Context-Aware Head-and-Eye Motion Generation with Diffusion Model

Yuxin Shen[1] Manjie Xu[2] Wei Liang[1,2] *

[1]Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing, China
[2]Beijing Institute of Technology

Abstract

In humanity's ongoing quest to craft natural and realistic avatars within virtual environments, the generation of authentic eye gaze behaviors stands paramount. Eye gaze not only serves as a primary non-verbal communication cue, but it also reflects cognitive processes, intent, and attentiveness, making it a crucial element in ensuring immersive interactions. However, automatically generating these intricate gaze behaviors presents significant challenges. Traditional methods can be both time-consuming and lack the precision to align gaze behaviors with the intricate nuances of the environment in which the avatar resides. To overcome these challenges, we introduce a novel two-stage approach to generate context-aware head-and-eye motions across diverse scenes. By harnessing the capabilities of advanced diffusion models, our approach adeptly produces contextually appropriate eye gaze points, further leading to the generation of natural head-and-eye movements. Utilizing Head-Mounted Display (HMD) eye-tracking technology, we also present a comprehensive dataset, which captures human eye gaze behaviors in tandem with associated scene features. We show that our approach consistently delivers intuitive and lifelike head-and-eye motions and demonstrates superior performance in terms of motion fluidity, alignment with contextual cues, and overall user satisfaction.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

1 Introduction

Realistic virtual avatars play a crucial role in enhancing user experience within virtual environments [16]. Recent research has explored the design of virtual characters with natural behavior, believability, and physical appearances [36]. In particular, researchers have early on recognized the significance of eye gaze. Beyond its non-verbal communicative role in human interactions [35], gaze also functions as an indicator of human cognitive processes [42], assisting in guiding behavior [2], recognizing user tasks [17], and revealing intentions [44]. Efforts have, therefore, been directed toward generating natural eye gaze in various domains, such as games and films. For instance, the game Team Bondi's *L.A. Noire* [3] employs gaze behavior to judge the truthfulness of a Non-Player Character (NPC), where an evasive NPC might avert their gaze, implying deceit.

Two primary challenges exist in the realm of eye gaze generation. On the one hand, traditional methods have heavily depended on manual and procedural methods, which are both time-consuming and labor-intensive. For instance, in the video clips of Civilization VI, the eye gaze behavior of the sovereign is meticulously crafted by designers, expressing their discontent [9]. Numerous studies have employed statistical approaches to simulate the gaze behavior of virtual humans by collecting data from eye-tracking images or videos. For example, Lee et al. [26] propose a statistical gaze model based on the analysis of eye-tracking videos. However, these statistical methods often fall short in terms of flexibility and contextual adaptability, leading to the second challenge: crafting realistic gaze behavior demands an in-depth understanding of the context in which the avatar exists, as gazes respond to scene stimuli. Previous methods have occasionally resulted in the Uncanny Valley effect [33], wherein the virtual human elicits an eerie and unsettling impression from the viewer. While a number of studies have explored learning-based models to simulate eye gaze movements in virtual avatars [22], these methods tend to concentrate on simple tasks, lacking in applicability and robustness, thereby confining them to specific scenarios.

Figure 1: **The head-and-eye motions generated by our approach**. Given a scene and a pre-defined travel trajectory, our approach can generate diverse and natural gaze fixation sequences together with the corresponding head-and-eye motions.

To overcome these challenges, we propose a novel learning-based approach to automatically generate virtual characters' eye gaze and head movement in the simulated environment. Utilizing the conditional diffusion model, the produced head-and-eye motion sequences are harmoniously aligned with the scene in which the avatar is situated while ensuring natural and realistic motions. Specifically, we use the fixation sequence as an intermediate representation to decompose the problem into two distinct stages: a *Gaze Fixation Generation Stage* and a *Head-and-Eye Motion Generation Stage*. The first stage ensures the proper, context-aware gaze points within the scene, while the second stage assures natural motion generation. The *Gaze Fixation Generation Stage* navigates through the scene's complexity, analyzing static and dynamic elements to predict plausible fixation points that an avatar would naturally engage with. Utilizing a rich set of features encapsulating spatial, temporal, and semantic information, it determines where the avatar's gaze should be directed at different moments, ensuring that the generated eye movements are contextually relevant and convincing. In the subsequent *Head-and-Eye Motion Generation Stage*, our system, informed by the previously established fixation points and the head position sequence, computes rotation parameters for the avatar's eyes and head. We also include a smoothing process to help generate smooth and inherently coordinated motions.

To facilitate a more profound understanding of human eye-gaze behaviors within a given scene, we also introduce **C**ontext-**A**ware **H**ead-and-**E**ye Motion Dataset (CAHE), which encompasses human eye-tracking and head movement data within immersive virtual

*e-mail: liangwei@bit.edu.cn (Corresponding author)

scenes, complemented by exhaustive annotations regarding the low- and high-level features of the environment. Subsequently, we assess our proposed approach utilizing CAHE. Through comparative user studies and quantitative experiments, we demonstrate that our approach surpasses various baselines in generating head-and-eye motions within a given scene. The produced gaze behaviors are context-aware, closely mirroring human behaviors.

The proposed approach gives rise to a number of practical applications. In the gaming domain, our technique enriches the player experience by generating avatars whose eye and head movements interact with in-game objects. In addition, in virtual museums, our approach creates realistic virtual tour guides that use eye gaze and head movements to guide visitors' focus towards highlighted exhibits, fostering human-avatar interaction.

The main contributions of our paper are:

- We underscore the critical importance of avatar eye gaze in alignment with the contextual cues within the environment. This is fundamental for enriching immersive experiences and facilitating genuine interpersonal interactions within virtual environments.
- We present a cutting-edge two-stage diffusion-based method tailored for the automatic generation of context-sensitive eye gaze points seamlessly integrated with organic head-and-eye movements. To validate the effectiveness of our approach, we also present a new dataset, **C**ontext-**A**ware **H**ead-and-**E**ye Motion Dataset (CAHE), comprising users' eye-tracking and head movement data alongside extensive annotations related to various features in virtual environments.

2 Related Work

2.1 Context-Aware Behavior Synthesis

Behavior synthesis aims to create models that synthesize behavior automatically. Many researchers have been investigating behavior synthesis for virtual humans [11, 15, 24] or pets [29]. Understanding how characters' behavior associates with the scene context has also gained a lot of attention recently [54, 58]. Previous approaches focus on generating human behavior using Two-dimensional (2D) raw scene images. Cao et al. [5] introduce a learning framework to predict human motion given a single scene image. Wang et al. [51] propose a scene-aware generative method to model the distribution of human motion in the given scene, which is represented as an RGB image. Recently, several methods utilize Three-dimensional (3D) scene context to generate context-aware human behavior. Liang et al. [29] propose to synthesize high-level and context-aware virtual pet behaviors in a real scene, which is captured by a 3D sensor. SAMP [11] is a data-driven motion synthesis method to generate scene-aware motion sequences of a character given the voxel representation of a scene object. HUMANISE [52] is another generative method to generate diverse human motions which are physically plausible in interacting with the scene represented by an RGB-colored point cloud. Inspired by these generative approaches, we propose a context-aware head-and-eye motions generative model to generate more realistic and contextually aligned character behaviors.

