IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 July 2024, accepted 11 July 2024, date of publication 18 July 2024, date of current version 30 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430835

RESEARCH ARTICLE

Gaze Generation for Avatars Using GANs

DAVID DEMBINSKY[1,2], KO WATANABE[1,2], ANDREAS DENGEL[1,2], AND SHOYA ISHIMARU[3,4], (Member, IEEE)

[1]Department of Computer Science, Rheinland-Pfälzische Technische Universität (RPTU) Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[2]DFKI GmbH, 67663 Kaiserslautern, Germany
[3]Graduate School of Informatics, Osaka Metropolitan University, Osaka 530-0001, Japan
[4]DFKI Laboratory Japan, Osaka 599-8231, Japan

Corresponding author: David Dembinsky (david.dembinsky@dfki.de)

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) International Call on Artificial Intelligence "Learning Cyclotron" under Project 442581111.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Graduate School of Informatics at Osaka Metropolitan University.

ABSTRACT The movement of our eyes during conversations plays a crucial role in our communication. Through a mixture of aimed and subconscious control of our gaze, we nonverbally manage turn-taking in conversations and convey information about our state of mind and even neurological disorders. For animated avatars or robots, it is hence of fundamental importance to exhibit realistic eye movement in conversations to withstand the scrutiny of an observer and not fall into the Uncanny Valley. Otherwise, they will be rejected by the observer as unnatural and possibly scary, provoking disapproval of the entire avatar. Although there exist many promising application areas for avatars and great attention has been given to the automatic animation of mouth and facial expressions, the animation of the eyes is often left to simplistic, rule-based models or ignored altogether. In this work, we aim to alleviate this limitation by leveraging Generative Adversarial Networks (GANs), a potent machine-learning approach, to synthesize eye movement. By focusing on a restricted scenario of face-to-monitor interaction, we can concentrate on the eyes, ignoring additional factors such as gestures, body movement, and spatial positioning of conversation partners. Using a recently published dataset on eye movements during conversation, we train two GANs and compare their performance against three statistical models with hand-crafted rules. We subject all five models to statistical analysis, comparing them to the ground-truth data. We find that the GANs produce the best data of the four models that synthesized reasonable eye movement (excluding the best-scoring model for generating absurd movements). Additionally, we perform a user study, comparing each model pairwise against the others based on 73 participants, resulting in a total of 1314 pairwise comparisons. It shows that the GANs achieve acceptance ratings of 55.3% and 43.7%, outperforming the baseline model with an acceptance rate of 34.0%. Although the best model reaches 67.0%, beating our GANs using a set of rules, we argue that this approach will not be feasible once information like emotions or speech is added to the input.

INDEX TERMS Avatars, eye gaze, GAN, human–computer-interaction.

I. INTRODUCTION

Eye movements are a fundamental part of nonverbal communication during conversations [1], [2], [3], [4]. Although most movements are induced and noticed unconsciously, they convey important information. The gaze is used to manage conversations and facilitate turn-taking, as has been thoroughly investigated [5]. It includes semantically context-free patterns like looking away during thinking or gazing at the conversation partner at the end of one's speech [6], [7], as well as context-sensitive interactions between the spoken word and eye movement [8]. Additionally, looking at partners is

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

used to express and reassure shared attention during conversation [9] or draw other's attention to objects of interest [10].

In conversational avatars, the eye movement they exhibit can have a considerable influence on the avatars' perception and humans' responses [11]. Appropriate gaze behavior increases the realism of conversational avatars, but at the same time, a more realistic appearance of avatars also increases the observer's expectation and demand for faithful eye movements [12]. Avatars failing to meet the observer's, usually unconscious, expectations regarding their behavior and movement are at risk of falling into the Uncanny Valley [13]. It will not only diminish the acceptance of the avatar but may induce negative associations towards that avatar, making it appear scary at worst. Such a reception is counterintuitive for most application scenarios, narrowing the benefit conversational avatars have in areas such as teaching [14], [15], [16].

One often neglected part of avatars' movements are the eyes, which are dealt with superficially most of the time, resorting to simplistic control or no eye movement at all [17], [18], [19], [20], [21]. Lee et al. compared the effects of different eye movements in conversational avatars among observers and found that reasonable movements, exhibited by a rule-based model, are superior to static eyes or random movement in engagement or livelihood [22]. However, generating high-quality gaze patterns for conversational avatars or robots remains an open field. To our knowledge, research in this area has come to a halt, with the last considerable model to generate eye movements being introduced over 20 years ago [22]. More recent works either rely on this same model [23], use similar methods without publishing their parameters [24], [25], built even more simplistic rule-based models [26], or focus on transferring real eye movement to an avatar [24], [25], [27].

In this work, we want to close the gap by leveraging Generative Adversarial Networks [28] (GANs) to control the eye movements of a conversational avatar while speaking and listening. GANs have seen a wide range of applications in avatar generation over the past years [20], [21], [29], [30], [31]. We separate the eye's control from the avatar's visualization, to be applicable to any visualization software and compare different eye movement generators in isolation. We believe the approach of generating visualization parameters and leaving the visualization to external software has its very important niche when it comes to avatars, allowing for a wide range of applications such as video calls, augmented reality, and video games or movies. The contributions of this research can be summarized as:

- Building a procedural model and training two GANs to generate eye movements that outperform the previous state-of-the-art from Lee et al., relying on a larger corpus of data than previous works.
- Evaluating the performances of those models in an extensive user study, collecting more feedback than related research, and providing insight into some of the decisive criteria in rejecting a given avatar.

TABLE 1. The statistics of the machine learning datasets and their respective windows. ω is chosen as 3seconds, to match the smallest windows for $W = 10$. Some randomness in the fixed dataset is introduced by splitting it into training and evaluation subsets, as we ensure that they share not one single fixation.

	fixed window $W = 10$		variable window $\omega \geq 3$s
Total windows:	2697.6	±45.0	2037
Min duration (s):	3.0	±0.2	3.0
Max duration (s):	22.3	±0	37.6
Mean duration (s):	5.6	±0	5.9

The structure of this work is as follows: We will introduce the current state-of-the-art in the generation of conversational avatars and their eye movement, and describe the dataset of eye movement collected by Dembinsky et al., which we will rely on in this work [32]. Afterward, we introduce the different gaze generators and their setup and outline our visualization pipeline. We evaluate all generators through a quantitative statistical analysis and a qualitative user study, before discussing the results.

II. RELATED WORK

A. GAZE GENERATION

To our knowledge, little attention has been given to the generation of eye movement for avatars so far. Most researchers investigating the animation of avatars' heads focused on mouth dubbing and to some extent facial animation, deliberately ignoring the eyes or relying on simplistic models [17], [18], [19], [20], [21].

The timely VASA-1 model by Xu et al. generates fully animated talking faces from a single image and an audio clip [33]. The avatars exhibit convincing head and mouth movement, facial expression, and also eye movement. Unfortunately, their evaluation does not include a user study focused on the individual parts of the talking face and thereby leaves it open, whether the eye movements themselves are faithful. Although they generate movements directly from speech, they explicitly assume the voice in the audio to belong to the avatar, thereby ignoring the listening case. Finally, their diffusion-based approach generates the video from a single image input, which sometimes leads to rendering artifacts like morphing teeth. We want to pursue a different approach, only generating the parameters of the movement and leaving the visualization to separate software.

One major contribution to the field of procedural models, generating only the parameters of eye movement, stems from Lee et al. [22]. They built a model that generates gaze based on a 9-minute sample video from one participant. They extracted the direction and magnitude of saccades as well as the length of fixations during talking and listening modes. To gain high-level control of the eyes, they differentiated between mutual-gaze with the eyes aimed at the conversational partner's face and gaze-away, alternating between those two states. Using a (compared to modern

standards) simple visualization, they evaluate their gaze model against a static model without movement and a model with random movement. Presenting a clip for each of the three models to 12 participants, they found their model to be superior to the other models in perceived interest, engagement, friendliness, and livelihood.

