EyEar: Learning Audio Synchronized Human Gaze Trajectory Based on Physics-Informed Dynamics

Xiaochuan Liu[*1], **Xin Cheng**[*1], **Yuchong Sun**[1], **Xiaoxue Wu**[1], **Ruihua Song**[1†], **Hao Sun**[1†], **Denghao Zhang**[2†]

[1]Gaoling School of Artifcial Intelligence, Renmin University of China, Beijing, China
[2]Department of Psychology, Renmin University of China, Beijing, China
{liuxiaochuan, chengxin000, ycsun, wuxiaoxue1102, rsong, haosun, zdh}@ruc.edu.cn

Abstract

Imitating how humans move their gaze in a visual scene is a vital research problem for both visual understanding and psychology, kindling crucial applications such as building alive virtual characters. Previous studies aim to predict gaze trajectories when humans are free-viewing an image, searching for required targets, or looking for clues to answer questions in an image. While these tasks focus on visual-centric scenarios, humans move their gaze also along with audio signal inputs in more common scenarios. To fill this gap, we introduce a new task that predicts human gaze trajectories in a visual scene with synchronized audio inputs and provide a new dataset containing 20k gaze points from 8 subjects. To effectively integrate audio information and simulate the dynamic process of human gaze motion, we propose a novel learning framework called EyEar (Eye moving while Ear listening) based on physics-informed dynamics, which considers three key factors to predict gazes: eye inherent motion tendency, vision salient attraction, and audio semantic attraction. We also propose a probability density score to overcome the high individual variability of gaze trajectories, thereby improving the stabilization of optimization and the reliability of the evaluation. Experimental results show that EyEar outperforms all the baselines in the context of all evaluation metrics, thanks to the proposed components in the learning model.

1 Introduction

Predicting human gaze trajectory within a visual scene is an important research problem for a range of research communities, including visual understanding, human behavior, psychology, etc. It is also crucial for some downstream applications, such as building virtual characters that are alive and more interactive. As shown in Figure 1, existing studies of gaze trajectory prediction aim to predict gaze sequences (a.k.a. scanpaths) when human are free-viewing images (Jiang et al. 2015) or webpages (Shen and Zhao 2014), searching for required objects in an image (Mondal et al. 2023; Yang et al. 2020), or collecting clues for answering visual questions (Chen, Jiang, and Zhao 2021).

While existing tasks focus on visual-centric scenarios, the gaze trajectory of humans can be affected by both visual and audio stimuli. To bridge this gap, we propose a new task: predicting a gaze trajectory in an audio synchronized viewing scene. We contribute to this task from both dataset and algorithm perspectives.

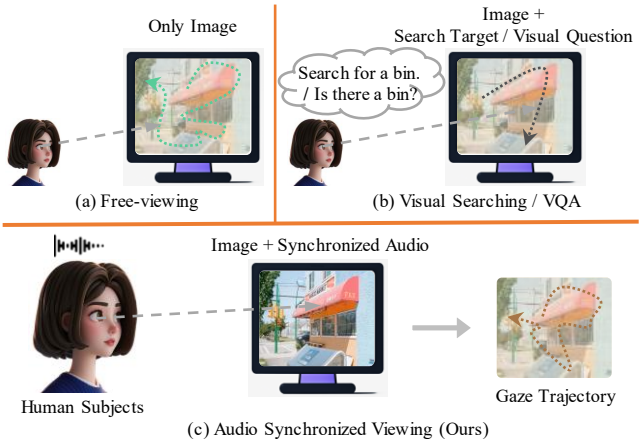


Figure 1: Different from existing tasks which predict gazes in situations such as (a) humans free-viewing an image, or (b) searching for required targets or clues to answer questions in an image, our proposed task (c) aims to predict human gaze in a more common scenario where humans receive synchronized audio signals when directing their gaze.

From the dataset perspective, we construct a new dataset with 20k gaze points from 8 subjects, using eye tracking devices to trace the subjects' scanpaths over the image when they hear an audio clip. Our dataset has a much longer average gaze sequence length and duration than existing datasets, making it more challenging and valuable for learning long-range human gaze trajectories.

