic recorded a compelling value of (4, 33) = 6.035, p = 0.001, with a substantial partial η^2 value of 0.422. These findings underscore how the collaborative influence of these variables plays a pivotal role in shaping users' perceptions and engagement in relation to head movement within talking head avatars.

Examining the Talking feature, we observed that the interaction between Model and

Character yielded significant results, with an F(2, 35) = 8.445, p = 0.001, and a partial η^2 value of .190. However, the Model * Style interaction did not exhibit significance. Notably, the Character * Style interaction displayed significance, underlined by an F(2, 35) = 5.209, p = 0.010, and a partial η^2 value of 0.229. Conversely, the three-way interaction involving Model, Character, and Style did not attain significance, as indicated by an F(4, 33) = 3.747, p = 0.013, and a partial η^2 value of 0.312.

For the analysis of VideoQuality, the interaction between Model and Character yielded remarkable significance, marked by an F(2, 35) = 11.020, p < 0.001, and a partial η^2 value of 0.234. Similarly, the Model * Style interaction displayed significance, with an F-statistic of (4, 33) = 3.398, p = .028, and a partial η^2 value of 0.273. Furthermore, the Character * Style interaction revealed significance, denoted by an F(2, 35) = 6.892, p = 0.003, and a partial η^2 value of 0.283. The three-way interaction encompassing Model, Character, and Style achieved considerable significance, recording an F(4, 33) = 4.119, p = 0.008, and a partial η^2 value of 0.333. These findings highlight the intricate interplay between these variables in influencing users' perception of video quality in talking head avatars.

Analyzing the LipSync feature, the interaction between Model and Character yielded profound significance, with an F(2, 35) = 15.612, p < 0.001, and a substantial partial η^2 value of 0.471. Similarly, the Model * Style interaction exhibited significance, as indicated by an F(4, 33) = 6.012, p = 0.001, and a partial η^2 value of 0.422. The Char * Style interaction also yielded noteworthy significance, marked by an F(2, 35) = 12.159, p < 0.001, and a partial η^2 value of 0.410. The highest level of significance, however, was achieved through the three-way Model * Character * Style interaction, with an F(4, 33) = 10.126, p < 0.001, and a substantial partial η^2 value of 0.551. These findings underscore the collaborative significance of these factors in shaping users' perceptions of lip sync accuracy within talking head avatars.

Finally, examining the feature of comfortability, the interaction between Model and Character demonstrated significant influence, with an F(2, 35) = 9.627, p < 0.001, and a noteworthy partial η^2 value of 0.355. Similarly, the Model * Style interaction exhibited significance, marked by an F(4, 33) = 5.032, p = 0.003, and a partial η^2 value of 0.379. Furthermore, the Char * Style interaction displayed significance, with an F(2, 35) = 5.209, p = 0.010, and a partial η^2 value of 0.229. The three-way interaction involving Model, Character, and Style also attained significance, as indicated by an F(4, 33) = 5.610, p = 0.001, and a substantial partial η^2 value of 0.405. These findings emphasize the collective impact of these variables on users' perceived comfort when interacting with talking head avatars.

This concise summary table encapsulates the findings from our analyses of various

avatar features. For a comprehensive overview of the significance levels, F-statistics, p-values, and partial η^2 values related to the interactions between Models, Characters, and Styles, look at table 5.4

Avatar Feature	Interaction Significance	F-Statistic	p-Value	Partial η^2 Value
HeadMovement	Model * Character	3.569	.040	.090
	Model * Style	Not Significant	-	-
	Character * Style	Not Significant	-	-
	Model * Character * Style	6.035	.001	.422
Talking	Model * Character	8.445	.001	.190
	Model * Style	Not Significant	-	-
	Character * Style	5.209	.010	.229
	Model * Character * Style	Not Significant	-	-
VideoQuality	Model * Character	11.020	<.001	.234
	Model * Style	3.398	.028	.273
	Character * Style	6.892	.003	.283
	Model * Character * Style	4.119	.008	.333
LipSync	Model * Character	15.612	<.001	.471
	Model * Style	6.012	.001	.422
	Character * Style	12.159	<.001	.410
	Model * Character * Style	10.126	<.001	.551
Comfortability	Model * Character	9.627	<.001	.355
	Model * Style	5.032	.003	.379
	Character * Style	5.209	.010	.229
	Model * Character * Style	5.610	.001	.405
OverallQoE	Model * Character	12.487	<.001	.258
	Model * Style	5.368	.002	.157
	Character * Style	6.002	.006	.255
	Model * Character * Style	4.810	.004	.368

Table 5.4: Interactions Summary for Avatar Features

5.5 Analyzing the Post-Test Questionnaire

One of the goals of this study was to examine the impact of the uncanny valley phenomenon on different types of avatar videos. For this purpose, three distinct styles were purposely chosen to obtain a comprehensive understanding of how the avatars were perceived and emotionally responded to by the participants.