2.2 Eye Gaze Research

Gaze behavior is an essential element of virtual characters, affected by diverse stimuli. Humans prefer to focus on specific areas of interest within visual scenes, filtering out irrelevant information [1, 4]. Factors like color, shape, size, and orientation significantly impact human gaze behavior [28, 55]. Warm hues attract more attention than cool tones [53], and stimuli with greater luminance and texture differences are more captivating [38]. Human attention tends to be captured by threat- and reward-related stimuli [27, 32], as well as distinct or abrupt sounds [31]. Research highlights a strong correlation between visual attention and task-relevant objects [37, 40], enhancing efficiency in goal achievement. In this work, we leverage these stimuli to generate head and eye movements of virtual avatars that are more realistic and closely resemble real humans.

However, designing gaze animations manually to guide virtual characters' attention toward these stimuli is time-consuming. Several researchers have been investigating training-free statistical methods. Lee et al. [26] and Yeo et al. [56] implement statistical models to generate the eye gaze animation. Some studies synthesize eye movements by modeling the relationship between gaze and head movements [30], or by mapping speech to eye movements [25]. Lan et al. [23] synthesize realistic eye movement behavior using a psychology-inspired generative model based on images and videos. With the development of deep learning techniques, Recurrent Neural Network (RNN) [22, 57] and Convolutional Neural Network (CNN) [18, 46] have proven to be very successful for the generation of eye movement. Recently, another approach applies a Variational Autoencoder (VAE) model to generate synthetic eye-tracking data based on images [7]. However, all these models neglect the high-level features and nuances of scene semantics. Our method empowers virtual characters to discern scene semantics using both low- and high-level panoramic segmentation images and adapt their eye gaze and head behaviors accordingly across varied scenes.

2.3 Diffusion Generative Model

Diffusion models [47] are a type of generative model that mimics the process of creating complex data distributions. They achieve this by sequentially applying reversible transformations to a basic initial distribution, gradually shaping it into the targeted intricate data distribution. Diffusion-based image generation models [13, 48] have become famous over the past few years. Dhariwal et al. [6] propose a conditional diffusion model by introducing Classifier-Free Guidance technical. The classifier-guide diffusion can balance fidelity and diversity and achieve better results [34]. It also significantly propels the human motion domain. Tevet et al. [49] propose the Motion Diffusion Model, which is a classifier-free diffusion model for motion generation. Huang et al. [21] present a general conditional generative model for motion generation. Drawing inspiration from these generative models, we employ a transformer-based diffusion model informed by our annotated CAHE dataset. We also adopt a classifier-free approach, enhancing both the quality and diversity of samples.

3 Context-Aware Head-and-Eye Motion Generator

We introduce a head-and-eye motion generator capable of generating sequences of realistic eye gaze and head movements conditioned on the scene in which the avatar is situated. As shown in Figure 2, our key idea is to leverage context-aware fixation points: first generating a fixation sequence from low- and high-level panoramic feature segmentations and then estimating the head-and-eye motion from the head position and previously generated fixation sequence. Our approach uses various features to capture spatial, temporal, and semantic details, enabling realistic and contextually relevant eye gaze and head movements.

3.1 Fixation Sequence Generation

Inspired by the recent success of the diffusion model in image generation [43], we deploy a Fixation Diffuser to generate fixation sequences with pre-defined length (five frames) conditioned on the given scene. Specifically, the scene is extracted into 11 unique feature segmentations and the Fixation Diffuser takes five frames of images as input, encompassing both low- and high-level panoramic feature segmentation.

Framework We use a conditional diffusion model based on the Motion Diffusion Model (MDM) [50] to generate the fixation sequence. The latent representations of the scene are extracted using an ImageNet [45]-pretrained Resnet50 [12] and then fed into the

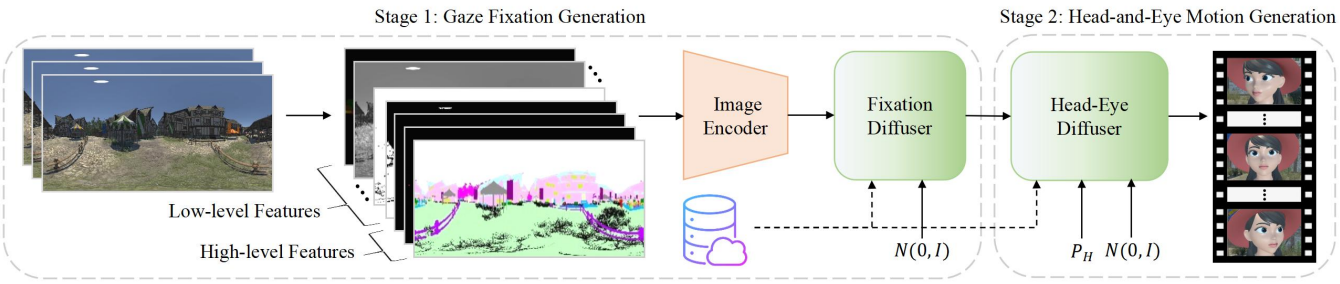

Figure 2: **The pipeline of our two-stage context-aware generation approach.** The approach first takes low- and high-level feature segmentations as input and predicts the fixation sequence using the Fixation Diffuser and smooth algorithm. The predicted fixation points are then fed to a Head-Eye Diffuser, followed by a similar smooth algorithm to generate the head-and-eye motion of the avatar.

generation module as condition code. To note, diffusion models are the latent variable models that map to the latent space using a fixed Markov chain. Conditional diffusion models work by destroying training data through the successive addition of Gaussian noise in the forward diffusion process and then learning to recover the data in the reverse denoising process. In the inferring phase, the data is generated by passing randomly sampled noise and condition code following the denoising process.

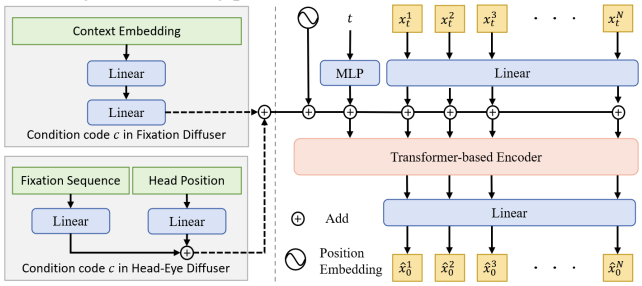

Figure 3: **Model architecture of the denoising network.** We use a transformer encoder-only architecture in the denoising network.

Instead of predicting ε_t as formulated by [13], we predict the signal $\hat{x}_0$ by using a straightforward transformer [41] encoder-only architecture in the denoising network which is shown in Figure 3. The transformer architecture is temporally aware, enabling learning arbitrary length motions [39, 49]. The condition code c, noise time-step t, and the noise input x_t of each time-step t are projected to the transformer dimension by a separate feed-forward network and concentrated with a standard positional embedding. Each frame of the noised input x_t is linearly projected into the transformer dimension. These embedding codes are then fed to the encoder. The output of the transformer is projected to the original fixation sequence dimensions by the feed-forward network and estimate x_0.

We train our transformer encoder using classifier-free guidance [14], which can trade-off diversity and fidelity. In practice, the encoder learns both the conditioned and the unconditioned distributions by randomly setting $c = \varnothing$ for 10% of the samples. When sampling the encoder G in an iterative manner, we predict the clean sample $\hat{x}_0 = G(x_t, t, c)$ and noise it back to x_{t-1}. This process is repeated from $t = T$ until x_0 is obtained.