From the algorithm perspective, we solve two major challenges. The first challenge is how to effectively integrate audio information in this multimodal scenario, with a proper framework that considers the dynamical physics of eye movement (*C1*). The second challenge is that different people have diverse gaze trajectories, making the collected data have high individual variability (*C2*). This makes the commonly used Mean Square Error (MSE) loss confused and raises a challenge for the stability of optimization. Existing evaluation metrics based on Euclidean distance also lack re-

[*]These authors contributed equally.
[†]Corresponding authors.

liability due to divergent targets.

We overcome the aforementioned challenges by proposing an EyEar framework with two novel designs. First, we build an audio-aware dynamical system (*C1*), inspired by the dynamical systems in physics which can model the complex evolution process using states and motion vectors (Birkhoff 1927). Specifically, we consider three kinds of forces to decide motion: an inherent motion tendency force to model the continuous transitions of eye movement, an audio semantic attraction force that models fine-grained audio-visual association to predict an attraction point in the image due to the presence of audio, and a vision salient attraction force that models attractions to salient image regions. Second, we propose a probability density score (PDS) for reliable evaluation and a corresponding probability density loss to ensure stable optimization (*C2*). PDS is based on mixed Gaussian distributions fitted from the gaze points of multiple subjects. It estimates the distribution using Gaussian kernel density estimation and measures how well the predicted gaze point fits the ground-truth distribution by the normalized value of its probability density.

Experiment results show that EyEar performs the best in all evaluation metrics. The improvements over the best in each metric are from 4% to 15%. Ablation studies show that our proposed components are all effective for improving performance. Our contributions can be summarized as follows:

- We propose a new task that predicts gaze trajectories in a visual scene under the stimulation of synchronized audio inputs. We also collect a dataset containing 20k gaze points from 8 subjects to facilitate the investigation.
- We propose an EyEar framework that models physics-informed dynamics by considering three possible forces that influence motion to effectively integrate audio information by fine-grained audio-vision association modeling and simulate the dynamics of human gaze.
- We propose a probability density score and loss based on mixed Gaussian distributions of multiple gaze trajectories, which not only stabilize model optimization but also evaluate different methods more reliably.

2 Related Work

Gaze Trajectory Prediction. Studies in the neuroscience field (Moschovakis, Gregoriou, and Savaki 2001; Krauzlis, Lovejoy, and Zénon 2013) show that when staring at a certain position on an image, human brain selects the next point to look at and then moves to it. To mimic this mechanism and generate human-like gaze sequences, researchers propose the gaze trajectory (a.k.a scanpath) prediction tasks. Different gaze trajectory prediction tasks focus on different specific tasks such as free-viewing (Judd, Durand, and Torralba 2012; Borji and Itti 2015; Jiang et al. 2015), viewing webpages (Shen and Zhao 2014), visual searching (Yang et al. 2020; Mondal et al. 2023), and visual question answering (Chen, Jiang, and Zhao 2021). Kümmerer and Bethge (2021) survey the models that predict gaze trajectories and classify previous models into four categories: biologically inspired models (Engbert et al. 2015; Zanca, Melacci, and Gori 2020), statistically inspired models (Xia et al. 2019; Lan, Scargill, and Gorlatova 2022), cognitively inspired models (Liu et al. 2013; Sun, Chen, and Wu 2021) and engineered models (Assens et al. 2017; Kümmerer, Bethge, and Wallis 2022). However, most existing works focus on silent tasks with well-defined requirements. Instead our work extends the gaze trajectory prediction task to a visual scene with synchronized audio inputs, and proposes a physics-informed model to learn a natural gaze trajectory.

Visual Grounding. Our task is also related to visual grounding task because people are expected to gaze at the semantically relevant regions of images when hearing narrations. Visual grounding task aims at locating objects queried by natural language in an image. It is a multi-modal task that requires understanding of the relationship between language and vision. Many advanced models like MDETR (Kamath et al. 2021), GLIP (Li et al. 2022), DQ-DETR (Liu et al. 2022), and Grounding DINO (Liu et al. 2023b) utilize the power of object detection techniques and pre-trained multi-modal models to achieve good performance. MITR (Meng et al. 2021) is a more relevant work to our task in terms of data sources. It is trained on Localized Narratives (LN) dataset (Pont-Tuset et al. 2020) with bounding boxes derived from mouse traces while annotators describing images. However, there exists substantial divergence between eye tracking data and mouse tracking data (Tavakoli et al. 2017). Despite being similar to visual grounding tasks, our new task has more challenges involving the large amount of ungroundable words, the unique motion process of gazes, and the high individual variability of gazes from different persons, which will be addressed in this paper.