The participants were asked the following question in the final questionnaire: "Generally, which avatars did have the most realistic appearance for you?" To present

the responses clearly, a donut plot was utilized 5.10, revealing that a significant 61% of the participants favored the original style, perceiving it as the most realistic compared to the other two styles.

The second general question inquired about the participants' overall preference regarding the avatars: "Generally, which avatars do you like the most?". Understanding users' emotional responses and preferences towards avatar styles holds significance, as these factors can influence trust and engagement during conversations. Surprisingly, 41% of the participants expressed a distinct fondness for the cartoony avatars, showcasing a notable departure from concerns related to the uncanny valley phenomenon. Additionally, approximately 39% of participants indicated a strong liking towards the original avatars, revealing a preference that closely rivalled the appeal of the cartoony style. These findings, presented in Figure 5.11 as a donut plot.

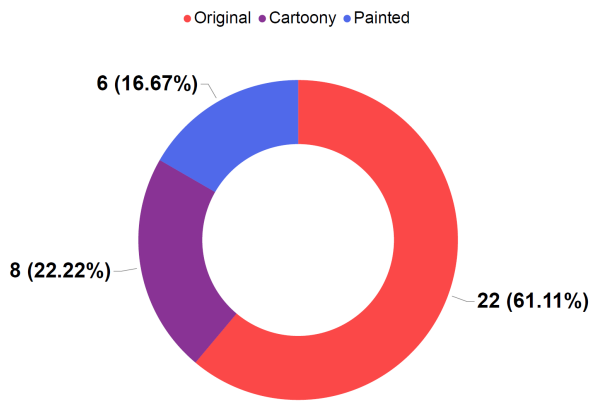


Figure 5.10: Perceived realism of avatars

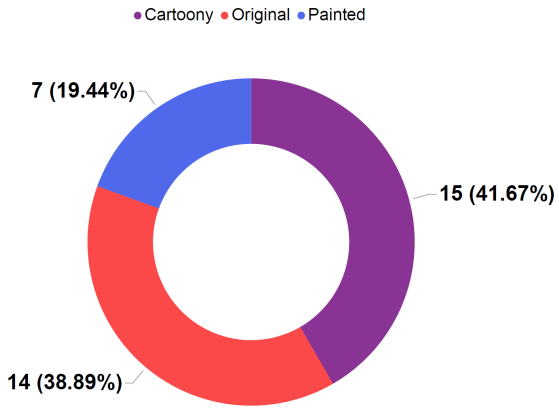


Figure 5.11: Overall favorite avatars

The third question, "For a conversation (talking to a computer), which avatars do you prefer to use?", inquired about the participants' preference for engaging in conversations with the different avatars. Our main goal was to determine which avatars they found most preferable for interactive communication with a computer. Interestingly, even though participants expressed a stronger liking for the cartoony avatars, a significant majority (66%) actually favored the original avatars when it came to engaging in conversations. This finding shed light on the participants' practical inclinations in selecting avatars for interactive purposes. The results are visually presented in Figure 5.12.

The next question asked participants, "Considering the avatars' talking, which were the most realistic?" This question focused on how real the avatars felt during

conversations, not just their appearance. Interestingly,, 52% of participants found the cartoony avatars to be the most realistic, while 25% preferred the original pictures, and 22% liked the painted style. The distribution of responses is shown in Figure 5.13, giving us the insight into how participants perceived the realism of different avatar styles during interactions.