The more recent approach of Canales et al. defined the task of predicting gaze and head movement as a regression task [34]. They leveraged Recurrent Neural Networks to predict new movement from the prior one on a database of around 36 minutes from four performers. Evaluating different inputs, they achieved better results with the binary labels of speaking and listening, rather than additional audio information in the form of pitch and intensity. They conducted a user study with stylized avatars and compared two of their models against captured data, no movement, and two procedural models based on [22]. With a total of 64 responses, ranking each of the models individually, their model ranged worse than the procedural models, leaving the model by [22] to be the current baseline.

B. DATASET

Whereas the previously mentioned works relied on rather small samples of eye movement, we make use of the very recent dataset by [32]. It is considerably larger, consisting of almost 4 hours of eye movement categorized into speaking and listening. It was collected from 19 participants in a dyadic video call, thereby effectively eliminating additional variables such as gestures or relative placement of participants. They compared their data to the findings of [22] and found similar results on the foundation of more data. Unfortunately, they report the gaze position only in terms of screen pixel, without taking into account what the participants are looking at and thereby are not differentiating between mutual-gaze and gaze-away.

We will use their dataset with extracted fixation points, which used Dispersion-Based Identification to combine subsequent gaze points that are locally close into a fixation. Therefore, each data point is a fixation, which is defined by its length, position, and gap before the next fixation. Unfortunately, their recordings contained various missing values (NaNs) for the positional data, when people looked outside the boundaries of the screen. Since it is not reasonable, to interpolate those gaps, it leaves the data fragmented into sections of fixations, ranging from one NaN gap to the next.

In their work, they construct a machine-learning dataset from the fixations, by using a sliding window of fixed length $W = 10$. Training a classifier to distinguish between speaking and listening samples, they achieve an accuracy of 88.1%. Hence we will use the same dataset and window size, to train our gaze generator. Additionally, we construct a second machine learning dataset, extracting windows that range from one NaN gap to the next, containing a varying amount of fixations. Based on the minimum duration of the fixed windows, we select a threshold for the second dataset, defined as a minimum duration $\omega \geq 3$ seconds per window. In Table 1 we present statistics of those two datasets.

Algorithm 1 Gaze Generator: Fuse

Require: target_duration D, conversation_mode M

$d = 0$, fixation_sequence $\leftarrow$ empty, state s = mutual

while $d \leq D$ **do**
  macro_duration = get_Fixation_Length$^{\text{macro}}(M, s)$
  macro_saccade = get_Saccade_Length$^{\text{macro}}()$
  macro_position = get_Fixation_Position$^{\text{macro}}(M, s)$

  $\hat{d} = 0$
  while $\hat{d}$ < macro_duration **do**
    F_d = get_Fixation_Length$^{\text{micro}}(M)$
    F_s = get_Saccade_Length$^{\text{micro}}()$
    F_x, F_y = get_Fixation_Position$^{\text{micro}}(M,$ macro_position$)$

    $F_s \leftarrow$ macro_saccade if last micro iteration
    fixation_sequence $\leftarrow$ add (F_d, F_s, F_x, F_y)
    $\hat{d} = \hat{d} + F_d + F_s$
  end while
  $d = d + \hat{d}$
  state $s \leftarrow$ switch mutual/away
end while

III. METHODOLOGY

A. GENERATORS

We built a total of five models to generate eye movement. Each generator takes two input parameters, the desired duration of the sequence of fixations and the conversation mode (either ''listening'' or ''speaking''). Each generator outputs a series of fixations of the desired time, defining each fixation's duration, the gap before the next fixation starts, and the position. For the position we rely on the same convention as [32], using screen coordinates in pixels, assuming a screen of size 637.35mm × 438.90mm at a head-to-screen distance of approximately 675mm, using a resolution of 2250px × 1500px. The first three models are procedural generators, relying on a fixed set of rules and the statistics of eye gaze reported in the different research. The final two models are deep learning generators, trained on either the dataset with fixed or variable window length.

1) PROCEDURAL MODELS

The first model relies on the earlier work by [22] and is therefore referred to as *Lee*-model. According to the original report, we differentiate between the two states of *mutual-gaze* and *gaze-away*. The first is looking directly at the primary position (the other participant's face) and the second is looking somewhere else. We always start the generated sequence in the primary position and from there, we alternate

between the two states. The exact parameters used for the probability distributions can be found in the supplements. The fixation length is sampled from a probability distribution based on the conversation mode and the gaze position state. For listening and mutual-gaze speaking we used the parameters provided. Unfortunately, the gaze-away speaking parameters they provided did not match a figure they included showing the function. Hence, we chose to rely on the figure provided, extracted the data from it, and used that to fit our function. They normalized all saccade lengths to 6 frames (at 30fps). Hence we fixed every saccade's duration to $\frac{6\text{frames}}{30\text{fps}} = 0.2$ seconds. The gaze-away position is calculated by sampling the saccade magnitude using the function provided and choosing a direction according to the reported distribution. Since each direction bin spans 45°, we add $\mathcal{N}(0, 5)^\circ$ to the bin center for more variability. For mutual-gaze, the position is set at the center position (1125, 750)px.

Second, the *Dembinsky*-model will use the statistics reported in the recent work of [32]. The general procedure is similar to the *Lee*-model, with the metrics defined in the supplementary material. In contrast to the previous model, we do not differentiate between mutual-gaze and gaze-away. Instead, each fixation point is calculated from the last one by adding a saccade. The time of a saccade is determined randomly from $\mathcal{N}(0.2, 0.04)$s since the gaps between fixations in the ground-truth data are not identical to straight saccades (e.g. through recording noise or eye-rolling). The saccade magnitude has to be determined first, for the saccade direction to be chosen according to the correct distribution since the directions vary for larger ($> 2^\circ$) and smaller ($\leq 2^\circ$) saccades.

Finally, we used a combination of the previous two models to rely on the strength of either model, accounting for the weaknesses of the other one, naming it the *Fuse*-model. Since the *Dembinsky*-model ignores mutual-gaze and gaze-away, it has no mechanism to restrict the fixation points, leading to fixation sequences wandering off far away without coming back to the other participant's face. On the other hand, the *Lee*-model exhibits only a few, very long fixations, with no movement in between. Therefore, we use the *Lee*-model to control the macro-movement, determining the duration of mutual-gaze and gaze-away, but use the statistics of the *Dembinsky*-model to enhance it with frequent micro-movement. The algorithm of this combined model is shown in Algorithm 1. First, the macro-movement is handled, deciding whether to do a mutual-gaze or gaze-away, determining the macro-position, as well as the duration for that position. The duration is calculated using the mutual-gaze and gaze-away statistics from *Lee*. The position is calculated using the *Dembinsky*-model, but only taking big saccades ($> 2^\circ$) into account. Afterwards, the macro-movement's time is divided into fixations with shorter duration. Each fixation's position is calculated by adding a small movement to the previously defined macro-position. Those small movements are calculated from the *Dembinsky*-model, only using small ($\leq 2^\circ$) saccades. The saccade durations are sampled from normal distributions, with those between macro movements being longer than those between micro-movements.