3 Task and Dataset

3.1 Problem Formulation

The input to our task consists of an image V and an audio clip A. Using speech recognition tools, we can obtain a sequence of words with their start and end time from the audio:

$$(w_1, t_1^s, t_1^e), \cdots, (w_i, t_i^s, t_i^e), \cdots, (w_n, t_n^s, t_n^e), \quad (1)$$

where w_i is the i-th word, t_i^s and t_i^e are the start and the end timestamps of word w_i in audio A respectively, n is the number of words in the audio.

In our task, human gaze well aligns with the audio timestamps. Subjects tend to gaze at a point until they hear the next word. Therefore, we predict gaze points at each end time. At the end time t_i^e, ground-truth gaze points from N subjects are recorded. Consequently, for the image-audio pair V-A, the recorded gaze trajectory of the j_{th} subject is:

$$S_j = \{s_1^j, \cdots, s_n^j\}, \quad (j = 1, \cdots, N), \quad (2)$$

where s denotes a coordinate on the image. The objective of our task is to predict a gaze trajectory $\hat{S} = \{\hat{s}_1, \cdots, \hat{s}_n\}$ that is most similar to those ground-truth gaze trajectories conditioned on image and audio inputs.

3.2 Dataset Construction

To facilitate our work and future research, we collect gaze tajectory data from eight subjects when they view images and simultaneously listen to corresponding audio clips. Our

Dataset	Task Scenario	Synchronized Stimuli	Avg. Sequence Length	Avg. Duration
OSIE (Xu et al. 2014)	Free-viewing	✗	9.36	2.01s
COCO-FreeView (Chen et al. 2022)	Free-viewing	✗	15.45	4.05s
COCO-Search18 (Chen et al. 2021)	Visual Searching	✗	3.77	0.92s
AiR (Chen et al. 2020)	Visual Question Answering	✗	10.16	2.89s
EyEar-20k (Ours)	Audio Synchronized Viewing	✓	**26.06**	**12.09s**

Table 1: Comparing our dataset with commonly used gaze trajectory prediction datasets. Different from free-viewing or visual searching, our task predicts gaze trajectory in a visual scene with synchronized audio stimuli. Our dataset has much longer sequence length and duration than existing datasets, enabling learning long-range human gaze dynamics from our dataset.

dataset contains a total of 20k gaze points. The detailed steps for constructing the dataset are as follows. (1) **Image Selection.** We select images from Unsplash.com website, known for providing high-quality, visually appealing images that are freely available to the public. To mirror the real world more accurately, we select images that one might encounter in everyday life. We also require the selected images containing multiple objects and rich details, while avoiding close-up shots. All images have a resolution of 1024*1024 pixels. (2) **Narrative Audio Designing.** We design the narrative text using a combination of automatic generation and manual refinement. In our task, text primarily serves to guide the subjects' attention to different areas of the images. Thus it would provide accurate and detailed descriptions of the images. For efficiency, we choose LLaVA-1.5 (Liu et al. 2023a) to automatically generate descriptions of the images. To ensure data quality, annotators are asked to manually modify the generated descriptions to avoid rigid patterns, such as always starting with "there is" or "in the image", increase the diversity of description expressions, correct mistakes, and supplement descriptions of objects that have not been described. The prepared narrative text is converted to audio with a TTS tool 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. Please note that all text and audio data in our dataset are in Chinese because all subjects participating in the data collection are native Chinese speakers. (3) **Subjects Instructions.** It is important to note that the subjects are not given explicit instructions. They are instructed to maintain the most natural state because we aim to predict gaze trajectories that occur in daily life rather than in a specific task scenario. We introduce other details of our dataset collection in Supplementary A.

3.3 Dataset Comparison

We compare our dataset with commonly used gaze trajectory datasets, as shown in Table 1. OSIE and COCO-FreeView only record human gaze trajectories when free-viewing images without stimuli of other modalities. While COCO-Search18 and AiR datasets contain textual information, they present the text before displaying images, the gaze trajectories reflect the searching process over images. In contrast, our dataset presents subjects images with synchronized auditory stimuli, closely mirroring real-world scenarios. Furthermore, the statistics indicate that our dataset features much longer sequence length and duration, making it more effective for learning long-range human gaze trajectories.