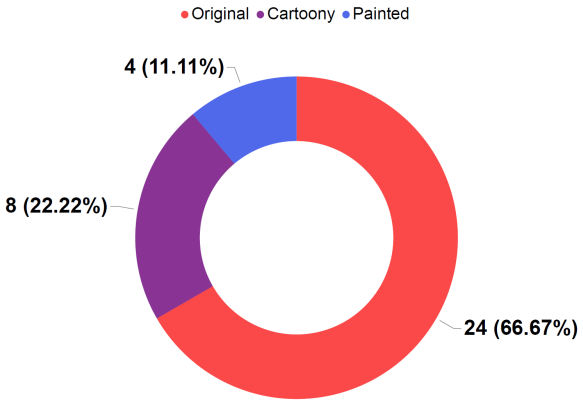


Figure 5.12: Preferred avatars for computer-mediated conversations

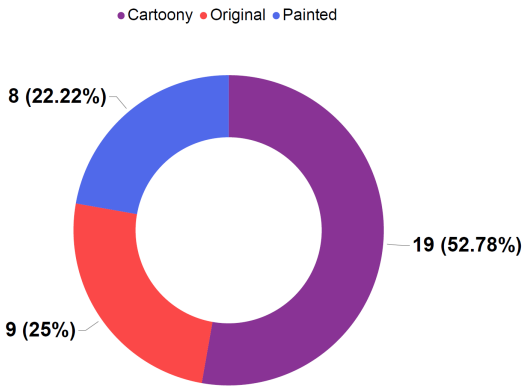


Figure 5.13: Perceived realism of avatars during speech

The last question, "Generally, which avatars were the most believable for you?", held significant importance in achieving the goals of this study. Given the application of our research in training psychologists and professionals to effectively communicate and interview children who have experienced trauma or injury, the believability of avatars becomes a crucial factor. The findings revealed that 52% of the participants found the original style to be the most believable, followed by 30% who favored the cartoony style, and 17% who found the painted style to be more believable. Figure 5.14 illustrates the distribution of responses among the participants, shedding light on their perceptions of avatar believability.

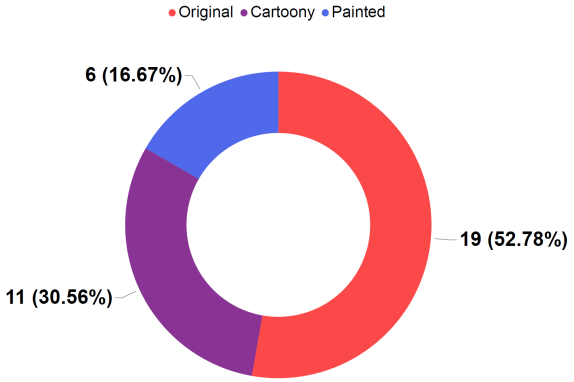


Figure 5.14: Most believable avatars

In summary, this chapter has presented all the obtained results from the questionnaires, offering a dual perspective. One approach involved the utilization of Power BI box plots to enhance the visualization of participant responses. However, the more significant aspect encompassed the analysis of answers through repeated measure ANOVA, employing the SPSS software. This analysis provided valuable insights that will be discussed in the subsequent chapter. The key effects of the model, character, style, and their interactions were considered. Finally, a comprehensive understanding of the uncanny valley was achieved through the visualization of donut plots in the latter part of the questionnaire.

Chapter 6

Discussion and Limitations

The outcomes and findings chapter of this study zoomed in on how different factors influenced what people thought about talking head avatars. The model used to create the avatars made a big difference videos made with the pose controllable model stood out as more convincing and appealing to participants. Interestingly, in some of the features, the style of the avatars also played a role in shaping the results for certain aspects. In this chapter, we will dig into the results and findings we uncovered. We will also compare our initial expectations and guesses with what actually happened in the study. Additionally, we will discuss the challenges we faced along the way. It is important to recognize that our study covers a specific part of a larger field. By addressing the limitations we encountered and making more accurate predictions in future work, we can build upon what we've learned and make our research more comprehensive.

6.1 Hypothesis and Discussions

As a reminder, our research questions were:

RQ1: Which models can generate avatars that generally give a better feeling to the audience?
RQ2: What distinctions emerge in user experiences when interacting with real, cartoony, and painted avatars? How do these differences influence viewers' perceptions and emotional responses toward these avatars?

We obtained a clear answer to the first question, with the pose controllable model outperforming others in all six features, as illustrated in the box plots in the previous section. This model prioritized refining both LipSync and HeadMovement, channeling emphasis into enhancing the avatar's overall pose. This emphasis on head movements and proficient LipSync likely contributed to improved user-avatar communication.

Moreover, the commendable video quality associated with this model could have further bolstered participants' engagement and experiences, according to the author's perspective.

In regard to the avatar's style, which was our second research question, we initially expected it to strongly influence all features, as the "uncanny valley" effect suggests that avatars resembling humans can feel strange. Although, the study showed that style did not make a huge difference. Notably, the painted, cartoony and original style avatars were the favorite among participants, showing that all avatars can still be well-received.

This finding supports the concept of the uncanny valley effect. People liked the original style avatars almost as much as the cartoony or painted styles and not more. It was the initial hypothesis that perhaps more realistic avatars with original pictures would receive higher scores. However, we observed that realism wasn't the sole determining factor in whether people liked the avatars or not. As highlighted in the donut plots from the previous chapter, people even experienced a sense of realism when interacting with avatars in a cartoony style. Interestingly, it's worth noting that the perception of realism wasn't exclusively tied to avatars with original pictures. Other elements, such as lipsync and head movement, played substantial roles in creating believable avatars. This suggests that the overall convincing quality of an avatar goes beyond its visual attributes alone.

