Supplementary for "EyEar: Learning Audio Synchronized Human Gaze Trajectory based on Physics-informed Dynamics"

A Dataset Collection

This section introduces the detailed process for collecting eye movement data, determining the number of subjects and data preprocessing. We also conduct word frequency analysis in narratives of our dataset to investigate word biases in the gaze trajectory data.

A.1 Eye Movement Data Collection.

Each image is displayed on the screen with a projected size of 0.2 meter by 0.2 meter, and the distance between the eyes and the screen is 0.5 meter. The prepared narrative text is converted to audio and then played to the subjects. A 5-second blank audio is added at the beginning, allowing the subjects to get familiar with the image. To maintain consistency, audio conversions are conducted using the same timbre from a TTS tool[1]. We employ an "Eyelink Portable" eye-tracking device to instantly capture the coordinates of where subjects' eyes land on the screen when they view the images while simultaneously listening to the audio.

A.2 Subject Number Selection.

To avoid randomness, it is necessary to collect eye movement data from multiple subjects for each image-text pair. Determining the number of subjects is crucial because too few subjects may result in poor data stability, while too many subjects are costly. We determine the optimal number of subjects based on the stability of distribution via a pilot trail on a set of 75 image-audio pairs before formally collecting data. For a given number of subjects, we can estimate the distribution formed by the recorded gaze points. We increase the number of subjects and calculate the changes between the adjacent distributions by KL divergence. As shown in Figure 6, we find that there is a turning point where the KL divergence goes flat and low after the number of subjects reaches 8. Therefore, we finally choose 8 subjects in collecting our dataset.

Demographic information. To mitigate selection bias and enhance the applicability of our dataset, we prioritize diversity during data collection by selecting subjects as broadly and randomly as possible. Our pilot trial involves 12 randomly recruited subjects, evenly split between male and female, with ages ranging from 18 to 28. The subjects have diverse academic backgrounds including 9 distinct fields, which ensures a broad demographic representation.

[1]https://ttsmaker.com/zh-cn

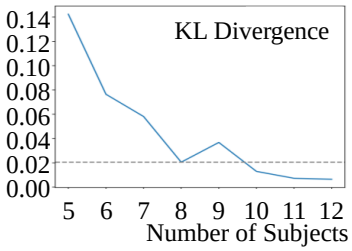

Figure 6: The changes in the distributions as the number of subjects increases.

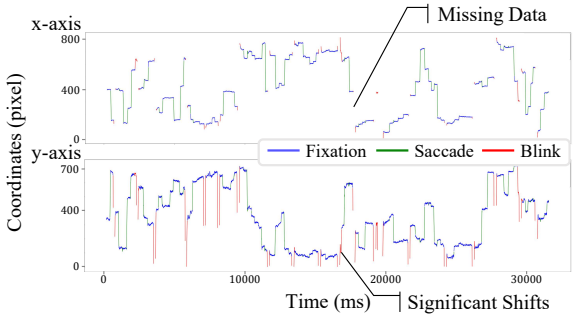

Figure 7: Original eye movement data. There are some data losses and significant shifts caused by blinks.

A.3 Data Preprocessing.

The original eye movement data exhibits some data losses and significant shifts, as shown in Figure 7. Blinking, a necessary behavior that replenishes tears in the cornea of eyes, is a primary factor causing these data losses and often coincides with two significant shifts due to closing and reopening the eyes. To obtain higher-quality data, it is necessary to identify and remove the blink segments. First, we calculate the increments of the eye movement in the X and Y directions for each sampling interval, e.g., 1ms in our dataset. When the absolute increment values exceed a certain threshold, e.g., 4 pixels in our experiment, the segment is considered non-fixation points. Second, missing values are labeled

Table 4: Distributions for the top 20 most frequent words and the top 10 most frequent content words in narratives of our dataset EyEar-20k.

Word	Frequency	Word	Frequency
Top 20 words			
of	263	have	90
a	49	at	44
on	32	-ing	28
is	27	being	26
beside	26	white	26
one	26	put	25
and	23	some	22
she	18	also	18
girl	18	far away	18
red	17	right	16
Top 10 content words			
white	26	she	18
girl	18	red	17
woman	14	house	14
person	13	he	11
man	11	clothes	10

as blink points, and all non-fixation points adjacent to blink points are considered blink behaviors. At last, by excluding all blink behaviors and utilizing linear interpolation to fill in the missing data gaps, we obtain higher-quality eye movement data.