4 Method

4.1 Framework Overview

As shown in Figure 2, our proposed framework EyEar is built on a physics-informed dynamical system that considers three kinds of forces to decide the motion of eyes (See Module 1): one force keeping the inherent motion tendency of eye, one force attracting eyes to salient part of image V no matter what the subject heard, and the other force attracting eyes to the audio semantic attraction point that is semantically relevant to the heard text. Specifically, we propose a multimodal attention mechanism to predict the audio semantic attraction point for each word (See Module 2). It can integrate three different types of information: image, text, and image patch coordinates into the coordinates of next audio semantic attraction point. Furthermore, EyEar introduces a probability density score function based on mixed Gaussian distributions fitted from multiple ground-truth gaze points as a loss function (See Module 3). It allows for better optimization and evaluation of a natural gaze trajectory, which has high individual variability across humans.

4.2 Audio-Aware Dynamical System

To capture the motion features of eyes, we propose an audio-aware dynamical system inspired by physics (Birkhoff 1927). In a dynamical system, there is a concept known as a state, which is a set of determinable real numbers. Tiny variations in the state correspond to tiny variations in these real numbers. The evolution of the dynamical system is governed by a set of functions, describing how future states depend on the current state. The rule is deterministic, which means, for a given time interval, only one future state can evolve from the current state. Herein, a state refers to the gaze location.

As shown in Module 1 of Figure 2, the formulation of our dynamical system is as follows:

$$\hat{s}_i = \hat{s}_{i-1} + \Delta t_i \cdot M_i, \tag{3}$$

where the current predicted gaze point $\hat{s}_i$ is calculated based on the previous gaze $\hat{s}_{i-1}$, the time interval $\Delta t_i = t_i^e - t_{i-1}^e$,