Algorithm 2 GAN Training Algorithm

Require: Hyperparameters n_{epochs}, n_{rounds}, L, σ_{latent},λ, α_D, α_G

for $e = 0, \ldots, n_{epochs}$ **do**
  for $r = 0, \ldots, n_{rounds}$ **do**
    for $i = 1, \ldots, m$ **do** ▷ Train Discriminator
      Sample real data $x \sim \mathbb{P}_r$
      Sample latent variable $z \sim \mathcal{N}\left(0, \sigma^2_{latent} \cdot \mathbf{I}_L\right)$
      Sample random variable $\epsilon \sim \mathcal{U}[0, 1]$
      $\tilde{x} = G_\theta(z)$
      $\hat{x} = \epsilon(x) + (1 - \epsilon)(\tilde{x})$
      $PG = \lambda\left(||\nabla_{\hat{x}} D_\omega(\hat{x})||_2 - 1\right)^2$
      $\mathcal{L}^{(i)}_D = D_\omega(\tilde{x}) - D_\omega(x) + PG$
    end for
    $\omega \leftarrow \text{Adam}(\nabla_\omega \frac{1}{m} \sum_{i=1}^m \mathcal{L}^{(i)}_D,\ \omega, \alpha_D,)$
  end for
  for $i = 1, \ldots, m$ **do** ▷ Train Generator
    Sample latent variable $z \sim \mathcal{N}\left(0, \sigma^2_{latent} \cdot \mathbf{I}_L\right)$
    $\tilde{x} = G_\theta(z)$
    $\mathcal{L}^{(i)}_G = -D_\omega(\tilde{x})$
  end for
  $\theta \leftarrow \text{Adam}(\nabla_\theta \frac{1}{m} \sum_{i=1}^m \mathcal{L}^{(i)}_G,\ \theta, \alpha_G)$
end for

We use the two datasets described in subsection II-B to train two different machine-learning generators. Whereas the fixed window dataset is suitable for any feed-forward architectures, the variable window dataset can only be used with a Recurrent Neural Network (RNN).

For training, we use a variation of the original Generative Adversarial Network (GAN) introduced by [28], the Wasserstein GAN [35] with Gradient Penalty [36]. It consists of two networks, the Generator[1] and Discriminator, which are trained alternately. The Generator takes a random variable, the latent vector, as input and outputs a sequence of fixations. The Discriminator consumes real and synthesized fixations and rates them, assigning high values to real data and low values to forged samples, to reflect the Wasserstein-1 distance between the distribution of real and fake data. The gradient penalty enforces a norm of 1 for all samples on straight lines between real and fake data, thereby satisfying the 1-Lipschitz constraint. It is calculated as the norm of the gradient on interpolated data from the real and fake data set which is added to the Discriminator's loss. The training algorithm is given in Algorithm 2. The GAN trains over n_{epochs} epochs, alternating between n_{rounds} consecutive updates to the Discriminator and one single update step to the Generator. The space from which we draw the latent

[1]In the context of this work "generator" will refer to gaze generators in general, whereas "Generator" refers to one of the two neural networks forming the GAN.

TABLE 2. The best found hyperparameters for the GANs trained on the two different datasets.

	"fixed window"	"variable window"
n_{epochs}	50 000	5 000
n_{rounds}	5	5
L	150	500
σ_{latent}	1	2
λ	2.5	20
α_G	0.000 05	0.000 01
α_D	0.000 1	0.000 01

variable z, the Generator's input, is a vector of length L, where each entry is drawn from a normal distribution with mean 0 and standard deviation σ_{latent}. We use Adam optimizer with learning rates α_G and α_D for Generator and Discriminator, respectively.

We performed an exhaustive search for architectures and hyperparameters and report the best results. As selection criteria, we used a subjective evaluation of the generated data, together with the Wasserstein-1 distance W_1 [37] and two-sided Kolmogorov-Smirnov test δ_{KS} [38], both given in Equation 1, comparing the individual distributions of the generated data to those reported in [32]. On the fixed window dataset, we trained a Convolutional Neural Network (CNN) for both the Generator and Discriminator, with two layers of feature-wise and timestep-wise convolutions side by side. The GAN on the dataset with variable window lengths uses a simple RNN with one Gated Recurrent Unit (GRU) layer. The Discriminator is an RNN with three stacked layers of GRUs. The architectures are described further in the supplementary material. We evaluated various methods of handling the desired label, either "speaking" or "listening". We found it most beneficial, to duplicate the architecture of the Generator, training one per label, and to add a classification output to the Discriminator, adding the binary cross-entropy classification loss to the Discriminator's loss. Incorporating the label into the input, by concatenating it with the fixational data showed fewer improvements. The best hyperparameters are reported in Table 2.

2) MACHINE LEARNING GENERATORS

Given two distributions μ, ν we sample the data u, v

Kolmogorov-Smironv test [38]

$$\delta_{KS}(\mu, \nu) = ||F_u - F_v|| = \sup_x |F_u(x) - F_v(x)| \quad (1)$$

Wasserstein-1 distance [37]

$$W_1(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^2} |x - y| \, d\gamma(x, y) \right)$$

with $\Gamma(\mu, \nu)$ the set of all couplings of μ, ν

We use the trained GANs as gaze generators, resulting in a CNN (from fixed window) and RNN (from variable window) model. Each Generator synthesizes a window of fixations, each one a 4-tuple, from a random latent vector.

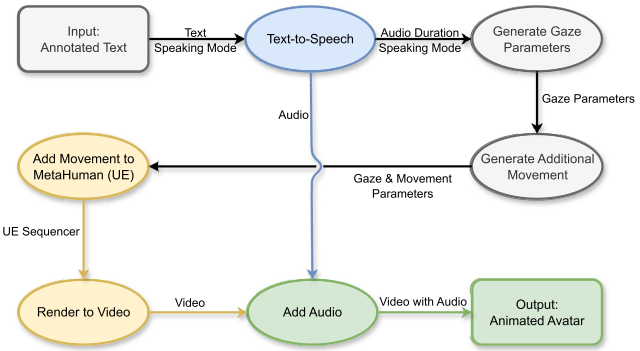


FIGURE 1. The avatar rendering pipeline encompasses various steps to create the finished video of an animated avatar (green) by generating audio (blue) and video (yellow) from a text input and numerical parameters (grey).

To receive a sequence of fixations with the desired length, multiple windows are concatenated. We clipped the output data, to ensure the fixational duration and pause between fixations to be at least one frame.

B. VISUALIZATION

The rendering pipeline encompasses various steps, which are presented in Figure 1. The input to the pipeline is a transcript of a dialogue, annotated with the speaker of that particular passage. We decided to only differentiate between a male ("m") and a female ("f") speaker, to make the difference as obvious as possible. An example could look like:

- ("m", "Hello, my name is John. I live in ..."),
- ("f", "Well hello, I'm Sarah. ..."),

In a Text-to-Speech module, the transcript is then converted to audio, using the voice matching the annotated gender. For each section, the duration of the generated audio is extracted and forwarded together with the information on whether the primary avatar (in our case always the male avatar) is speaking ("m") or listening ("f"). We used a text-to-speech model from HuggingFace Inc.[2] based on SpeechT5 [39] to produce audio from the transcripts. Given duration and speaking mode, the parameters for gaze and additional movements are then generated and forwarded to the visualization module.

In an intermediary step, the parameters get converted to a format suitable for keyframe animation. The fixations are a 4-tuple containing fixation and saccade durations in seconds and the x- and y-positions in pixels. Given an fps value, the durations get converted from seconds to frames, adding a key with the same position for the start of the fixation and one for the end. The gaze angles have to be converted to coordinates, using the data collection measures. Additionally, the alignment has to change, as in the eye-tracker data, the origin is at the bottom left of the screen, and for the UE environment, the origin is at the center (looking straight). The conversions are shown in Equation 2.

[2]https://huggingface.co/microsoft/speecht5_tts

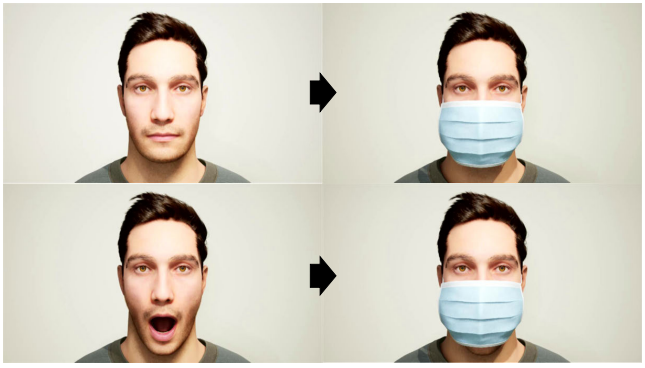


FIGURE 2. An example of the mask covering the mouth to not distract participants through unfaithful mouth movement. It is still possible to distinguish between listening (mouth constantly closed) and speaking (mouth opens and closes).