A.4 Word frequency analysis.

We provide distributions for the top 20 most frequent words in narratives of our dataset EyEar-20k. As shown in Table 4, function words make up a large portion, echoing the challenge emphasized in the Related Work section. We also report the top 10 most frequent content words in Table 4. It reflects our preference for daily life scenes when selecting images, which aligns with our motivation of building alive virtual characters in realistic settings. The similar frequencies of these content words also highlight the diversity of image selection.

B Implementation Details

B.1 Network Structure.

For the coarse-grained background information, we segment the image into 4×4 patches. Text and image embeddings are both obtained through the pre-trained BriVL model (Huo et al. 2021), with dimensions of 2048. The number of layers for the GRU is set to 3. In the attention mechanism, the size of the matrix W_q is set to 2048×32, compressing the text embeddings to 32 dimensions to avoid consuming excessive computational resources. Similarly, the size of W_k is consistent with W_q to maintain uniformity between text and image. The size of the matrix W_v is set to 2×2. The feedforward neural network is a two-layer MLP network with a ReLU activation function, and the hidden layer has a dimension of 32. In the audio-aware dynamical system, neural networks are all set to two-layer MLP networks with a ReLU activation function, and the hidden layer has a dimension of 16.

B.2 Training Settings.

We find that it is slow to optimize the probability density loss, although it is more precise. Therefore, we adopt a two-stage training process. In the first stage of training, we use MSE loss to optimize the model. After reaching a plateau in the loss, we continue training using the probability density loss function to refine the model. This two-stage training strategy strikes a balance between training efficiency and effectiveness. For the probability density loss function, we use the default setting of kernel parameters in the "scipy" library. All components of the model are jointly trained with the learning rate of 1e-4 and the optimizer is set to AdamW. Additionally, teacher forcing is utilized during training for sequence prediction. The initial teacher forcing ratio is set to 0.5 and gradually reduced to 0. It ensures that there are no discrepancies between the training and testing stages. All the above hyperparameters are fine-tuned based on the validation set.

C Baselines Construction

In this section, we provide a detailed explanation of how each baseline is constructed.

(1) **Pre-trained image-text models.** We choose BriVL (Huo et al. 2021) and CLIP (Radford et al. 2021), two popular pre-trained image-text models, as baselines. CLIP uses a text encoder and an image encoder to extract features from text and images, and employs contrastive learning to learn image-text alignment. BriVL is a variant of CLIP specifically designed for Chinese text. We employ these two models for encoding patches and words, and take the center coordinates of the patch with the highest cosine similarity to the current word as the predicted gaze point.

(2) **Visual grounding models.** We choose Grounding DINO (Liu et al. 2023) and MITR (Meng et al. 2021) as baselines. Grounding DINO is the state-of-the-art visual grounding model that marries Transformer-based detector DINO (Zhang et al. 2022) with grounded pre-training. MITR is a mirrored transformer architecture trained on Localized Narratives (LN) dataset (Pont-Tuset et al. 2020). The LN dataset contains sentence-level text and bounding boxes derived from mouse traces. Therefore, the setup of MITR is more similar to our task, and it is expected to achieve better performance. We use these two models to predict bounding boxes for all words and connect their centers as the predicted gaze trajectory.

(3) **Gaze trajectory prediction models.** In the visual searching task, subjects are instructed to search for a given visual target within an image. We choose the latest models Chen et al for VS (Chen, Jiang, and Zhao 2021) and Gazeformer (Mondal et al. 2023) as baselines. We construct the baselines by regarding each word as a search target and concatenating the results of all words to form the gaze trajectory. In the visual question answering task, subjects need to view the image based on the given question. We choose Chen et al for VQA (Chen, Jiang, and Zhao 2021) as the baseline,

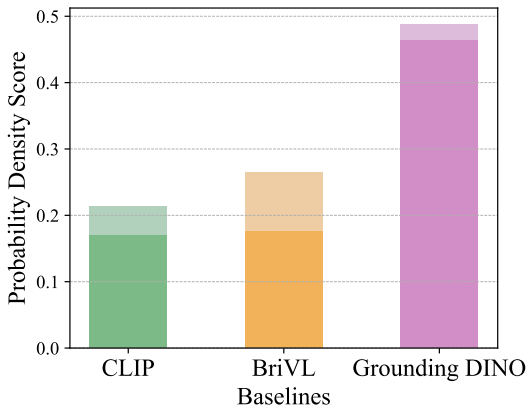

Figure 8: Baselines performance comparisons between incorporating the dynamical system and without. The light-colored bars represent models with the dynamical system. Incorporating the dynamical system improves the performance of all baseline models.