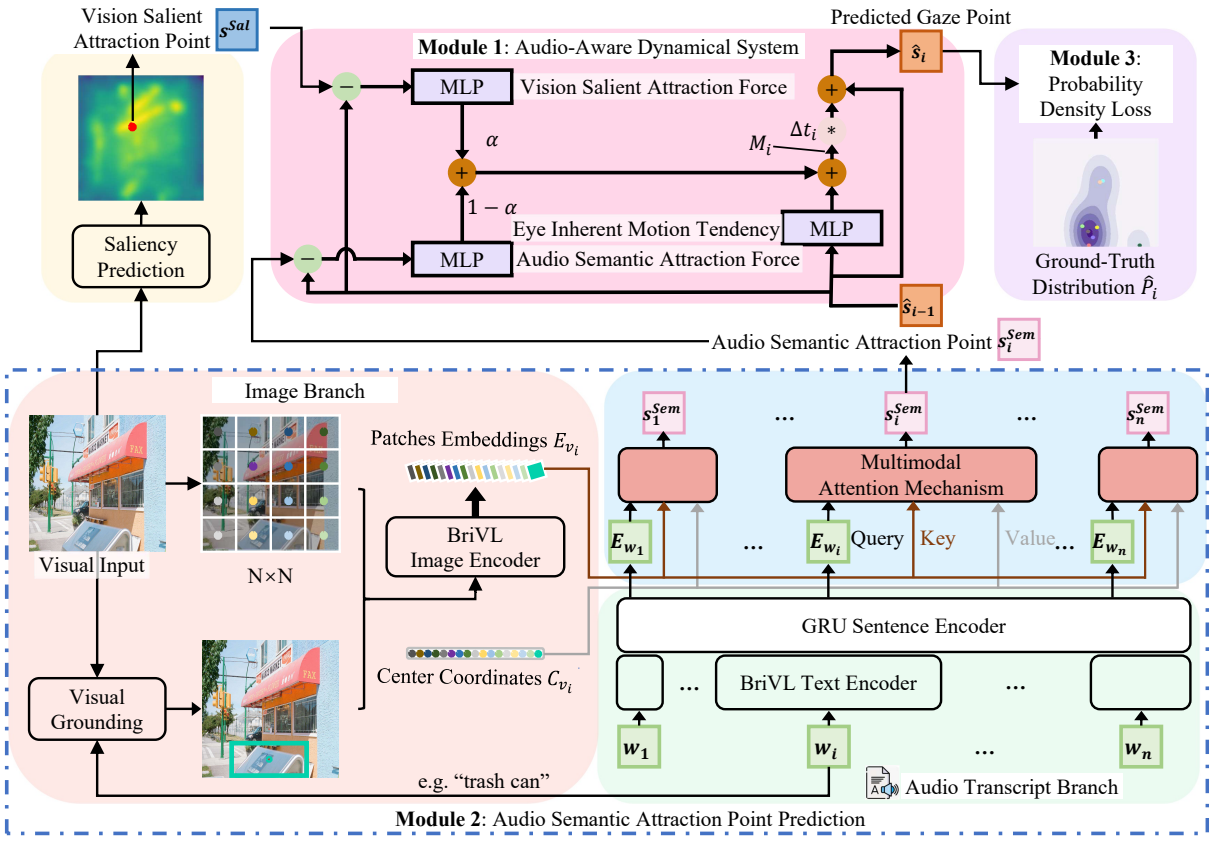


Figure 2: **EyEar overview.** The core component is a physics-informed audio-aware dynamical system that simulates the motion of eyes (See Module 1). The next predicted gaze point is calculated from the current gaze point, the time interval, and a motion vector. The motion vector is influenced by three kinds of forces. The most important force, audio semantic attraction force is predicted by Module 2. We propose probability density loss (See Module 3) to train the model.

and a motion vector M_i. Specifically, M_i is calculated as:

$$M_i = \mathit{MLP}_A(\hat{s}_{i-1}) + \alpha \cdot \mathit{MLP}_B(s^{Sal} - \hat{s}_{i-1}) + (1-\alpha) \cdot \mathit{MLP}_C(s_i^{Sem} - \hat{s}_{i-1}), \quad (4)$$

where s^{Sal} means vision salient attraction point and s_i^{Sem} means the i_{th} predicted audio semantic attraction point. We comprehensively consider three sources of force influencing the motion vector that represents both the direction and speed of motion. The neural networks in Eq.(4) correspond to the set of functions (motion components caused by three forces) in the dynamical system. The first term in Eq.(4) represents the motion component caused by the force that keeps the inherent motion tendency of eye at the current position $\hat{s}_{i-1}$, regardless of any stimuli. The second term in Eq.(4) represents the motion component caused by the force that attracts the gaze to the most salient point in the image. s^{Sal} is obtained by utilizing the DeepGaze IIE (Linardos et al. 2021) model. We consider this force because human attention may sometimes be completely captivated by the image, in particular the salient part. The third term in Eq.(4) represents the motion component caused by the force that attracts the gaze to the audio semantic attraction point s_i^{Sem}. This term considers human attention influenced by auditory stimuli. Intuitively, when a human hears some words, she/he pays attention to the semantically relevant part. Finally, the learnable weighting parameter α measures the extent to which human attention is drawn to the image itself rather than being influenced by the heard.

4.3 Audio Semantic Attraction Point Prediction

As shown in Module 2 of Figure 2, to measure the broad semantic relations between image regions and the heard words, we carefully design image branch, audio transcript branch, and a multimodal attention mechanism to integrate different types of information and predict the next audio semantic attraction point.

Image Branch. In our scenario, both coarse-grained background information and fine-grained visual grounding information in image V are found helpful in identifying audio semantic attraction points. For the background information, we segment the image into $N \times N$ patches. We choose $N = 4$ for its best performance in our experiments. We also record the center coordinates of these patches for subsequent use. For the fine-grained visual grounding information, we apply the state-of-the-art model Grounding DINO (Liu et al. 2023b) to identify the most relevant re-

gion in the image for each word w_i. This identified region is regarded as a special patch V_{D_i} and we record its center coordinates too. When no identified region output by Grounding DINO, we set the embedding of V_{D_i} as all zeros, and the center coordinates as $(0, 0)$. Next, we employ a large-scale pre-trained model to align image and text embedding in the same space. Given the Chinese nature of our narrations, we choose BriVL (Huo et al. 2021) (a Chinese CLIP-like model) for this purpose. All $N \times N$ patches plus the special patch identified by Grounding DINO are finally encoded into embeddings: $E_{v_i} = IE(V_{N\times N}, V_{D_i})$, where IE is the abbreviation of BriVL's Image Encoder.