Temporal Conversion

$$\text{frame}^i_{\text{start}} = \begin{cases} 0, & \text{if } i = 0 \\ \sum_{j=0}^{i-1} (F^j_d + F^j_s), & \text{else} \end{cases}$$
$$\text{frame}^i_{\text{end}} = \text{frame}^i_{\text{start}} + F^i_d \tag{2}$$

Positional Conversion

$$p^{UE}_x = p^{px}_x \cdot \frac{637.35\text{mm}}{2250\text{px}} - \frac{637.35\text{mm}}{2}$$
$$p^{UE}_y = p^{px}_y \cdot \frac{637.35\text{mm}}{2250\text{px}} - \frac{438.9\text{mm}}{2}$$

We render the animated avatar using Unreal Engine (UE) [3] v5.1.1 and its MetaHumans.[4] This framework allows us to focus on the parameters that control the avatar's movement rather than computer graphics. Inside UE we mimicked the study setup, letting the avatar look at an invisible computer screen and placing the camera inside this screen, at the reported distance. Finally, the audio created earlier has to be added to the video, completing our visualization pipeline.

To add more realism to the rendered avatar, we implement additional movements, using simplistic rules. First, we add blinks to the eyes which are independent of the fixations. For this we randomly calculate the duration between blinks, the duration the eye takes to open and close, and the duration the eye is closed. Additionally, we implement mouth movements during speaking to give a visual cue to an observer, from which conversation mode a gaze pattern is supposed to estimate. During speaking we randomly open and close the mouth of the avatar and keep it closed during listening. Because we don't want observers to be distracted by the random, and hence inaccurate, mouth movement, we obscured it through a medical mask. We assumed this would be least obstructive, as most people are accommodated to talking with someone wearing a medical mask since the COVID-19 pandemic. This way, it is still possible to see whether the mouth is moving to distinguish between listening and speaking, but it doesn't distract the observer's attention. Figure 2 presents an example of the mask application, demonstrating how it is still possible to identify whether the mouth is moving or not.

We implemented our pipeline offline, to allow us to inspect each generated video for the study visually, although we did not perform any filtering on the generated outputs. Nevertheless, our GAN Generators are lightweight enough to ensure generation in online scenarios. We tested the performance using PyTorch[5] v2.1.1 on a 2.30GHz processor[6] with 16GB RAM, one thread, and without using GPU acceleration, while the computer in question was doing other tasks simultaneously. The trained CNN and RNN Generators took 3.3ms and 39.5ms on average to generate 100 windows of fixations, each covering around 1 minute, resulting in an estimated processing time of 5.3μs and 69μs per frame at 90fps (one frame covers $\sim$ 11.1ms), respectively. This should be more than enough, to allow for real-time generation.

TABLE 3. The scores from the statistical analysis are computed by assigning higher scores to low distances between the generated and real data for each parameter.

Dembinsky	CNN	RNN	Fuse	Lee
16.69	16.53	14.31	9.06	3.30

IV. EVALUATION

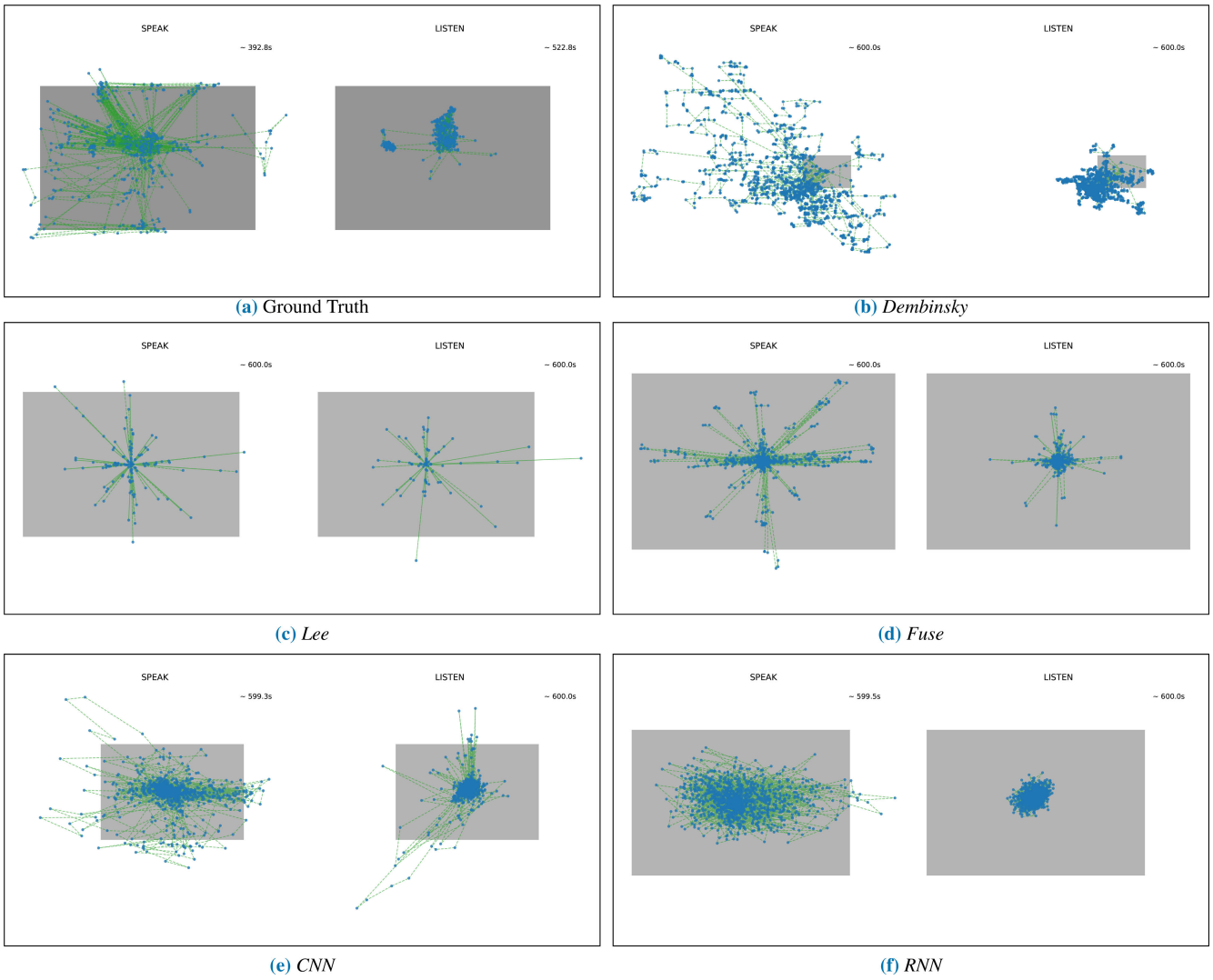
A. STATISTICAL ANALYSIS

To assess the performance of the five generators quantitatively, we compare their created data to the ground truth of the dataset. Therefore, we let each model synthesize 10 minutes of gaze for both speaking and listening. To give an easy-to-understand intuition of the gaze patterns, we present a visualization of the original gaze and the generated fixation points in Figure 3. We assess the similarity through W_1 and δ_{KS} (see Equation 1), comparing the following five properties: Fixation and saccade duration, saccade magnitude, and the distribution of the magnitudes for small ($\leq 2°$) and large ($> 2°$) saccades. This results in a total of 20 numerical ratings per model. For each of the 20 different property-metric pairs, we assign a score to each model, where the best one receives a score of 1 per pair and the worst one receives a score of 0. We min-max normalize the values per pair across all models, to account for how close two values are, using the formula in Equation 3. The resulting ordering is presented in Table 3.

$$\text{score}(\text{model}[i]) = \sum_{pm} s_{pm}(\text{model}[i])$$
$$s_{pm}(\text{model}[i]) = \frac{\max[v_{pm}] - v_{pm}(\text{model}[i])}{\max[v_{pm}] - \min[v_{pm}]} \tag{3}$$

with value v of all property-metric pairs pm

[3]Unreal Engine: https://www.unrealengine.com.
[4]MetaHuman: https://www.unrealengine.com/metahuman.
[5]https://pytorch.org/ [40]
[6]Intel i5-6200U.