Learning To train the network, we use our annotated CAHE dataset. Following random selection, we split the dataset into a separate training and evaluating dataset. The feature segmentations in the training set have a size of 512×256 and are fed to the network. Our model is trained with a batch size of 128 for 4000 epochs. We set the initial learning rate to $1e-5$.

Loss The object function $\mathscr{L}_{simple}$ of the diffusion model we use is following [13] to predict the signal itself, *i.e.*, $\hat{x}_0 = G(x_t, t, c)$.

$$\mathscr{L}_{simple} = E_{x_0 \sim q(x_0|c), t \sim [1,T]}[||x_0 - G(x_t, t, c)||_2^2] \tag{1}$$

In the above objective function, $G(x_t, t, c)$ represents the predicted signal, which is equal to $\hat{x}_0$. t is the time step of the reverse denoising process and c is the condition code. $q(x_0|c)$ denotes the predicted signal x_0 given the condition code c.

To train the fixation generation module, we employ the Mean Square Error (MSE) $\mathscr{L}_{mse}$ to force the predicted fixation sequence close to its corresponding ground truth. To generate smooth fixation sequences, we also use the L2 loss term $\mathscr{L}_{dis}$ to ensure that the distance between consecutive frames is reasonable.

$$\mathscr{L}_{mse} = \frac{1}{N} \sum_{i=1}^{N} ||P_i - \hat{P}_i||_2^2 \tag{2}$$

$$\mathscr{L}_{dis} = \frac{1}{N-1} \sum_{i=1}^{N-1} ||(P_{i+1} - P_i) - (\hat{P}_{i+1} - \hat{P}_i)||_2^2 \tag{3}$$

where N is the length of the fixation sequence. P_i and $\hat{P}_i$ represent the ground truth and predicted fixation point at frame i.

Overall, the training loss of our fixation sequence module is

$$\mathscr{L} = \mathscr{L}_{simple} + \mathscr{L}_{mse} + \mathscr{L}_{dis} \tag{4}$$

3.2 Head-and-Eye Motion Generation

Predicting both eye direction and head rotation from the head position and fixation point is not a one-to-one mapping problem due to the various combinations of eye direction and head rotation that can be used to focus on the same point. We also formulate the task using a conditional diffusion model called Head-Eye Diffuser. The architecture of the Head-Eye Diffuser is similar to the Fixation Diffuser. The inputs of the diffuser consist of the head position sequence $P_H = (x_H, y_H, z_H)$ and fixation points $P_F = (x_F, y_F)$. The Head-Eye Diffuser generates five frames of eye gaze and head movement parameters, matching the length of the input. We use a separate feed-forward network to map the input data to the transformer dimension. This is then passed into the diffuse process as conditional code c. We train the Head-Eye Diffuser on our annotated dataset CAHE and use the same training setting as the Fixation Diffuser.

3.3 Smooth Processing

The direct outputs of both diffusers are sequences of a pre-defined length, specifically five frames. To create longer, continuous sequences of fixation or head-and-gaze motion, we concatenate these short segments. Additionally, we introduce a smoothing algorithm designed to eliminate noise, such as unrealistic jitter, ensuring a more natural and coherent sequence.

The main idea of the algorithm is to use Dijkstra's algorithm to find the shortest path from the source node to a destination node in a map. In our approach, a graph node represents a generated fragment containing a fixed-length sequence, and the edge represents the loss between two nodes. Firstly, to generate the graph, we generate n sequence fragments using our diffuser for each fragment of the entire sequence. Then, we calculate the loss between the last frame of fragment t and the first frame of fragment $t+1$. To prevent unchanged head-and-eye motion of virtual avatars, we experimentally set the

Figure 4: **Demonstration of immersive virtual scenes included in the proposed CAHE dataset.** Three different virtual scenes with varying styles are used for data collection: the forest (left), the village (middle), and the town (right).

range of loss to avoid the distance of consecutive fixation points or the gaze-and-head parameters being too small. If the loss value of an edge falls within the pre-defined range, it will be retained while discarded. Note that we set different loss ranges for the outputs of the Fixation Diffuser and Head-Eye Diffuser. The diagram of the smoothing process can be found in supplementary material.

In the searching phase, we select the first and the last fragment of the entire sequence. Our algorithm finds the shortest path among the paths from the source node to its adjacent nodes first. The terminal node of the path is then regarded as an intermediate node. The algorithm then searches for the following shortest path from the intermediate node to its adjacent nodes. The iteration continues until every node is traversed. The total loss between the starting fragment and the destination fragment is obtained.

We utilize different loss functions for both stages to constrain the fixation and head-and-gaze sequences separately. In stage 1, the smoothness loss L^t_{fix} between two frames is the distance of two adjacent fixation points as follows:

$$L^t_{fix} = \sqrt{(x^{t+1}_i - x^t_{i+N})^2 + (y^{t+1}_i - y^t_{i+N})^2} \tag{5}$$

where x^t_i and y^t_i denote the horizontal and the vertical value of the fixation point at frame i for the t^{th} fragment. N represents the length of each fragment.

In stage 2, the overall smoothness loss comprises a head rotation loss and left and right eye forward loss. We use the quaternion representation to express the head rotation, and for both eye rotations, we use the forward direction. We measure the distance between the rotation quaternions of two successive frames for head rotation loss. The objective of the head rotation loss, denoted as L^t_{head}, can be expressed as follows:

$$L^t_{head} = \arccos(Q^{t+1}_i \cdot Q^t_{i+N}) \tag{6}$$

where Q^t_i denotes the head rotation quaternion at frame i for the t^{th} fragment.

Similarly, for eye rotation loss L^t_{eye}, we compute the angle distance of the eye forward direction between consecutive frames, as shown in the following functions:

$$L^t_{eye} = \arccos(D^{t+1}_i \cdot D^t_{i+N}) \tag{7}$$

where D^t_i denotes the eye forward direction at frame i for the t^{th} fragment.

For stage 2, the overall loss function of the smooth algorithm can be summarized as follows:

$$L^t = L^t_{head} + L^t_{eyeL} + L^t_{eyeR} \tag{8}$$

After searching the shortest path between the source fragment and the destination fragment, we choose the lowest loss path as the output of our smoothing processing.

4 Context-Aware Head-and-Eye Motion Dataset (CAHE)

We present **C**ontext-**A**ware **H**ead-and-**E**ye Motion Dataset (CAHE) to facilitate our study on eye gaze and head movements and to aid in the training of models. CAHE is a dataset of human head and eye motion that provides humans' eye-tracking and head movement data in immersive virtual scenes, also containing extensive annotations about low- and high-level feature segmentations in environments.

Figure 5: **The dataset collection processing.** The red arrows indicate the path designated for the user to follow. We highlight four user positions, where the yellow dots represent the user's fixation points projected in the panoramic image. We also display the user's real-time eye direction and head pose.

4.1 Scenes

The CAHE encompasses three distinct immersive virtual scenes, comprising the forest, the village, and the modern town, as depicted in Figure 4. Each scene exhibits a unique style and is constructed using the Unity Engine. Given the varying impacts of static and dynamic stimuli on human visual attention, we also incorporate both stationary and dynamic objects within these environments, focusing on elements routinely utilized in virtual environment research. In these settings, static entities such as stones, houses, and bridges are strategically dispersed, and vegetables are randomly placed when creating a scene. Dynamic elements comprise four main categories: plants, rivers, animals, and humans. We create a Unity script to simulate wind effects, influencing the movement of grasses, trees, and river currents. We source 3D models from the Unity Asset Store and Mixamo to model humans and animals. We set their movement paths using our custom Unity script, enabling them to wander randomly within the virtual environments. To cater to the intricacies of the human visual system, we introduce variations in color, texture, and shape for similar types of objects. We employ three types of virtual humans in each virtual scene: strangers, acquaintances, and enemies. Enemies are strategically positioned in hidden locations, like behind bushes or houses. Enemies also appear and disappear randomly.