Audio Transcript Branch. We first use the text encoder of BriVL to encode each word in the audio: $E'_{w_i} = TE(w_i)$, where TE is the abbreviation of Text Encoder. For a large number of ungroundable words, the correlation between their embeddings and the image embeddings is minimal. To achieve fine-grained alignment between audio and visual stimuli, we use the GRU for additional sentence encoding: $E_{w_i}, H_i = GRU(H_{i-1}, E'_{w_i})$, where H_i represents the hidden state of the GRU at the i-th time step and E_{w_i} represents the final embedding for word w_i. We choose the GRU instead of Transformer (Vaswani et al. 2017) for two reasons: i) we observe that the semantic attraction point is sensitive to the nearer heard words, which well aligns with the characteristics of RNN models (Zaremba, Sutskever, and Vinyals 2014). They can effectively utilize short-range context information; and ii) compared to the Transformer, GRU requires much fewer computational resources and training data.

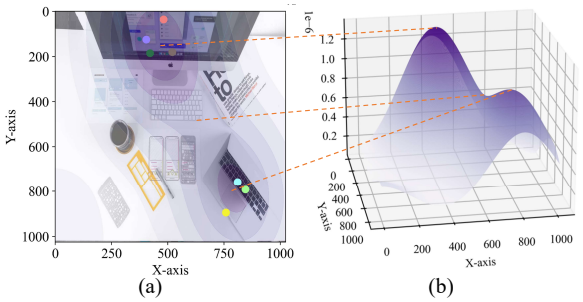
Prediction with Multimodal Attention Mechanism. Intuitively, when a person is viewing an image and listening to narrative audio at the same time, various relevant parts of the image exert attractive forces to her/him. The audio semantic attraction point should be a comprehensive manifestation of these attractive forces. Therefore, we design a multimodal attention mechanism to integrate information from image and text, and finally predict the audio semantic attraction point. Specifically, we treat text embedding E_{w_i} as query, each image embedding in E_{v_i} as key, and the center coordinates of each image patch as value:

$$s_i^{Sem} = Attention(E_{w_i}, E_{v_i}, C_{v_i}), \tag{5}$$

where C_{v_i} represents the center coordinates of patches, and s_i^{Sem} represents the audio semantic attraction point for word w_i. The network structure of our attention mechanism is the same as the Transformer block (Vaswani et al. 2017).

4.4 Probability Density Loss

The high individual variability of ground-truth gaze trajectories raises a challenge to optimization and evaluation. As the example in Figure 3 (a) shows, when hearing "the computer", the gaze points of subjects stay on computers. However, there are two computers in the image, resulting in the gaze points being split into two groups. Such diverse targets make the commonly used Mean Square Error (MSE) loss confused. In the example, a middle point between two groups minimizes the MSE loss, which is undesired as no computer is there. Therefore, we need better objectives to measure whether a gaze trajectory is human-like.



(a) (b)

Figure 3: Illustration of our probability density score. (a) An example image with the gaze points of multiple subjects when they heard "the computer". (b) Its corresponding ground-truth distribution $\hat{P}_i$ visualized in a 3D way.

We propose a distribution-based measure, called Probability Density Score (PDS), instead of point-based measures like Euclidean Distance. First, we estimate the distribution $\hat{P}_i$ formed by multiple ground-truth gaze points (using Gaussian kernel density estimation) as the ground-truth distribution for word w_i, as shown in Figure 3 (b). Second, for a predicted gaze point $\hat{s}_i$, we measure how well it fits the ground-truth distribution by the normalized value of its probability density on this distribution:

$$PDS(\hat{s}_i) = \frac{\hat{P}_i(\hat{s}_i)}{\max_s \hat{P}_i(s)}. \tag{6}$$

In the example, a point near to the center of any group can get higher score than the middle point between groups. For a predicted gaze trajectory, we average the scores for all gaze points within the trajectory and get the final trajectory PDS. We finally utilize negative trajectory PDS as the training loss, i.e., probability density (PD) loss: $L(\hat{S}) = -\sum_{i=1}^{n} PDS(\hat{s}_i)$. PDS can also be used as a measurement, which shows excellent discriminative power in our comparison experiments.