(a) Ground Truth (b) *Dembinsky* (c) *Lee* (d) *Fuse* (e) *CNN* (f) *RNN*

FIGURE 3. The different gaze patterns created by the individual generators. Each blue dot is one fixation connected by green lined saccades. The computer screen is represented through the gray rectangle (compare [32]). The *Dembinsky*-model's gaze wanders off aimlessly. The procedural models *Lee* and *Fuse* create star-like shapes through mutual-gaze and gaze-away control. The machine learning models *CNN* and *RNN* produce the most reasonable output, with the first showing more variety and nuances.

Unsurprisingly, *Dembinsky* performs best, as it is deliberately designed to mimic the statistics of the ground truth dataset. All of its individual properties are extremely close to those of the original dataset, except for saccade duration, which was designed to only loosely match the statistics. However, inspecting the generated data visually (see Figure 3), it is apparent that the model's lack of high-level control (like mutual-gaze and gaze-away) results in the gaze wandering aimlessly around.1 Hence, statistical accuracy does not necessarily imply meaningful gaze patterns. The *CNN*-model however can achieve an almost identical score, while producing reasonable gaze patterns at the same time. The gaze visualization looks somewhat similar to the one presented by the original data, with most fixations near the center, and more movement during speaking than listening. The *RNN*-model proves slightly inferior, reaching worse scores in the individual properties, as reflected by the cumulative score and being the weaker of the two machine learning models. The *Fuse* and *Lee*-models rely on the additional information from [22], especially the mutual-gaze and gaze-away statistics, hence their bad score comes to little surprise. The mutual-gaze and gaze-away mechanic is reflected in the gaze patterns, producing an ''exploding star'', with the mutual-gaze at the center.

Overall, it can be stated, that the machine learning generators outperform those procedural models capable of creating meaningful data. Since the *Dembinsky*-model failed to produce reasonable output, we discarded it from further analysis.

B. USER STUDY

1) SETUP

To collect meaningful qualitative feedback, we conduct a user study on the perceived naturalness of the different avatars. Therefore, we used our visualization pipeline (subsection III-B) and generated conversational avatars using the different eye gaze generators. We wrote three dialogues between a male (''John'') and a female (''Sarah'') avatar, with varying proportions of speaking and listening, named ''Speak'', ''Listen'' and ''Dialogue'', according to the main activity[7] of our avatar, John. We showed participants two videos side by side and tasked them, to select the less natural-looking one, concerning eye movement. Each participant watched a total of 18 video pairs, given by the $6(=3!)$ pairings of the used models and 3 audios. Since the models are nondeterministic, we generated three videos per audio and model, to get a more robust testing corpus and each time selected one of those videos at random. We decided to use pairwise comparison to make the decision easier, eliminating the subjectiveness of a scoring system, and thereby making the results more robust. Similar approaches were used in related studies [19], [41], [42]. Participants were allowed to rewatch each pair as often as they wanted and skip a pair if they couldn't decide.

We evaluated the reliability of each participant, by presenting three video pairs, with the task of selecting the model which exhibits eye movement. The test consisted of the pairings: (Lee, Static), (Fuse, Static), (Lee, Fuse), with ''Static'' showing no eye movement at all.

2) PARTICIPANTS

We set up the survey as a website and disseminated the link to our university, research group, and beyond. A total of 73 participants finished the study. Participants could opt-in to take part in a raffle, and in total 33 were eligible by finishing the survey, opting in, and not being employed at our research group. We randomly distributed 10(30%) Amazon gift cards worth 10€ each.

All participants voluntarily engaged in the survey after being provided with detailed information about the study's objectives and procedures on the website. Before commencing the survey, participants were required to enter personal information, indicating their willingness to participate.

Figure 4 presents the demographics of those participants who completed the survey. Most participants are Germans (30.1%), and second most are Indians and Japanese (21.9%, 20.5%). The mean age of our participants is around 26, with most participants in the age group of $18-29$ (69.9%). There is an imbalance in the gender distribution, with most participants identifying as male (63.0%).

[7]Main activity means that in ''Speak'', John is talking most of the time, but also listening brief periods.

3) RESULTS

Although some participants did fail the reliability test, incorrectly selecting the static avatar or skipping it, we did not exclude participants from the analysis. We will establish in the following subsubsection IV-B4 that this does not influence the results but increases the statistical interpretability.

To evaluate each model's qualitative performance, we calculated the aggregated preference percentage and the pairwise binomial sign test with Bonferroni correction to estimate the validity of the results (following [19], [43]). The first is the number of times, a model was selected, divided by the times it was eligible, excluding ties. The latter compares the models' performances for all pairwise comparisons and calculates the statistical significance of the results, reporting the significance level $\alpha = 2 \cdot P(X \leq k | p = 0.5)$. Because the survey asked participants to select the avatar that was more **un-**natural, we inverted the results, to compute the acceptance rate, i.e. which avatar is more natural.

We collected a total of 1314 pairwise comparisons, stemming from 73 participants, presenting each 18 different pairs. In Figure 5 we present the results of the survey, using the two named metrics. The *Fuse*-model achieves the best results among all models with an approval rate of 67.0%, followed by the GAN models, with 55.3% and 43.7% for the CNN and RNN generators, respectively. The least preferred model is the *Lee*-model, which was our baseline, reaching an approval rate of 34.0%. From the pairwise comparison we see, that participants preferred both ML models over the *Lee*-model. This proves, that GANs are capable of beating the naïve procedural model. However, it is also evident that they are not able, to beat the *Fuse*-model with carefully designed rules, relying on the same data as the ML generators. Between the two GAN-based generators, the CNN model outperforms the RNN model, both directly, as well as compared to the *Fuse*-model, and is similar in performance against the weaker *Lee*-model. The α values are small enough to assume the statistical relevance of the results.

4) ABLATION

a: RELIABILITY TEST

First, we demonstrate why we chose to include all participants, regardless of their performance in the reliability test for the main results. In the reliability test, we saw, that the majority of participants were able to correctly identify the video exhibiting movement for rounds 1 and 3 (''Fuse-Static'' and ''Lee-Static'', respectively). However, in the second round (''Fuse-Lee'') only a small fraction (27.8%) selected ''Tie'', as it would have been correct. We speculate, that the movement in the *Lee* video was too small and participants didn't expect both videos to exhibit eye movement. Based on the results of the first and third rounds, we define five different filter modes, to select participants for analysis based on their performance in the reliability test, ranging from loose to strict filtering:

- **Mode 0:** Don't filter.