4.2 Apparatus

Data in CAHE is collected in a 4m $\times$ 4m acoustically-isolated laboratory room, with a Varjo XR-3 HMD to display the virtual environment. Participants' eye gaze and head movement data are collected at a high resolution of 2880 $\times$ 2720 pixels, a refresh rate of 90 Hz, and a field view of 115°. Although the HMD offers inside-out tracking via RGB video pass-through cameras, we install two SteamVR Base Stations 2.0 in the corner of the room to enhance the tracking accuracy. The virtual scenes are hosted on Intel(R) Core(TM) i9-12900KF CPU and Nvidia GeForce RTX 4080 GPU.

4.3 Head-and-Eye Motion Data Collection

We recruit 15 participants from our existing participant pool to help create CAHE. Participants' ages ranged from 18 to 28 years ($M = 23.2$, $SD = 1.97$), with 3 females and 12 males, and are paid an hourly wage of $20. All participants have normal or corrected-to-normal vision; 13 wear glass, and 2 wear contact lenses during the experiment. None of the participants report visual or vestibular disorder, such as color blindness or dyschromatopsia. 15 participants have prior experience using the Xbox controller, with 11 (73%) having previously encountered the Virtual Reality (VR) HMD in a VR course. Tutorials are prepared to acquaint subjects with the VR HMD and Xbox controller. Before starting the experiment, the participants are requested to complete the one-point calibration. We strategically arrange the HMD cables to reduce potential disturbances.

Table 1: Feature segmentations used in our approach.

#	Features	Description
1	Color	Warm color in raw panoramic image is masked with white, cold with black.
2	Grayscale	Grayscale image is extracted from raw panoramic image.
3	Category	Different types of objects are masked in different colors.
4	Enemy	Enemy is masked in white color, and other humans in black.
5	Strangers and acquaintances	Acquaintances are marked in blue, strangers in white, and background in black.
6	Dynamic objects	Dynamic objects whose position did not change are masked in black color, and others in white.
7	Moving objects	Moving objects whose position change are masked white color, and others in black.
8	Speed	Objects having different speeds are masked in different colors.
9	Orientation for dynamic objects	The optical flow image of dynamic objects is extracted from the dynamic feature segmentation image.
10	Orientation for moving objects	The optical flow image of moving objects is extracted from the moving feature segmentation image.
11	Voice	Objects with normal and abnormal sounds are masked in white and blue color, respectively, with others in black.

Each participant is randomly assigned one of the immersive virtual scenes. Their task involves identifying lurking enemies while exploring the scene, with actions confined to rotating their heads and patrolling along designated paths. Striking an enemy makes it vanish, increasing the reward given to the participant by one point. Participants receive penalties when they make mistakes. The collection processing is shown in Figure 5.

After exploring the virtual scenes, participants are asked to fill out a post-experiment questionnaire detailing the objects that captured their interest, which serves as a metric for assessing performance.

4.4 Data Processing

We process the eye gaze and head motion data collected from the Varjo headset to extract meaningful patterns and facilitate future research. Besides, we employ the Unity Engine offline to obtain both low- and high-level feature segmentations per frame. It should be noted that although the sampling frequency of the HMD is 200 frames/second, we down-sample the gaze and head data as well as the feature segmentations to 10 frames/second to maintain a balance between data resolution and computational efficiency.

Eye and Head Data Extraction All gaze data is transformed from the Unity world coordinates to head coordinates. The gaze direction of each eye is represented as normalized vectors in the head coordinate system. Utilizing the gaze ray's origin and direction, our custom Unity scripts capture the 3D fixation points within the VR space. To obtain the projection fixation points in 2D panoramic images, we use a coordinate conversion method from world coordinates to image coordinates. We also record the status of both individual and combined eye states to identify instances of blinking and unreliable data. A status of 0 indicates unavailable eye-tracking data, which could be due to the transfer problem of the device or blinking, while a status of 2 indicates valid data. Ultimately, we capture gaze data for a total of 13 distinct parameters, which are stored in CSV format. Please refer to the supplementary materials for the description and notations of data collected from the HMD.

Multi-Level Feature Segmentations Feature segmentations provide coarse-grained and fine-grained visual features of the scene. We select 11 unique features and generate the corresponding feature segmentations from the ground truth of the scene in Unity. Further details on the feature segmentations are provided in Table 1. We categorize these features into two types: semantic-related and non-semantic. Specifically, Features 3 to 5, which correspond to particular object categories, are semantic-related. In contrast, the rest are classified as non-semantic. Additionally, these features are subdivided into dynamic and non-dynamic categories. Features 6 to 11 are dynamic, considering their changeable nature, whereas the others are non-dynamic. Notably, voice is deemed a dynamic feature, based on our assumption that only moving entities like animals, humans, or rustling grass can produce sound. We employ these dual classification strategies in our ablation study for a comprehensive analysis. A cubemap consists of six square textures that together depict the reflections on an environment from a central viewpoint. Each face of the cubemap corresponds to a view along one of the world's axes: up, down, left, right, forward, and back. Every face covers a 90° field of view in both horizontal and vertical directions. Using cylindrical projection, the images from all six directions are merged into a panoramic format. Therefore, given the recorded head and eye gaze data, we first set up separate cameras within the VR scenes to capture images in all six directions based on the provided head and gaze data. Subsequently, we employ image stitching technology to seamlessly combine these images into a single panoramic view. Of all the attributes, the color segmentation is annotated from raw panoramic images. The grayscale segmentation is derived by converting the raw RGB image to grayscale. The orientation features of dynamic and moving objects are extracted using the Gunnar Farneback algorithm [8] from dynamic and moving objects' segmentation, a method for computing dense optical flow. Other segmentations are captured from the VR scenes using our custom Unity script.

Data Analysis Due to the differences in each participant's gaming experience, we analyze the collected scores to avoid inaccuracy or invalid data. To evaluate participants' performance, we compute the percentage of enemies found by users. The results show that no outlier data are to be removed. Please see the supplementary material for more details about the analysis of user performance.

In total, CAHE contains approximately 150 minutes of head-and-eye motion data, including scene information, eye-tracking data, and head-movement trajectory in three virtual environments. Each trial data contains gaze data and 11 types of feature segmentations at a sampling rate of 10Hz. CAHE can be publicly accessed at https://sites.google.com/view/context-aware-generation.

5 Quantitative Experiment and Analysis

We use 105 min data (70%) from CAHE dataset for training while retaining the remaining data for validation purposes. We conduct experiments on the three scenes: forest, village, and town. The three scenes are allocated equal portions in the validation set (each 10% in the whole dataset). The collected data of the "town" virtual scene only existed in the validation dataset to evaluate the out-of-distribution generation of our approach. Firstly, the trajectory of the virtual character is randomly chosen from the validation dataset. The feature segmentations are fed into the Fixation Diffuser to generate multiple short-term fixation sequences. We use the smooth algorithm to obtain an entire sequence. The fixation points are visualized in the panoramic images. Secondly, given the head position and fixation point, the head-and-gaze motions are synthesized using the Head-Eye Diffuser, as well as smooth processing. Finally, the generated head-and-eye parameters with a specific length are fed into a virtual character in the virtual scene.