Moreover, the character's gender did not seem to have a strong influence on people's preferences. Both boy and girl avatars were equally favored by the participants. This outcome was expected, as the boy character, despite being generated with GAN and not representing a real person, demonstrated remarkable performance, making it difficult for the human eye to discern any differences between the deepfake character and an actual individual. Therefore, our prediction proved accurate, and there were no significant discernible differences between the boy and girl characters.

The video made by the PC-AVS model, named PKO, along with Kian's original style, emerged as the winner in terms of overallQoE. This feature was of utmost importance in our evaluation. Therefore, we can confidently assert that PKO was the best video among all. Additionally, it also outperformed others in the comfortability and lipsync features. However, in the HeadMove feature, the video created by the PC-AVS model, PDP, featuring Donya in a painted style, achieved significantly better results, with PKO securing the second position in this category. Furthermore, PKO's superior performance in the talking feature, coupled with its commendable video quality, solidified its position as the overall winner in this study.

Our findings lead to the conclusion that in the realm of talking head avatars designed

for both children and adults, the observation consistently highlights the significance of model quality. This quality directly influences factors like lip sync performance, audio synchronization, video quality, and, most importantly, the overall comprehensive user experience.

Focusing on avatars created specifically for child abuse interviewing, a previous study within the child abuse avatar project revealed that experts emphasize the substantial impact of realism on outcomes(Hassan, Salehi, Røed et al., 2022). Aligning this with our broader observations, it's apparent that directing attention towards realism entails enhancing multiple facets of the avatars. This encompasses not only the visual appearance but also other features predominantly influenced by the models, which significantly contribute to achieving a heightened sense of realism.

Overall, it can be concluded that focusing on enhancing the models is a promising approach to improving talking head avatars. The PC-AVS model required additional inputs compared to the MakeItTalk and FOMM model models. It necessitated a video for training the person's pose in the final generated video. In contrast, the MakeItTalk and FOMM model models only required a single picture and an audio input for producing the output video. Considering this disparity in input requirements, it was reasonable to expect better results from the PC-AVS model. The availability of more detailed information about the person's pose through the training video likely contributed to the enhanced performance of the PC-AVS model, making it more capable of generating convincing and accurate talking head avatars. Utilizing new methods that require more inputs can lead to the development of even better talking head avatars, thus helping us achieve our goals of creating highly realistic and effective avatars for various applications.

6.2 Difficulties and Limitations

Throughout the process of running the models, we encountered various difficulties and encountered challenges. Initially, we compiled a list of models to execute and observe their outcomes. However, each model came with its own set of limitations. Some were solely documented in research papers without accompanying code, while others had GitHub repositories with codes that proved ineffective or unhelpful. Certain codebases were incomplete, and issues were reported, indicating problems in running them. Additionally, some codes lacked the flexibility to incorporate custom data for video generation, restricting us to using predefined materials. This posed a challenge to our study's objective of comparing videos generated from various models using consistent characters and styles. Details of certain models from our shortlist that were not utilized in Chapter 3 were elaborated upon, and a summary table was

included there..

Ultimately, we identified models that were suitable for our benchmarking task based on their accessibility and ease of use via Google Colab links. Despite encountering challenges when executing these selected models, most issues were solvable. For instance, some models were coded with outdated library commands that required modification to function properly with newer library versions. These conflicts were identified and resolved, ensuring the successful execution of the code. While our journey presented numerous hurdles, our commitment to addressing and overcoming each challenge allowed us to advance through the benchmarking process.

Another challenge revolved around the selection of characters for our study. The decision-making process involved choosing between utilizing real individuals or employing characters generated by GAN style which at the end, we have used both of them. Various characters were tested in conjunction with the videos to identify the most suitable match. A crucial aspect that contributed to achieving favorable outcomes was the preference for images where the person was looking directly forward. Furthermore, the task of selecting appropriate styles posed its own set of challenges. The chosen styles needed to exhibit believability while simultaneously exhibiting distinct differences from one another.

Initially, our attempt to source styles from a GitHub repository proved unfavorable, as the resulting output suffered from poor quality. Ultimately, the solution emerged in the form of generating cartoony and painted images through a dedicated website, which we introduce in the third chapter of this thesis. This strategic shift in approach facilitated the creation of styles that met the desired criteria and contributed positively to the overall quality and credibility of our study's results.

After videos were created, challenges were encountered in devising an effective questionnaire for obtaining participant responses. A form was required that could shuffle the videos to prevent bias from a single order. The complication arose from initial questions needing to precede the rest of the questionnaire, requiring video shuffling afterwards. Initially, Google Forms was chosen for the questionnaire creation, but it failed to fulfill the video shuffling requirement. Subsequently, an alternative solution was sought, leading to the adoption of Microsoft's platform, which not only offered user-friendly form creation but also facilitated video shuffling. Furthermore, due to being the short thesis, there was limited time available to delve extensively into this domain. Consequently, both the number of models and videos remained constrained.