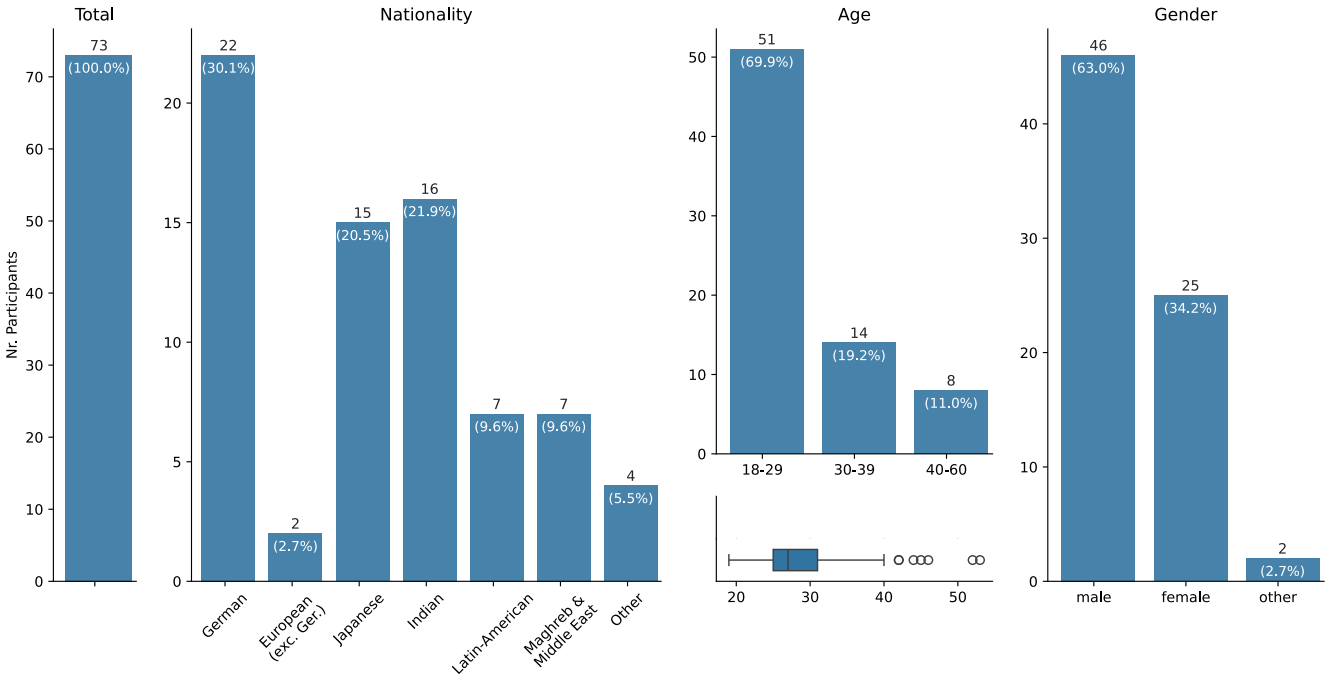


FIGURE 4. The demographics of all participants that finished the survey. Participants who did not finish were excluded from the analysis and raffle.

- **Mode 1:** Discard participants, selecting "Static" two times.
- **Mode 2:** Discard participants, selecting "Static" at least once.
- **Mode 3:** Discard participants, selecting "Static" at least once or "Tie" twice.
- **Mode 4:** Discard participants, selecting "Static" or "Tie" at least once.

Table 4 shows the aggregated preferences of each model per filter mode, together with the number of participants included in this subset and the average significance value $\hat{\alpha}$ among all six pairwise comparisons. It is evident that with stricter filtering, the number of participants and therefore the meaningfulness of the results decreases, which is reflected by an increase in $\hat{\alpha}$, albeit the performance results are very similar across all different filter setups. We conclude, that the performance in the reliability test is not related to the actual performance during the real survey. Therefore, we decided to report the results for all participants, regardless of their respective performance during the reliability study, to include as many participants as possible and increase the statistical validity of the data.

b: CONVERSATION MODE

Next, we compare the results for the three different conversation setups we had in our study: "Speak", "Listen" and "Dialogue". Table 5 reveals that the ordering between the different models is the same for all three modes, given as:

$$\text{Fuse} > \text{CNN} > \text{RNN} > \text{Lee}.$$

However, there are some differences between the individual results. During dialogues, the performance reflects the average results, but during speaking, the *Fuse*-model achieves worse results, with the CNN model close behind. For listening, RNN and *Lee* perform worse than in the other modes. Overall, we observe some variations in the exact preference rates, but it is hard to attribute those results to specific characteristics of the generators. Additionally, the ordering between the generators is the same for each individual mode and the overall results, proving their expressiveness.

5) QUESTIONNAIRES

Each participant was asked six times (once per pairing) to justify the selected video. This way we aimed to collect descriptive feedback on the factors that contribute to the perception of naturalness in eye movement.

In our analysis of participant feedback, several key observations emerged regarding their engagement with the study. Firstly, we noted a limited capacity among participants to focus on eye movements, likely due to its infrequency in everyday conversations. Many struggled to articulate precise reasons for their avatar selections, offering vague explanations such as "A felt more unnatural than B." This attention limitation also led to difficulties in detecting subtle movements, as seen in the reliability test where participants frequently misidentified eye movement. Despite the difficulty in pinpointing specific flaws, participants expressed strong discomfort with unnatural gaze behaviors, aligning with the Uncanny Valley theory's emphasis on realistic gaze representations.

TABLE 4. The aggregated preferences are almost identical for either filter mode. However, the number of participants and the statistical validity (increasing $\hat{\alpha}$) decrease with the strictness of the filtering.

	RNN	CNN	Fuse	Lee	Participants	$\hat{\alpha}$
Filter 0	43.7%	55.3%	67.0%	34.0%	73	0.00211
Filter 1	43.6%	54.8%	67.4%	34.5%	71	0.00521
Filter 2	43.1%	54.6%	67.5%	34.9%	66	0.01667
Filter 3	43.0%	55.2%	67.6%	34.3%	61	0.01321
Filter 4	42.3%	57.5%	67.3%	33.1%	40	0.01128

TABLE 5. Aggregated preference per conversation mode although there are some differences in the exact values between the individual modes, the overall trend stays the same as in the overall results.

	RNN	CNN	Fuse	Lee
Dialogue	41.5%	50.0%	73.0%	35.1%
Speak	49.8%	54.9%	57.9%	37.6%
Listen	39.9%	61.4%	70.1%	29.3%
All	43.7%	55.3%	67.0%	34.0%

Moreover, participants provided rich descriptive feedback, associating eye movements with character traits or mental states of the avatars, demonstrating the potential for gaze behavior to influence perceived personality. However, the varied and sometimes conflicting feedback hindered the confident assessment of individual models' strengths and weaknesses. Common criticisms included a lack of accompanying head movements and extremes in eye movement frequency (''too little'' or ''too much''), stressing the need for balance to avoid distracting viewers. Additionally, participants noted a disconnect between eye movements and speech content across all models, which is one of the major shortcomings of our models.

V. DISCUSSION

The results from the statistical analysis and survey show many similarities and some differences.

Comparing the machine learning models, the CNN was better than the RNN in both evaluations. It seems, that the higher stability provided by fixed window lengths produces more reliable output than the RNN trained on varying sequence lengths. With the CNN we were able to design an architecture, that computes timestep-wise and channel-wise features simultaneously. However, the RNN's flexibility to generate eye movements of arbitrary length is superior when we think of application scenarios. We tried several architectures but restricted ourselves to simpler models, including multilayer perceptrons, CNNs, and RNNs. A more extensive search for architectures like coupling CNN and RNN may reveal better-suited models. Another approach would be to use a rolling-window-like approach, letting the model predict the next several fixations and using the last few as input, without necessarily relying on an RNN (similar to [34]). Furthermore, we tried several handlings of the conversation mode label by the GANs, namely incorporating it to the input for both Generator and Discriminator, duplicating the Generators and Discriminators architecture and training one per label, and asking the Discriminator to additionally classify the presented data. We can think of more possibilities that could be evaluated in the future, like completely ignoring the label with the Discriminator.

In both rankings, the *Lee*-model performs worst. Its simplistic rule set ignores small eye movements completely and keeps the gaze fixated on the conversational partner most of the time. User feedback we gathered during the evaluation study revealed, that this constant staring is far from natural eye movement and sometimes even made observers feel uncomfortable. Nevertheless, [34] reported the *Lee*-model to outperform their approach. Hence it served as the baseline model for our study and was outperformed by every other model, especially our two GAN models.

Interestingly, the *Fuse*-model was worse in mimicking the statistics of the dataset and far behind the GANs' scores. However, the meticulously handcrafted rules found greater acceptance among study participants. This shows how blindly trying to mimic some statistics does not necessarily result in a reproduction of the underlying structures. We saw this most noticeably with the *Dembinsky*-model, which almost perfectly matched the statistical properties of real gaze data, but generated eye movement too unbelievable to be included in the user study.