Our approach generates diverse and realistic head-and-eye motions in real time using segmentation features and head position input. It is worth noting that the virtual character primarily directs its attention towards concealed areas, such as bushes, corners, and windows, where enemies are more likely to hide. We also find that

(a) Ground Truth (b) Ours (c) W/O-Semantic (d) W/O-Dynamic

Figure 6: **Examples of fixation points generated by different ablation models.** We show fixation points generated with (a) Ground Truth, (b) Ours, (c) W/O-Semantic features, and (d) W/O-Dynamic features. The yellow box indicates the object that the user is focusing on.

Table 2: Angular errors on the validation dataset and cross-user.

	Total scenes	Overlapping scenes	New scene	Cross-User
AE	$1.14^\circ \pm 0.82^\circ$	$0.99^\circ \pm 0.73^\circ$	$1.53^\circ \pm 0.91^\circ$	$1.13^\circ \pm 0.82^\circ$

the virtual character sometimes looked at dynamic or brightly colored objects, not just the objects related to the task. The avatar's head and eye movements are natural and robust in different virtual environments. The direction of the head and eye movements align with the fixation point. Please see the supplementary video for the visualization of the head-and-eye motions used in our experiments.

We evaluate the effectiveness of our approach quantitatively and conduct an ablation study to verify the selection of input features and the loss design of our architecture.

5.1 Metrics

Angular error (*AE*) In line with previous studies [18–20], we employ the angular error between the ground truth and the predicted fixation position as one evaluation metric for the Gaze Fixation Generation Stage. A smaller angular error indicates a higher accuracy in prediction performance.

Cosine similarity We also leverage a distance measure to explore the influence of semantic and dynamic features for the Gaze Fixation Generation Stage. Specifically, we employ cosine similarity to compare the object distributions observed by the fixation points in both the ground truth and our predictions. A higher cosine similarity indicates a closer alignment between the model's predictions and the object distribution in the ground truth.

Orientation errors (*OE*) To evaluate the effectiveness of the Head-and-Eye Motion Generation Stage, we use the orientation errors of the head, denoted as O_{head}, and the eyes, denoted as O_{eyes}, as key metrics. These errors are quantified by computing the Frobenius norm of the difference between the predicted and ground truth rotation matrices: $||R_{pred}R_{gt}^{-1} - I||_2$. Here, $Rpred$ represents the predicted rotation matrix, and R_{gt} denotes the ground truth rotation matrix. A smaller value of this orientation error indicates superior model performance.

5.2 Gaze Fixation Generation Evaluation

We partition the validation dataset into two sections: 'village & forest' and 'town.' The Gaze Fixation Generation Stage is evaluated using angular error, a standard metric for fixation prediction. This error is calculated for total, overlapping, and new scenes in the validation datasets, providing insights into our approach's performance and generalization ability.

To examine our model's generalization across different users, we conduct a three-fold cross-user evaluation. The dataset is evenly divided into three parts based on user variance, with two parts used for training and the third for testing. This process is repeated three times, rotating the test set in each iteration. The results from these tests are aggregated, and we compute the mean and standard deviation of the angular errors to assess overall performance and consistency.

The outcomes are detailed in Table 2. Our findings reveal a slight decrease in performance on new scenes compared to overlapping ones, aligning with our expectations due to the presence of novel objects in the new scenes. However, this decline is minor, underscoring the robust generalization capability of our method. Overall, our approach demonstrates strong performance across the entire validation dataset. Moreover, it shows impressive cross-user performance, nearly mirroring the results obtained on the full dataset. These results highlight the robustness of our model and its ability to adapt to varying user behaviors.

5.3 Ablation Study

We conduct two ablation experiments: the first feature ablation study validated the effectiveness of the feature selection on both semantic- and dynamic-related information for our approach, and the second loss ablation study analyze the influence of distance loss in both stages of our architecture. The analysis of feature ablation further reveals the elements that absorb human attention more easily.

5.3.1 Feature Ablation

We conduct a semantic feature ablation by removing all the semantic features from the model while keeping other elements intact. To ensure a balanced comparison, we maintain consistent architecture and parameters with the original model, with the only alteration being in the linear layer during the feature extraction phase due to changes in input dimensions. The training configuration remains consistent with our primary architecture.

Semantic feature ablation To assess the impact of the semantic features, we ablate them collectively. For a balanced comparison, we maintain consistent architecture and parameters. The only variation is in the use of the linear layer of the feature extraction phase, necessitated by a change in the input dimensions. The training setting is the same as our architecture. Then, we count the number of objects in scenes based on gaze points predicted by two different approaches, as well as the ground truth. We convert the counts of different objects into a category vector and normalize it.

Table 3: Angular errors for our model and two feature ablated models.

	Ours	W/O-Semantic features	W/O-Dynamic features
AE	$1.14^\circ \pm 0.82^\circ$	$1.53^\circ \pm 0.85^\circ$	$1.37^\circ \pm 0.82^\circ$

Dynamic feature ablation A parallel experiment is executed for dynamic features. By systematically excluding the dynamic features and maintaining the rest of the model as constant, we intend to assess the effect of dynamic features on our approach. Again, the architecture, parameters, and training conditions are consistent to guarantee a fair comparison. We count the frequency at which users focused on dynamic objects in both the original and ablated models, as well as ground truth. The counts of different dynamic objects are transformed and normalized into a dynamic vector.

Result analysis We compute angular errors for two ablated approaches and our approach, which are shown in Table 3. Our approach outperforms all ablated approaches by a large margin, with an improvement of 25.5% for the semantic feature ablated model and 16.8% for the dynamic feature ablated model. The results for cosine similarity metric indicate that our approach (96.9×10^{-2}) achieved a higher cosine similarity than the semantic feature ablated approach (96.5×10^{-2}). Similarly, our approach (97.1×10^{-2}) achieves a higher cosine similarity than the non-dynamic approach (96.8×10^{-2}). These disparities further underscore the critical role of semantic and dynamic features in our approach's performance and the influence of different objects on guiding user attention.

We show qualitative results in Figure 6, illustrating the outputs of three models in the same head position and environmental context. In this case, when removing the semantic features, the avatar more easily focuses on the concealed place without semantic objects like humans. Similarly, the avatar can not focus on the dynamic objects when we remove the dynamic features, which is inconsistent with the ground truth. The result also emphasizes the importance of careful feature selection to mimic human attention patterns in virtual environments accurately.

Comparing results between our approach and ground truth, we notice that our approach generates diverse outcomes, which is attributed to the classifier-free guidance mechanism. In this case, user focuses on a person in the ground truth and our approach directs attention to another person in the village scene. In the town scenario, our approach directs attention toward the moving object (enemy), which is similar to ground truth (chicken). This illustrates that our model can simulate human attention patterns in diverse scenarios.

Our approach establishes a correlation between these attribute features of scenes and the eye gaze and head movements of a virtual avatar by extracting both low- and high-level semantic information features from the scenes, rather than relying on a direct mapping from underlying scene features to the output. When introducing a new scene, our approach can extract these features from the scene, thereby ensuring the generalization of our approach.