Chapter 7

Conclusion

The primary aim of this thesis is to thoroughly evaluate talking head models based on user experiences and the level of realism they produce. This goal carries a critical purpose: to enhance the realism of a platform, which is intended for use by law enforcement and child protective service professionals to practice interview training with children who have endured trauma or abuse. This effort constitutes a segment of a larger project, wherein the development of talking head avatars serves as a platform for expert training before conducting actual interviews. The importance of this training stems from the sensitivity associated with interviewing such children. It is hoped that the results of this study can contribute, even if in a small way, to the enhancement of talking head avatars' communication capabilities with users.

7.1 Major Takeaways

- The scope of the study was refined to assess user experiences and perceptions, achieved through the utilization of three models for crafting talking head avatar videos featuring childlike characters. The models encompassed MakeItTalk, FOMM model, and PC-AVS model. Two characters, named Donya and Kian, were animated in three distinct styles: cartoony, painted, and original photo. In total, the generation of 18 videos, each lasting 10 seconds, served as the groundwork for a subsequent user study. Analysis revealed the model type had the most significant impact on differentiating the videos. Specifically, the PC-AVS model demonstrated superior performance over the other two across various metrics. The character itself did not produce any significant main effects. However, style exhibited a more complex interplay of effects depending on the factor being evaluated. As talking head avatar technology continues advancing, more sophisticated models are emerging that aim to heighten user satisfaction. A

key consideration is mitigating the uncanny valley effect, even for cartoonish or painted styles. This study highlighted the substantial influence of the uncanny valley phenomenon. Surprisingly, the cartoon and painted styles could compete with the more traditional original photo style by circumventing the uncanny valley.

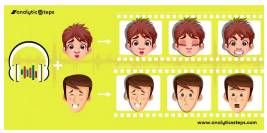
- One of the early resolutions formed during the brainstorming stage of this study was to compile a list of video-driven models and subsequently compare them with audio-driven models by conducting another user study. This comparative analysis aimed to shed light on the differences between these two model types. An additional intention was to conduct a separate user study to enable a comprehensive comparison of the outcomes derived from these distinct model groups. However, due to the time-consuming nature of these tasks, this phase had to be omitted from the current investigation.
- In exploring this field, there are several aspects that experts could think about. One thing to think about is adding more different models or trying out different looks for the same character. This could be interesting for future research. Also, our study mainly looked at how people felt when using these avatars. But there's another possibility we can explore: looking at how well the models work using some extra ways to measure their performance that we have. This could give us more insights. This idea would help us see how the models perform compared to how users feel about them. This would help us understand more about this area in a deeper way.

Appendix A

Questionnaire 1

Here is the link for the first part of the questionnaire[1].

[1] https://forms.office.com/r/2b3teM8EBQ



Benchmarking Talking-Heads Models:

(First Part)

Welcome!

The aim of this study is to get your opinion on different artificial faces. Imagine a scenario where you are speaking with a robot that has an artificial face. You will watch a series of videos of these avatars and then you are asked to answer some questions. Your responses to the questionnaires will be used for scientific research, please treat them very seriously to contribute with valid results. There is no right or wrong answer as we are looking for your opinion. The estimated time duration of the study is 15-20 minutes.

Please participate in this study if:

-Your device has a speaker or a headphone.
-You are older than 18 years old.
-You have a fair level of English to answer the questionnaires.
-You have no (strong) relevant visual constraint (e.g., color blindness).
-You agree that all data collected in the study gets stored and used anonymously for scientific analysis.

Procedure: The experiment consists of three sub-tasks.
- **Pre-test Questionnaire:** Instruction and answering demographic questions.
- **Test Section :** Watch 19 Talking Faces videos and answer a questionnaire and submit the form.
- **Post-test Questionnaire:** Answering final questions and submitting the form. This will be in another form.

1

Please insert the day of your birthday: *

Number must be between 1 ~ 31

2

Please choose the last digit of your mobile phone: *

0	1	2	3	4	5	6	7	8	9	10

3

How old are you?
Leave blank if you don't want to answer!

4

What is your gender?
Leave blank if you don't want to answer!

5

What is your job?
If student, please specify the field of study.
Leave blank if you don't want to answer!

6

Have you participated in any avatar studies before?
*

- ◯ Yes
- ◯ No

7

Please write down the 5-digit number from the audio file to confirm that you have access to a proper speaker or headphone! *

8

In this section, you will watch 19 different videos from combinations of different voices and avatars, and then you are asked to rate 4 questions for each video.
The duration of each video is around 10 seconds, but you can watch each video for multiple times.
When you are ready to start the test, Choose ready and go a head.