Although the GANs did not beat the procedural *Fuse*-model in the current stage, they hold the most promising opportunities for future research. All models currently lack the opportunity to somehow incorporate the meaning of a given speech into the generation of eye movement. Context or speech can't be added to procedural models (without great effort and expert knowledge), whereas it is possible to include this feature in the GAN models. They can be trained using audio input or a transcript of the text, to generate eye movements precisely fitting the speech, which could include movements like looking up for thinking and looking at the partner when asking a question. We postulate that such a model would beat the statistical models relying solely on the conversation mode. We advocate for future studies to refine this aspect, relying on data that contains eye movement together with the corresponding audio, which is unfortunately not true for the data by [32]. Future work should also experiment with different data formats, directly predicting

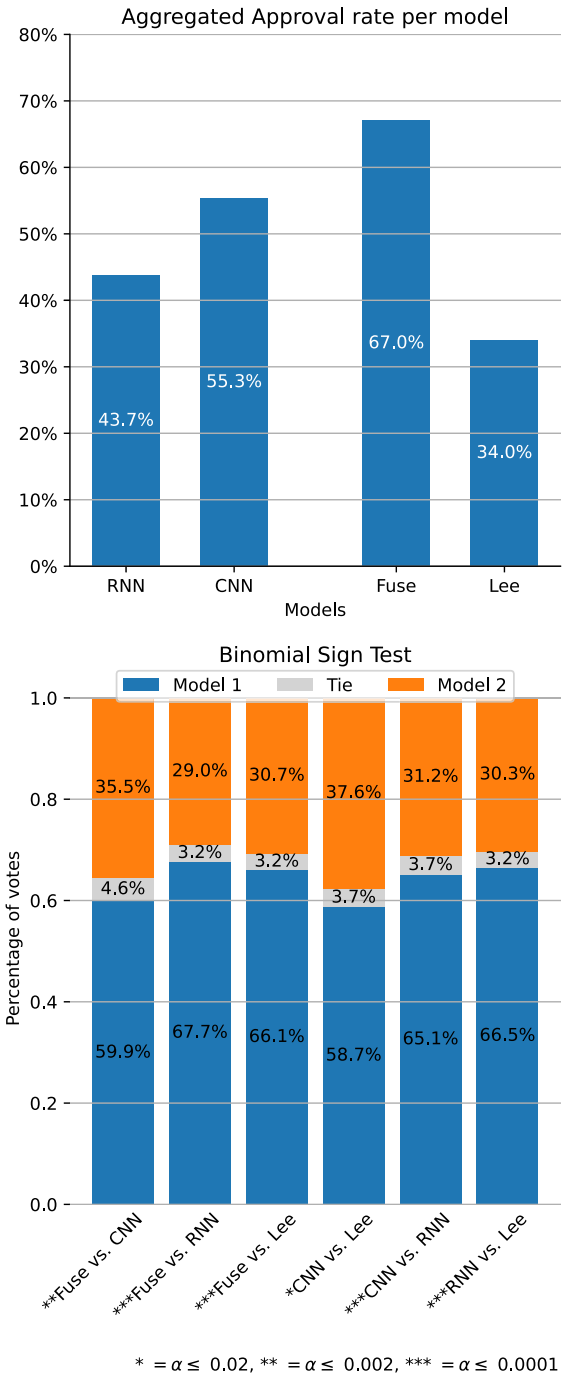


FIGURE 5. **The results of our qualitative user study. The upper one shows the aggregated approval rate per model (excluding ties), revealing a clear trend between the models, given as Fuse > CNN > RNN > Lee. Below reports the performance of each model's pairings with the associated significance value α. *Fuse* beats every other model, whereas *Lee* is beaten by every other model and CNN beats RNN.**

gaze position or gaze angles (like [34]), rather than relying on screen coordinates.

Finally, our visualization pipeline is currently not applicable to real-time visualization but can only be used offline. However, we have shown that the GANs are lightweight enough to perform online. Future work may build a complete framework, leveraging Natural Language Processing to generate speech, which will then be transformed into audio and eye movements in real-time.

VI. CONCLUSION

In this work, we trained two Generative Adversarial Networks (GANs) to synthesize eye movement in the form of fixation points for a conversational avatar. We showed, that both GANs, relying on different approaches, were capable of mimicking the statistics of the ground truth dataset, scoring higher than two of the procedural models. We assessed the quality of the generated gazes, by conducting a user study. 73 participants compared videos of conversational avatars during talking, listening, and dialogue, that relied on our two GANs and two procedural models. We showed that the GANs are able to outperform the procedural baseline model, reaching higher approval rates of 55.3% and 43.7% respectively, compared to 34% by the baseline model. However, a second procedural model with carefully designed parameters outperformed the GANs with an approval rate of 67.0%. Nevertheless, this work provides a novel step on the path of generating gazes for conversational avatars. We advocate for future researchers, to refine the machine learning generators to additionally incorporate audio features.

REFERENCES

[1] L. S. Bohannon, A. M. Herbert, J. B. Pelz, and E. M. Rantanen, "Eye contact and video-mediated communication: A review," *Displays*, vol. 34, no. 2, pp. 177–185, Apr. 2013.
[2] K. Watanabe, Y. Soneda, Y. Matsuda, Y. Nakamura, Y. Arakawa, A. Dengel, and S. Ishimaru, "DisCaaS: Micro behavior analysis on discussion by camera as a sensor," *Sensors*, vol. 21, no. 17, p. 5719, Aug. 2021.
[3] C. Chen, Y. Arakawa, K. Watanabe, and S. Ishimaru, "Quantitative evaluation system for online meetings based on multimodal microbehavior analysis," *Sensors Mater.*, vol. 34, no. 8, p. 3017, 2022.
[4] K. Watanabe, T. Sathyanarayana, A. Dengel, and S. Ishimaru, "EnGauge: Engagement gauge of meeting participants estimated by facial expression and deep neural network," *IEEE Access*, vol. 11, pp. 52886–52898, 2023.
[5] Z. Degutyte and A. Astell, "The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings," *Frontiers Psychol.*, vol. 12, Apr. 2021, Art. no. 616471.
[6] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psycholog.*, vol. 26, pp. 22–63, Jan. 1967.
[7] K. Jokinen, M. Nishida, and S. Yamamoto, "Eye-gaze experiments for conversation monitoring," in *Proc. 3rd Int. Universal Commun. Symp.*, New York, NY, USA, Dec. 2009, pp. 303–308.
[8] M. M. Egbert, "Context-sensitivity in conversation: Eye gaze and the German repair initiatorbitte?" *Lang. Soc.*, vol. 25, no. 4, pp. 587–612, Dec. 1996.
[9] S. Wohltjen and T. Wheatley, "Eye contact marks the rise and fall of shared attention in conversation," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 37, Sep. 2021, Art. no. e2106645118.
[10] D. Gergle and A. T. Clark, "See what I'm saying: Using dyadic mobile eye tracking to study collaborative reference," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, New York, NY, USA, Mar. 2011, pp. 435–444.
[11] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception," *Comput. Graph. Forum*, vol. 34, no. 6, pp. 299–326, Sep. 2015.

[12] M. Garau, M. Slater, V. Vinayagamoorthy, A. Brogni, A. Steed, and M. A. Sasse, "The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, Apr. 2003, pp. 529–536.

[13] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 98–100, Jun. 2012.

[14] P. Pataranutaporn, J. Leong, V. Danry, A. P. Lawson, P. Maes, and M. Sra, "AI-generated virtual instructors based on liked or admired people can improve motivation and foster positive emotions for learning," in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2022, pp. 1–9.

[15] T. Amemiya, K. Aoyama, and K. Ito, "Effect of face appearance of a teacher avatar on active participation during online live class," in *Human Interface and the Management of Information: Applications in Complex Technological Environments*, S. Yamamoto and H. Mori, Eds., Cham, Switzerland: Springer, 2022, pp. 99–110.