Table 5: Image information ablation on EyEar. The best results are in **bold**.

Model	PDS ↑	ED ↓	DTW ↓
EyEar	**0.614**	**221.6**	**201.3**
w/o patchify	0.544	242.6	212.3
w/o visual grounding	0.526	239.5	219.3

and construct the baseline by regarding the narrations in our task as the questions.

D Additional Results and Discussions

D.1 The plug-and-play capability of the dynamical system.

Our proposed audio-aware dynamical system can effectively learn the motion characteristics of objects by simply being provided the possible forces acting on them. Therefore, it is expected to have the plug-and-play capability. To validate this capability, we further introduce our dynamical system into the baselines (CLIP, BriVL, and Grounding DINO). We consider the output of these baselines as the semantic attraction force, as they all only consider the connection between text and images. As shown in Figure 8, incorporating the dynamical system improves the performance of all baseline models, indicating that our proposed dynamical system is universally applicable and plug-and-play. One only needs to comprehensively consider the possible forces and provide them accurately, and the dynamical system can help to precisely learn the characteristics of motion.

D.2 Additional Ablation Study.

In this section, we provide further ablation on EyEar, investigating the impact of different image information, different saliency models and different distribution-based loss.

Table 6: The impact of different saliency models on the performance of EyEar. The best results are in **bold**.

Used Saliency Models	PDS ↑
DeepGaze IIE	**0.614**
SalFBNet	**0.614**
Unisal	0.612
None	0.602

Table 7: The impact of different distribution-based loss. The best results are in **bold**.

Loss	ED ↓	DTW ↓	SMatch ↑
Gaussian KDE	**221.6**	**201.3**	**0.464**
Epanechnikov KDE	227.5	207.0	0.459
Exponential KDE	228.1	207.0	0.455
Parameter Estimation	227.9	226.8	0.415
MSE	237.3	237.3	0.397

Image information ablation. To demonstrate the importance of both coarse-grained (patchify) and fine-grained (visual grounding) information in the image branch, we perform the ablation studies by removing each and observing the impact on the performance. As shown in Table 5, removing either type of the image information results in a comparable performance drop, indicating that both are equally important for our task. Fine-grained visual grounding information provides detailed image context in response to auditory stimuli, while coarse-grained background information offers rich background context for the entire image, both of which can influence human gaze.

The impact of different saliency models. We introduce different saliency models into our framework to understand how different saliency models affect the outcome. We select the top 3 models from the famous MIT/Tuebingen Saliency Benchmark: DeepGaze IIE (Linardos et al. 2021), SalFBNet (Ding et al. 2022), and Unisal (Droste, Jiao, and Noble 2020). Their sAUC scores are 0.794, 0.786, and 0.784, respectively. As shown in Table 6, using different saliency models consistently improves the performance. The performance of EyEar also shows a positive correlation with that of the saliency models, with a Pearson correlation coefficient of 0.636, which further supports the positive impact of saliency prediction. However, since saliency prediction is a minor component of our model, the choice of different saliency models has a relatively small impact.

The impact of different distribution-based loss. To further demonstrate the effectiveness of our proposed PD loss, we explore three alternative methods: parameter estimation (assuming the data follows a Gaussian distribution), Exponential KDE and Epanechnikov KDE to observe the impact of different distribution-based loss. As shown in Table 7, probability-based loss shows clear superiority over MSE loss. Non-parametric estimation outperforms parameter estimation, and Gaussian KDE (our choice) achieves the

best results. Our PD loss utilizes widely used Gaussian KDE to estimate the distribution, which is highly suitable for our data with high individual variability.

E Practical applications

One of the most direct applications of our task is in virtual characters. Our model can enable virtual characters to move their eyes in a human-like manner when receiving both auditory and visual stimuli. To demonstrate this, we apply our model, EyEar, within a virtual character engine. A demo of the virtual character can be found at https://github.com/XiaochuanLiu-ruc/EyEar. By applying our model, virtual characters can exhibit more flexible and natural eye movements, representing an advancement in making virtual characters come to life.