◯ Ready

AC: Please answer these questions after watching the video bellow: *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	◯	◯	◯	◯	◯
How accurately did the lips move in sync with the audio?	◯	◯	◯	◯	◯
How do you rate the naturalness of head movement?	◯	◯	◯	◯	◯
How realistic the avatar was talking?	◯	◯	◯	◯	◯

was talking?

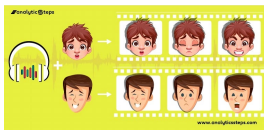
How do you rate the overall quality of this video?	◯	◯	◯	◯	◯
How comfortable do you feel conversing with this avatar?	◯	◯	◯	◯	◯

Appendix B

Questionnaire 2

Here is the link for the second part of the questionnaire[1].

[1] https://forms.office.com/r/0zWNMGJTVC



Benchmarking Talking-Heads Models

(second part)

Post-test Questionnaire: Answering final questions and submitting the form.

* Required

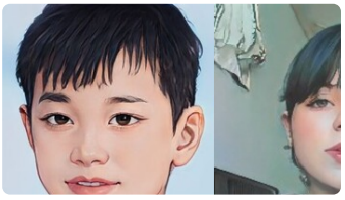
1. Please choose the last digit of your mobile phone: *

0	1	2	3	4	5	6	7	8	9	10

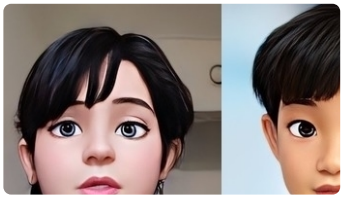
2. Please insert the day of your birthday: *

Number must be between 1 ~ 31

3. Considering the avatars' talking, which were the most realistic? *



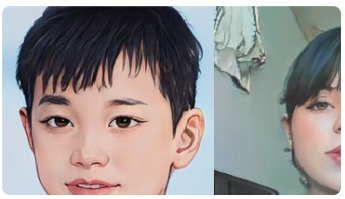
◯ Videos with painted avatar



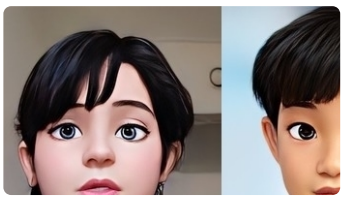
◯ Videos with cartoony avatar

◯ Videos with real picture

4. For a conversation (talking to a computer), which avatars do you prefer to use? *



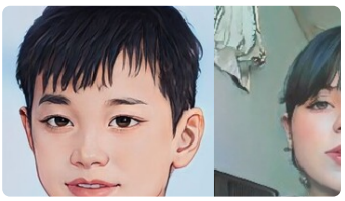
◯ Videos with painted avatar



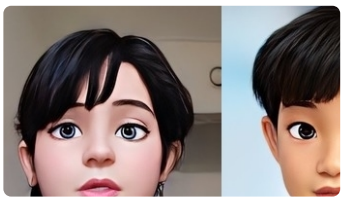
◯ Videos with cartoony avatar

◯ Videos with real picture

5. Generally, which avatars do you like the most? *



◯ Videos with painted avatar

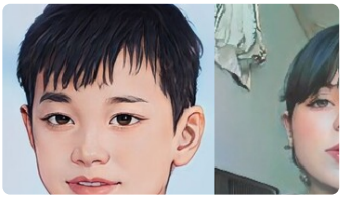


◯ Videos with cartoony avatar

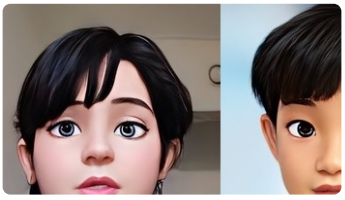
◯ Videos with real picture

6. Generally, which avatars did have the most realistic appearance for you? *

◯ Videos with real picture

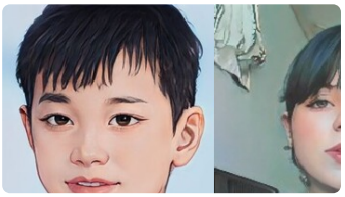


◯ Videos with painted avatar

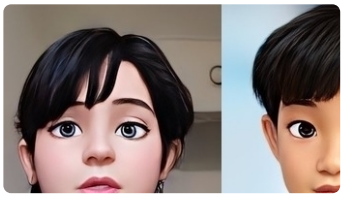


◯ Videos with cartoony avatar

7. Generally, which avatars were the most believable for you? *



◯ Videos with painted avatar



◯ Videos with cartoony avatar

◯ Videos with real picture

Bibliography

Abdullah, M., Madain, A., & Jararweh, Y. (2022). Chatgpt: Fundamentals, applications and social impacts. *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 1–8. https://doi.org/10.1109/SNAMS58071.2022.10062688

Barman, N., Reznik, Y., & Martini, M. (2023). Datasheet for subjective and objective quality assessment datasets. *arXiv preprint arXiv:2305.02142*.