5.3.2 Loss Ablation

To validate the effectiveness of the distance loss Equation 4 in the training phase, we generate results after removing the distance loss. We use the same architecture and parameters for a fair comparison. We compute the angular error of the fixation points for both our approach and the ablated model. We then compare the orientation errors, assessing the head and eye rotation generated by the Head-and-Eye Motion Generation Stage. This evaluation takes our predicted fixation points and the ground truth fixation points as input. Table 4 shows that the design of the distance loss for Stage 1 is effective to improve fixation prediction results. We showcase that our approach yields more accurate head and both eyes rotation prediction results, compared to the ablated model. We also show that the ground truth fixation points improve the head rotation prediction, indicating that by developing methods that predict more accurate head rotation, the full eye gaze and head movements prediction can be further improved.

Table 4: Loss ablation study for our model and the loss ablated model.

	Ours	W/O-$\mathscr{L}_{dis}$
AE	$1.14^\circ \pm 0.82^\circ$	$1.18^\circ \pm 0.83^\circ$
O_{head}	1.539	1.543
O_{eyes}	0.269	0.270
O_{head}/GT	1.127	1.146
O_{eyes}/GT	0.257	0.258

We further check the visualization result of the fixation point and head-and-eye motions. The lack of distance loss leads to sudden moving between the consecutive fixation points, thus making the avatar's eye gaze and head movements unnatural. Our original approach generates a smoother fixation sequence and more realistic head-and-eye motion. An example of the compared effect for both approaches can be found in Figure 7.

6 User Study

We conduct a user study to validate the effectiveness of our approach and investigate whether the synthesized head-and-eye motion is realistic and reasonable. We compare the head-and-eye sequence generated by our context-aware approach to the ones generated by two other approaches. The compared approaches are:

Free-viewing approach Performing a given fixation point at the location of an object that is randomly selected using a uniform distribution among the possible focusing objects. For example, in the forest scene, the fixation point could be on 16 possible objects, such as a tree, rabbit, bridge, or enemy. If the fixation point is initiated, the virtual character might focus on the tree, which is chosen with a probability of 0.06. The virtual character's eyes move to the object at a random speed, followed by the head-turning at a speed determined by the distance and angle. After holding its gaze on the tree for a random amount of time, the character selects the next gaze point.

Procedural approach The procedural approach simulates a virtual character's periodic eye gaze and head movements. It begins by setting an initial head direction and randomly generating five gaze directions. The character then focuses on a point in the first gaze direction. The head-and-eye movement strategy used in this approach is similar to the free-viewing approach. Once the character has looked in one direction for a random amount of time, it will move on to the next direction. This approach is commonly seen in games, e.g., periodic head-and-gaze movements of characters in Red Dead Redemption [10].

6.1 Experiment Settings

Subjects We recruit 20 subjects to take part in the user study. The subjects include 16 females and 4 males whose ages ranged from 21 to 50. All the subjects report normal or corrected-to-normal vision with no color-blindness. 7 subjects report that they do not have experience in 3D movies, animation, or games.

Data We use each approach to generate head-and-eye motions of a virtual character using the same head position sequence. We then collect a total of 10 groups of animation clips. Each group consists of animation clips generated by three different approaches. The animation clip contains the avatar's head-and-eye motion video from a third-person perspective, as well as a first-person perspective video of the virtual avatar where fixation points are visualized in real time. Please refer to the supplementary material for a snapshot of the generated animation clips.

Procedure Participants are presented with a series of animated clips and tasked with rating the generation. They are randomly assigned six videos, featuring two from each virtual scene, with the entire activity taking approximately 15 minutes to complete. The

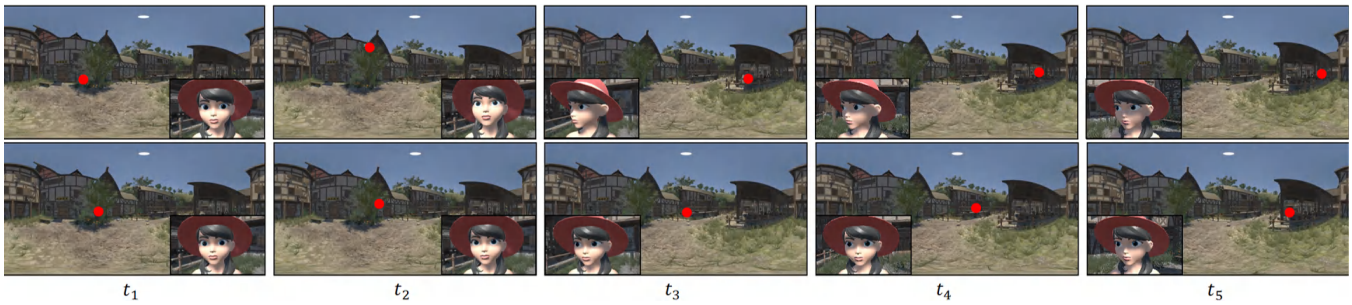

Figure 7: **Comparison result of loss ablation study.** Fixation sequence and head-and-eye motions of a virtual avatar generated by the approach without the distance loss (**the top**) and our approach (**the bottom**). The lack of distance loss leads to sudden moving between the consecutive fixation points.

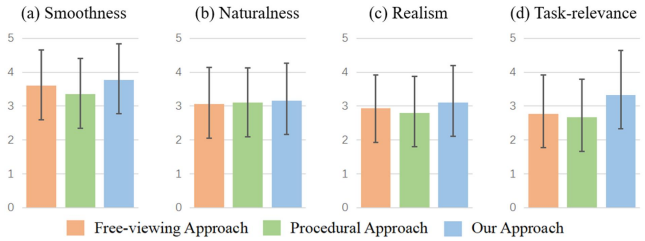

Figure 8: **Comparison between our proposed approach and other baselines based on human evaluation.** We show smoothness, naturalness, realism, and task-relevance evaluation scores for the head-and-eye motion generated by the free-viewing, procedural, and our approach. Error bars denote the standard errors of the mean.

order of the three approaches is randomly shuffled for each group. After each clip group, participants are instructed to complete a five-point Likert scale, where a rating of 1 means "strongly disagree" and 5 means "strongly agree." Participants are prompted to rate the animated clips based on the following four questions:

1. *Smoothness.* Is the eye gaze movement smooth?
2. *Naturalness.* Is the eye gaze movement natural?
3. *Realism.* Is the overall effect of virtual avatar realistic?
4. *Task-Relevance.* Assuming the avatar is tasked with patrolling the virtual environment, is its attention demonstrably task-related?

During the experiment, the participants are not explicitly informed which approach we used to generate the avatar's animation.

6.2 Statistical Analysis

We compare the results generated by the three approaches on four metrics. The mean and standard deviation of the users' ratings are shown in Figure 8. We also conducted a repeated-measures ANOVA test on the participants' ratings in each matric individually to analyze the significance. Please refer to the supplementary materials for the detailed number of statistics test results.

Smoothness evaluation The purpose of this evaluation is to verify the smoothness of avatars' eye gaze movement. Our approach obtains the highest rating ($M = 3.78$, $SD = 1.05$), followed by the free-viewing approach ($M = 3.60, SD = 1.05$) and procedural approach ($M = 3.35$, $SD = 1.05$). The ANOVA test results indicate that there is no significant difference in smoothness scores between our approach and the free-viewing approach ($p > .05$), as well as between the procedural approach and the free-viewing approach ($p > .05$). However, the smoothness score is significantly higher at our approach than at the procedural approach ($p < .05$).