[16] Y.-L. Theng and P. Aung, "Investigating effects of avatars on primary school children's affective responses to learning," *J. Multimodal User Interface*, vol. 5, nos. 1–2, pp. 45–52, Mar. 2012.

[17] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10093–10103.

[18] G. Tian, Y. Yuan, and Y. Liu, "Audio2Face: Generating speech/face animation from single audio with attention-based bidirectional LSTM networks," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2019, pp. 366–371.

[19] M. V. Aylagas, H. A. Leon, M. Teye, and K. Tollmar, "Voice2Face: Audio-driven facial and tongue rig animations with cVAEs," *Comput. Graph. Forum*, vol. 41, no. 8, pp. 255–265, Dec. 2022.

[20] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven facial animation with temporal GANs," 2018, *arXiv:1805.09313*.

[21] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1398–1413, May 2020.

[22] S. P. Lee, J. B. Badler, and N. I. Badler, "Eyes alive," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 637–644, Jul. 2002.

[23] V. Vinayagamoorthy, M. Garau, A. Steed, and M. Slater, "An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience," *Comput. Graph. Forum*, vol. 23, no. 1, pp. 1–11, Mar. 2004.

[24] G. Bente, F. Eschenburg, and N. C. Krämer, "Virtual gaze. A pilot study on the effects of computer simulated gaze in avatar-based conversations," in *Proc. Int. Conf. Virtual Reality*, Beijing, China. Cham, Switzerland: Springer, Jul. 2007, pp. 185–194.

[25] G. Bente, F. Eschenburg, and L. Aelker, "Effects of simulated gaze on social presence, person perception and personality attribution in avatar-mediated communication," in *Proc. 10th Annu. Int. Workshop Presence*, Barcelona, Spain, Oct. 2007, pp. 207–214.

[26] M. Kipp and P. Gebhard, "IGaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions," in *Proc. Int. Workshop Intell. Virtual Agents*. Cham, Switzerland: Springer, 2008, pp. 191–199.

[27] W. Steptoe, R. Wolff, A. Murgia, E. Guimaraes, J. Rae, P. Sharkey, D. Roberts, and A. Steed, "Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, Nov. 2008, pp. 197–200.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014, pp. 1–11.

[29] X. Zhao, L. Wang, J. Sun, H. Zhang, J. Suo, and Y. Liu, "HAvatar: High-fidelity head avatar via facial model conditioned neural radiance field," *ACM Trans. Graph.*, vol. 43, no. 1, pp. 1–16, Feb. 2024.

[30] A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, A. Ghosh, and S. Zafeiriou, "AvatarMe++: Facial shape and BRDF inference with photorealistic rendering-aware GANs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9269–9284, Dec. 2022.

[31] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.

[32] D. Dembinsky, K. Watanabe, A. Dengel, and S. Ishimaru, "Eye movement in a controlled dialogue setting," in *Proc. Symp. Eye Tracking Res. Appl.*, Jun. 2024, p. 7.

[33] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "Vasa-1: Lifelike audio-driven talking faces generated in real time," 2024, *arXiv:2404.10667*.

[34] R. Canales, E. Jain, and S. Jörg, "Real-time conversational gaze synthesis for avatars," in *Proc. ACM SIGGRAPH Conf. Motion, Interact. Games*, New York, NY, USA, Nov. 2023, pp. 1–7.

[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[36] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[37] A. Ramdas, N. Trillos, and M. Cuturi, "On Wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, Jan. 2017.

[38] J. L. Hodges, "The significance probability of the Smirnov two-sample test," *Arkiv För Matematik*, vol. 3, no. 5, pp. 469–486, Jan. 1958.

[39] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder–decoder pre-training for spoken language processing," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2022, pp. 5723–5738.

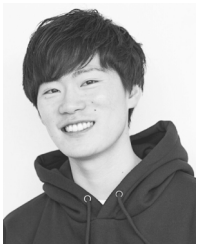
[40] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–7.

[41] B. H. Le, X. Ma, and Z. Deng, "Live speech driven Head-and-Eye motion generators," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 11, pp. 1902–1914, Nov. 2012.

[42] X. Ma and Z. Deng, "Natural eye motion synthesis by modeling gaze-head coupling," in *Proc. IEEE Virtual Reality Conf.*, Mar. 2009, pp. 143–150.

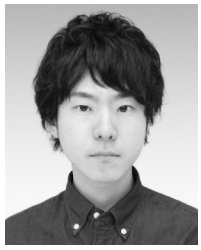
[43] P. Jonell, T. Kucherenko, G. E. Henter, and J. Beskow, "Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings," in *Proc. 20th ACM Int. Conf. Intell. Virtual Agents*, New York, NY, USA, Oct. 2020, pp. 1–8.

DAVID DEMBINSKY was born in Kaiserslautern, Germany, in 1999. He received the B.Sc. degree in computer science from the Technical University of Kaiserslautern (TUK), Germany, in 2021. He is currently pursuing the M.Sc. degree with the University of Kaiserslautern-Landau (formerly TUK, which merged with the University of Koblenz and Landau, in 2023, and formed RPTU). Since 2019, he has been a Research Assistant with German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern. His study focuses on machine learning and deep learning. His current research interest includes the employment and generation of animated avatars using machine learning.



KO WATANABE was born in Hiroshima, Japan, in 1994. He received the B.E. degree in mechanical engineering from Tokyo University of Agricultural and Technology, Tokyo, Japan, in 2017, and the M.E. degree from Nara Institute of Science and Technology, Japan, in 2019. He is currently pursuing the Ph.D. degree in computer science with the University of Kaiserslautern-Landau (appointed by the University of Kaiserslautern, in 2021, and its campus was merged with the University of Koblenz and Landau, in 2023). He was a Software Engineer with DeNA, Tokyo. His current research interest includes the investigation of technologies that augment human intellect.

ANDREAS DENGEL received the Diploma degree in CS from the University of Kaiserslautern and the Ph.D. degree from the University of Stuttgart. He is currently a Scientific Director of DFKI GmbH, Kaiserslautern. In 1993, he became a Professor in computer science with the University of Kaiserslautern (renamed to University of Kaiserslautern-Landau, since 2023), where he holds the Chair of Knowledge-Based Systems. Since 2009, he has been appointed as a Professor (Kyakuin) with the Department of Computer Science and Information Systems, Osaka Prefecture University. He was with IBM, Siemens, and Xerox Parc. He is a member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is a co-editor of international computer science journals and has written or edited 12 books. He is the author of more than 300 peer-reviewed scientific publications and supervised more than 170 master's and Ph.D. theses. He is a fellow of IAPR and received many prominent international awards. His main scientific emphasis is in the areas of pattern recognition, document understanding, information retrieval, multimedia mining, semantic technologies, and social media.



SHOYA ISHIMARU (Member, IEEE) was born in Ehime, Japan, in 1991. He received the B.E. and M.E. degrees in electrical engineering and information science from Osaka Prefecture University, Japan, in 2014 and 2016, respectively, and the Ph.D. degree in engineering (summa cum laude) from the University of Kaiserslautern, Germany, in 2019.

He has been a Project Professor with the Department of Computer Science, Osaka Metropolitan University, Japan, since 2023. In addition, he has been an Associate Director of Japan Laboratory of German Research Center for Artificial Intelligence (DFKI Laboratory, Japan), since 2023, and a Researcher with the Keio Media Design Research Institute, since 2014. He was a Junior Professor with the University of Kaiserslautern-Landau, Germany, from 2021 to 2023, and was a Senior Researcher with DFKI, from 2019 to 2023. His research interests include human–computer interaction, machine learning, and cognitive psychology with the aim of amplifying human intelligence.

Prof. Ishimaru's awards and honors include the Best Presentation Award at Asian CHI Symposium, in 2020, Poster Track Honorable Mention at UbiComp/ISWC, in 2018, and MITOU Super Creator, which is a title given to outstanding software developers (around ten people per year) by the Ministry of Economy, Trade, and Industry in Japan.