Baugerud, G. A., Johnson, M. S., Klingenberg Røed, R., Lamb, M. E., Powell, M., Thambawita, V., Hicks, S. A., Salehi, P., Hassan, S. Z., Halvorsen, P., & Riegler, M. A. (2021). Multimodal virtual avatars for investigative interviews with children. *Proceedings of the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, 2–8. https://doi.org/10.1145/3463944.3469269

Bertram, D. (2007). Likert scales. *Retrieved November*, *2*(10), 1–10.

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making*, *7*(4), 223–242.

Chen, L., Cui, G., Kou, Z., Zheng, H., & Xu, C. (2020). What comprises a good talking-head video generation?: A survey and benchmark.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia: Case studies on organization and retrieval* (pp. 21–49). Springer.

Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018). Unsupervised learning based on artificial neural network: A review. *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 322–327. https://doi.org/10.1109/CBS.2018.8612259

Eskimez, S. E., Zhang, Y., & Duan, Z. (2021). Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, *24*, 3480–3490.

Ettinger, T. R. (2022). Children's needs during disclosures of abuse. *SN Social Sciences*, *2*(7), 101. https://doi.org/10.1007/s43545-022-00397-6

Girden, E. R. (1992). *Anova: Repeated measures*. sage.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks.

Hamzelou, J. (2017). Robots could help children give evidence in child abuse cases. *New Scientist*, *233*(3118), 20. https://doi.org/10.1016/S0262-4079(17)30305-3

Hassan, S. Z., Salehi, P., Riegler, M. A., Johnson, M. S., Baugerud, G. A., Halvorsen, P., & Sabet, S. S. (2022). A virtual reality talking avatar for investigative interviews of maltreat children. *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, 201–204. https://doi.org/10.1145/3549555.3549572

Hassan, S. Z., Salehi, P., Røed, R. K., Halvorsen, P., Baugerud, G. A., Johnson, M. S., Lison, P., Riegler, M., Lamb, M. E., Griwodz, C., & Sabet, S. S. (2022). Towards an ai-driven talking avatar in virtual reality for investigative interviews of children. *Proceedings of the 2nd Workshop on Games Systems*, 9–15. https://doi.org/10.1145/3534085.3534340

Htike, K. (2017). A review on data-driven learning of a talking head model. *International Journal of Intelligent Systems Technologies and Applications*, *16*, 169. https://doi.org/10.1504/IJISTA.2017.084239

Igaue, T., & Hayashi, R. (2023). Signatures of the uncanny valley effect in an artificial neural network. *Computers in Human Behavior*, *146*, 107811. https://doi.org/https://doi.org/10.1016/j.chb.2023.107811

Kammoun, A., Slama, R., Tabia, H., Ouni, T., & Abid, M. (2022). Generative adversarial networks for face generation: A survey. *ACM Comput. Surv.*, *55*(5). https://doi.org/10.1145/3527850

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307–392. https://doi.org/10.1561/2200000056

Labs, E. (2023). Eleven labs - text to speech api [[Online; accessed 27-March-2023]].

Lambie, G. W. (2005). Child abuse and neglect: A practical guide for professional school counselors. *Professional School Counseling*, *8*(3), 249–258. Retrieved March 22, 2023, from http://www.jstor.org/stable/42732466

Lu, Y., Chai, J., & Cao, X. (2021). Live Speech Portraits: Real-time photorealistic talking-head animation. *ACM Transactions on Graphics*, *40*(6). https://doi.org/10.1145/3478513.3480484

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? uncanny responses to computer generated faces [Including the Special Issue: Enabling elderly users to create and share self authored multimedia content].

Computers in Human Behavior, *25*(3), 695–710. https://doi.org/https://doi.org/10.1016/j.chb.2008.12.026

Meskys, E., Kalpokiene, J., Jurcys, P., & Liaudanskas, A. (2020). Regulating deep fakes: Legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, *15*(1), 24–31.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, *65*(1), 99–106. https://doi.org/10.1145/3503250

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), 98–100.

Norman, R. E., Byambaa, M., De, R., Butchart, A., Scott, J., & Vos, T. (2012). The long-term health consequences of child physical abuse, emotional abuse, and neglect: A systematic review and meta-analysis. *PLoS Medicine*, *9*.

Perarnau, G., van de Weijer, J., Raducanu, B., & Álvarez, J. M. (2016). Invertible Conditional GANs for image editing. *NIPS Workshop on Adversarial Training*.