Our approach considers unexpected stimuli such as sudden sounds or objects appearing in the user's peripheral vision, causing the avatar to turn its head towards them quickly. However, the free-viewing approach, which is characterized by randomness in navigation, is not statistically distinguishable from our approach in terms of smoothness. This suggests that the designed smooth processing and the distance loss $\mathscr{L}_{dis}$ of our approach solves this problem successfully and mimics a smooth eye gaze movement similar to when an avatar engages in random navigation. This result also indicates that, despite considering multi-level elements in the scene, our approach generates more purposeful and rational behavior, making its performance relatively consistent compared to the free-viewing approach.

Additionally, our approach shows a significant difference in eye gaze movement smoothness compared to the procedural approach. The procedural approach rigidly follows a set procedure. If the predefined directions are not smooth enough, it will cause unsmooth eye movements. Repetitive motions also result in relatively uniform and rigid eye movements. In contrast, our approach allows the avatar to respond flexibly to objects that have sudden appearances, and the synthesized head-and-eye motion remains smooth. It is noteworthy that our approach achieves flexible attention adjustment in dynamic scenes when introducing multi-level scene segmentation features.

Naturalness evaluation In the naturalness evaluation, our approach achieves the highest score ($M = 3.16$, $SD = 1.10$), closely following the results composed by the procedural approach ($M = 3.10$, $SD = 1.03$) and the free-viewing Approach ($M = 3.06$, $SD = 1.08$). The results of the ANOVA test suggest that there are no statistically significant differences among the three approaches for the naturalness of the avatar's eye gaze movement.

The procedural and free-viewing methods emulate specific gaze behaviors, while our approach forms a connection between gaze behavior and the scene. This allows gaze behavior to be influenced by various scene factors, potentially leading to more significant gaze disturbances compared to the other two techniques. Despite this, all three methods are perceived to have similar levels of naturalness, showcasing the effectiveness of our approach.

Various factors, such as scene semantics, behavioral habits, and task contents, can influence the naturalness of eye movements. While different methods may exhibit performance variations in certain aspects, they can still achieve a comparable level of naturalness in eye movements. All three approaches appear to be proficient at simulating natural eye gaze movement, making it challenging for participants to discern which is more natural.

Realism evaluation The realism evaluation aims to validate the impact of the generated eye gaze and head movements on the overall realism of the virtual avatar. A higher score implies that the overall realism of the virtual avatar is enhanced. Our approach attains the highest score ($M = 3.11$, $SD = 1.09$), followed by the free-viewing approach ($M = 2.93$, $SD = 0.99$) and the procedural approach ($M = 2.8$, $SD = 1.07$). After conducting the ANOVA test, the result is consistent with the smoothness evaluation, where our approach's realism score is significantly higher than the procedural approach ($p < .05$), with no significant difference compared to the free-viewing approaches ($p > .05$).

Our approach produces more realistic eye gaze and head movements compared to procedural approaches. However, there is no significant difference in naturalness for eye gaze movement alone. The reason may be that the procedural approach might provide a

sense of structure and predictability, which in some situations can be seen as "natural." When these cyclical movements are paired with coordinated head movements, the overall behavior might appear too mechanical or repetitive, diminishing the perceived realism. In real-life situations, our head movements are rarely rigid or strictly patterned and instead adapt and respond to our environment.

Our approach shows no significant difference when compared to the free-viewing approach. This could be because the free-viewing approach may be closer to the way people observe in daily life. However, our approach may offer a more structured and contextually relevant way, resulting in limitations in the freedom and spontaneity of the generated head and eye movements. Exploring and modeling the finer-grained relationship between eye gaze and head movements can further enhance performance.

Task-relevance evaluation The main goal of this evaluation is to test whether the avatar's gaze and head behaviors are related to their current tasks. Our approach achieves the highest score ($M = 3.33$, $SD = 1.32$), which is much higher than the free-viewing approach ($M = 2.78$, $SD = 1.14$) and the procedural approach ($M = 2.67$, $SD = 1.13$). The results of the ANOVA test indicate a significant increase in task-relevance score for our approach compared to both the free-viewing approach ($p < .05$) and the procedural approach ($p < .05$).

The results indicate that the head-and-eye motions synthesized by our approach are more task-related than the free-viewing and procedural approaches. Our approach can associate the eye gaze and head movements of the virtual avatar with the current task, emphasizing the effectiveness of these selected low- and high-level scene segmentation features. By integrating these features, our model grasps both coarse- and fine-grained features and obtains a deeper understanding of the simulated environment. The significant improvement results are meaningful, indicating that our model can generate eye gaze and head movements that focus on task-relevant key elements in the scene, stimulating human vision patterns and intentions, rather than just free-viewing or repetitive motions.

6.3 User Feedback

In our study, we gathered and examined user feedback, uncovering several intriguing insights. The majority of participants found the eye gaze and head movements of the virtual avatar, created using our method, to be interesting, appealing, and vivid. They suggested that incorporating this technique in game character design could notably enrich the gaming experience through authentic and contextually relevant eye movements. Traditional manual methods have struggled to replicate realistic gaze behaviors, often hindered by their complexity and inefficiency. Our approach effectively overcomes these challenges, offering designers a valuable tool to produce realistic eye movements, thereby significantly improving the overall user experience.

Several participants observed that the virtual avatar frequently focused on concealed areas like bushes, boxes, corners, or windows. For instance, one participant remarked, "The avatar doesn't just perceive a cube as a mere box, but recognizes it as potential cover." Another participant, reflecting on our approach, noted, "It's fascinating how the virtual avatar acts as if it were me, instinctively searching for hidden spots, almost as though it's predicting my actions," after experiencing the three different approaches. Overall, most participants could discern how the virtual character's perception of objects varied across the three methods. Interestingly, despite not explicitly feeding the model information about hidden areas, it learned to understand the semantics of these areas through multi-level segmentation features. This underscores the effectiveness of context-based features in our approach, revealing a significant insight into the avatar's behavior and the model's learning capabilities.

On the other hand, Some participants expressed interest in using non-human characters, like virtual cats, as avatars. This feedback has inspired us to broaden our approach to include a variety of characters, including fictional ones. We plan to enhance our system by incorporating a diverse range of head-and-eye motion generators, using expanded datasets of human interactions. This will allow us to train our generators for various tasks, making our virtual avatars more versatile and realistic.

7 Conclusions

In this paper, we introduce a novel computational approach to automatically generate natural and realistic head-and-eye motions of virtual characters based on contextual cues within the environment. Guide by the high- and low-level feature segmentations, our approach first simulates human visual attention mechanisms to generate the fixation sequence. Then, we leverage the fixation sequence as an intermediate representation to associate head-and-eye behavior with a real scene. To address the lack of datasets that contain context information and head-and-eye motion, we also introduce an annotated dataset, CAHE, which contains users' eye-tracking and head movement data as well as different levels of feature segmentations of virtual environments. Experiments show that our approach can synthesize realism and context-aware head-and-eye motions of virtual characters.

Our approach leads to a number of potential applications. For example, our approach enables game designers to rapidly generate realistic head-and-eye motions of game NPCs for a specific scene without time-consuming manual design. By creating the trajectory of the NPC and capturing the panoramic feature segmentation using the game engine, our approach automatically synthesizes the eye gaze and head movements of NPC. In addition, our approach enables virtual characters to respond to users' vocal cues or gestures in virtual conferences or educational platforms by utilizing the platform's camera and semantic segmentation technology. Moreover, in virtual museums, a virtual guide mainly focuses on important exhibits or areas along the way. Our approach generates corresponding head and eye movements based on the guide's position and pre-defined points of interest.

Limitations and future work Our current approach primarily relies on visual segmentation images as scene information. Real-world scenarios often involve multiple modalities, including auditory or text cues, which our model does not take into account. This might limit the richness and authenticity of the synthesized head-and-eye motions in more complex environments. Thus, our approach could incorporate multiple modalities to make the synthesized head-and-eye motions of the virtual avatar more authentic and aligned with real-world human behavior.

For our study, we only choose three commonly used virtual environments: forest, village, and town scenes. However, virtual scenes are complex and diverse in practice. We notice that town scenes with new objects have only a slight chance of being accurately predicted by our approach. Thus, our approach might not generalize well for every potential virtual environment, especially those that introduce novel objects and interactions. It is beneficial to expand our dataset and approach to cater to a broader range of virtual environments.

Our current implementation is centered around a single task – search. However, in real life, users engage in various tasks, such as interacting with virtual objects or characters. Our model does not capture the nuances of these diverse interactions. Integrating datasets and interaction models can significantly enhance synthetic behavior's diversity and authenticity.

Acknowledgments

The authors wish to thank students and teachers from Visual Perception & Human-computer Interaction Lab, Beijing Institute of Technology, for their great support in the experiments conducted in this work. This work is supported in part by the National Key R&D Program of China (2022ZD0114900) and the NSFC (62172043).

REFERENCES

[1] J. R. Anderson and J. Crawford. *Cognitive psychology and its implications*. Wh Freeman San Francisco, 1980. 2

[2] M. Argyle, M. Cook, and D. Cramer. Gaze and mutual gaze. *The British Journal of Psychiatry*, 165(6):848–850, 1994. 1

[3] T. Bondi, W. R. Games, and W. P. C. Poland. La noire. *Rockstar Games*, 2011. 1

[4] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li. Salient object detection: A survey. *Computational Visual Media*, 5:117–150, 2019. 2

[5] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 387–404, 2020. 2

[6] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 2

[7] M. Elbattah, C. Loughnane, J.-L. Guérin, R. Carette, F. Cilia, and G. Dequen. Variational autoencoder for image-based augmentation of eye-tracking data. *Journal of Imaging*, 7(5):83, 2021. 2

[8] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pp. 363–370, 2003. 5

[9] F. Games. Civilization vi. *Game [PC]. 2K Games*, 2016. 1

[10] R. Games. *Red dead redemption*. Rockstar Games, 2010. 7

[11] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. J. Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11374–11384, 2021. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 2

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2, 3

[14] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[15] D. Holden, J. Saito, and T. Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[16] S. Hou, Y. Wang, B. Ning, and W. Liang. Climaxing vr character with scene-aware aesthetic dress synthesis. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 57–64, 2021. 1

[17] Z. Hu, A. Bulling, S. Li, and G. Wang. Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2021. 1

[18] Z. Hu, A. Bulling, S. Li, and G. Wang. Fixationnet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 27(5):2681–2690, 2021. 2, 6

[19] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 26(5):1902–1911, 2020. 6

[20] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. Sgaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 25(5):2002–2010, 2019. 6

[21] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16750–16761, 2023. 2

[22] A. Klein, Z. Yumak, A. Beij, and A. F. van der Stappen. Data-driven gaze animation using recurrent neural networks. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11, 2019. 1, 2

[23] G. Lan, T. Scargill, and M. Gorlatova. Eyesyn: Psychology-inspired eye movement synthesis for gaze-based activity recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 233–246, 2022. 2

[24] Y. Lang, W. Liang, and L.-F. Yu. Virtual agent positioning driven by scene semantics in mixed reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 767–775, 2019. 2

[25] B. H. Le, X. Ma, and Z. Deng. Live speech driven head-and-eye motion generators. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 18(11):1902–1914, 2012. 2

[26] S. P. Lee, J. B. Badler, and N. I. Badler. Eyes alive. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 637–644, 2002. 1, 2

[27] C. Li, W. Liang, C. Quigley, Y. Zhao, and L.-F. Yu. Earthquake safety training through virtual drills. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 23(4):1275–1284, 2017. 2

[28] W. Liang, L. Wang, X. Yu, C. Li, R. Alghofaili, Y. Lang, and L.-F. Yu. Optimizing product placement for virtual stores. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 336–346, 2023. 2

[29] W. Liang, X. Yu, R. Alghofaili, Y. Lang, and L.-F. Yu. Scene-aware behavior synthesis for virtual pets in mixed reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021. 2

[30] X. Ma and Z. Deng. Natural eye motion synthesis by modeling gaze-head coupling. In *2009 IEEE Virtual Reality Conference*, pp. 143–150, 2009. 2

[31] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the Tenth ACM International Conference on Multimedia*, pp. 533–542, 2002. 2

[32] A. Mohanty and T. J. Sussman. Top-down modulation of attention by emotion, 2013. 2

[33] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics and Automation Magazine*, 19(2):98–100, 2012. 1

[34] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[35] M. Nixon, S. DiPaola, and U. Bernardet. An eye gaze model for controlling the display of social status in believable virtual humans. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8, 2018. 1

[36] N. Norouzi, K. Kim, J. Hochreiter, M. Lee, S. Daher, G. Bruder, and G. Welch. A systematic survey of 15 years of user studies published in the intelligent virtual agents conference. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA)*, pp. 17–22, 2018. 1

[37] J. L. Orquin and S. M. Loose. Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, 144(1):190–206, 2013. 2

[38] D. J. Parkhurst and E. Niebur. Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, 19(3):783–789, 2004. 2

[39] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10985–10995, 2021. 3

[40] R. Pieters and M. Wedel. Goal control of attention to advertising: The yarbus implication. *Journal of Consumer Research*, 34(2):224–233, 2007. 2

[41] M. I. Posner, Y. Cohen, et al. Components of visual orienting. *Attention and Performance X: Control of Language Processes*, 32:531–556, 1984. 3

[42] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, and N. Avouris. Using eye gaze data and visual activities to infer human cognitive styles: method and feasibility studies. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP)*, pp. 164–173, 2017. 1

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-*

nition (CVPR), pp. 10684–10695, 2022. 2

[44] K. Ruhland, S. Andrist, J. Badler, C. Peters, N. Badler, M. Gleicher, B. Mutlu, and R. Mcdonnell. Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In *Eurographics 2014-State of the Art Reports*, pp. 69–91, 2014. 1

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2

[46] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 24(4):1633–1642, 2018. 2

[47] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, 2015. 2

[48] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[49] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3

[50] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[51] J. Wang, S. Yan, B. Dai, and D. Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12206–12215, 2021. 2

[52] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14959–14971, 2022. 2

[53] L. Wheeler. Color and human responses: Aspects of light and color bearing on the reactions of living things and the welfare of human beings by faber birren. *Leonardo*, 13(4):334–335, 1980. 2

[54] M. Xu, G. Jiang, W. Liang, C. Zhang, and Y. Zhu. Active reasoning in an open-world environment. *arXiv preprint arXiv:2311.02018*, 2023. 2

[55] B. Yang and H. Li. A visual attention model based on eye tracking in 3d scene maps. *ISPRS International Journal of Geo-Information*, 10(10):664, 2021. 2

[56] S. H. Yeo, M. Lesmana, D. R. Neog, and D. K. Pai. Eyecatch: Simulating visuomotor coordination for object interception. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 2

[57] R. Zemblys, D. C. Niehorster, and K. Holmqvist. gazenet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, 51:840–864, 2019. 2

[58] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 2