Powell, M. B., Guadagno, B., & Benson, M. (2016). Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society*, *26*(4), 365–374. https://doi.org/10.1080/10439463.2014.942850

Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. (2020). A lip sync expert is all you need for speech to lip generation in the wild. *Proceedings of the 28th ACM International Conference on Multimedia*, 484–492. https://doi.org/10.1145/3394171.3413532

Ren, Y., Li, G., Chen, Y., Li, T. H., & Liu, S. (2021). Pirenderer: Controllable portrait image generation via semantic neural rendering.

Salehi, P., Hassan, S. Z., Lammerse, M., Sabet, S. S., Riiser, I., Røed, R. K., Johnson, M. S., Thambawita, V., Hicks, S. A., Powell, M., Lamb, M. E., Baugerud, G. A., Halvorsen, P., & Riegler, M. A. (2022a). Synthesizing a talking child avatar to train interviewers working with maltreated children. *Big Data and Cognitive Computing*, *6*(2), 62. https://doi.org/10.3390/bdcc6020062

Salehi, P., Hassan, S. Z., Lammerse, M., Sabet, S. S., Riiser, I., Røed, R. K., Johnson, M. S., Thambawita, V., Hicks, S. A., Powell, M., Lamb, M. E., Baugerud, G. A., Halvorsen, P., & Riegler, M. A. (2022b). Synthesizing a talking child avatar to train interviewers working with maltreated children. *Big Data and Cognitive Computing*, *6*(2). https://doi.org/10.3390/bdcc6020062

Salehi, P., Hassan, S. Z., Shafiee Sabet, S., Astrid Baugerud, G., Sinkerud Johnson, M., Halvorsen, P., & Riegler, M. A. (2022). Is more realistic better? a comparison of game engine and gan-based avatars for investigative interviews of children. *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, 41–49.

Schmidt, S., Zadtootaghaj, S., Möller, S., Metzger, F., Hirth, M., & Suznjevic, M. (2018). Subjective evaluation methods for gaming quality (p. game).

Sethi, D., Bellis, M., Hughes, K., Gilbert, R., Mitis, F., & Galea, G. (2013). *European report on preventing child maltreatment*. World Health Organization. Regional Office for Europe.

Shao, H., Yao, S., Sun, D., Zhang, A., Liu, S., Liu, D., Wang, J., & Abdelzaher, T. (2020). Controlvae: Controllable variational autoencoder. *International Conference on Machine Learning*, 8655–8664.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019a). First order motion model for image animation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019b). First order motion model for image animation. *Advances in neural information processing systems*, *32*.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, *36*(4). https://doi.org/10.1145/3072959.3073640

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395.

This Person Does Not Exist [Accessed on March 28, 2023]. (n.d.).

Toshpulatov, M., Lee, W., & Lee, S. (2023). Talking human face generation: A survey. *Expert Systems with Applications*, 119678.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 2053951717743530.

Wang, H.-n., Liu, N., Zhang, Y.-y., Feng, D.-w., Huang, F., Li, D.-s., & Zhang, Y.-m. (2020). Deep reinforcement learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, *21*(12), 1726–1744.

Wang, T.-C., Mallya, A., & Liu, M.-Y. (2021). One-shot free-view neural talking-head synthesis for video conferencing. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Wang, X., & Yu, J. (2020). Learning to cartoonize using white-box cartoon representations. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8090–8099.

Wilson, E., Hewett, D. G., Jolly, B. C., Janssens, S., & Beckmann, M. M. (2018). Is that realistic? the development of a realism assessment questionnaire and

its application in appraising three simulators for a gynaecology procedure. *Advances in Simulation*, *3*(1), 1–7.

World Health Organization. (2021). Child maltreatment [Accessed on March 22, 2023].

Yang, S., Jiang, L., Liu, Z., & Loy, C. C. (2022). Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)*, *41*(6), 1–15.

Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Budagavi, M., & Guo, X. (2021a). Facial: Synthesizing dynamic talking face with implicit attribute learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3867–3876.

Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Budagavi, M., & Guo, X. (2021b). FACIAL: synthesizing dynamic talking face with implicit attribute learning. *CoRR*, *abs/2108.07938*. https://arxiv.org/abs/2108.07938

Zhang, T., Deng, L., Zhang, L., & Dang, X. (2020). Deep learning in face synthesis: A survey on deepfakes. *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*, 67–70. https://doi.org/10.1109/CCET50901.2020.9213159

Zhou, H., Sun, Y., Wu, W., Loy, C. C., Wang, X., & Liu, Z. (2021). Pose-controllable talking face generation by implicitly modularized audio-visual representation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4176–4186.

Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., & Li, D. (2020). Makelttalk: Speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, *39*(6), 1–15.