5 Experiments

5.1 Experiment Setup

Implementation Details. We randomly split our collected data into training, validation, and test sets in an 8:1:1 ratio. We find that it is slow to optimize the PD loss, although it is more precise. Therefore, we adopt a two-stage training process for efficiency. In the first stage, we use MSE loss to optimize the model. After reaching a plateau in the loss, we continue training using our PD loss. All components of the model are jointly trained with the learning rate of 1e-4 and the optimizer is set to AdamW. Teacher forcing is utilized during training for sequence prediction. For more details, please refer to the Supplementary B. [1]

Baselines. Since our proposed task is new, there is no baseline model addressing this exact task. However, there are three types of methods that can be applied to solve our problem. (1) **Pre-trained image-text models.** Our task can be

[1]Our code and data are available at https://github.com/XiaochuanLiu-ruc/EyEar.

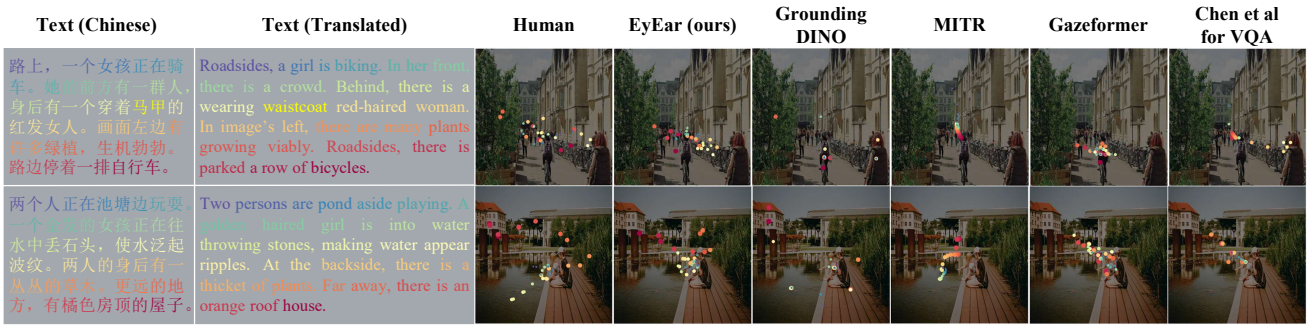


Figure 4: Visualization of the predicted gaze trajectories of different models and the ground-truth human gaze trajectories. Best viewed in color. We provide *word-to-word* translations for better understanding.

considered as a downstream task of text-visual alignment, so we choose BriVL (Huo et al. 2021) and CLIP (Radford et al. 2021) as baselines. (2) **Visual grounding models.** Visual grounding, which aims at precisely locating objects queried by natural language, can be seen as a highly simplified version of our task. We choose Grounding DINO (G-DINO) (Liu et al. 2023b) and MITR (Meng et al. 2021) as baselines. (3) **Gaze trajectory prediction models.** Some gaze trajectory prediction models in other task scenarios, such as visual searching (VS) and visual question answering (VQA), can also be applied to our task, because they have text-image multimodal inputs. We choose recent high-performer works Chen et al for VS (Chen, Jiang, and Zhao 2021), Gazeformer (Mondal et al. 2023) and Chen et al for VQA (Chen, Jiang, and Zhao 2021) as baselines. We provide the detailed explanation of how each baseline is constructed in Supplementary C. In addition, following previous research, we also use human gaze trajectories to compare with each other to obtain **Human** inter-subject similarity as the upper bound of model performance.

Metrics. In previous works, almost all widely-used metrics are **point-based metrics.** They treat a gaze trajectory as a sequence of gaze points and compare the similarity between the predicted and ground-truth sequences. Euclidean Distance (ED), Dynamic Time Wraping (DTW) (Müller 2007), and ScanMatch (Cristino et al. 2010) are widely used metrics. DTW uses a dynamic programming algorithm to discover an optimal alignment between two sequences, while ScanMatch uses the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). For point-based metrics, we average the scores compared to 8 ground-truth gaze trajectories as the final score. Our proposed PDS is a **distribution-based metric**. PDS uses distribution fitted from multiple ground-truth gaze points to evaluate the predicted gaze points.

5.2 Main Results

We compare EyEar with the baselines and present quantitative results in Table 2. It shows that EyEar consistently outperforms all the baselines in all metrics. The improvements over the best baselines are from 4% to 15% in different metrics. It indicates that although these baselines can be applied to our task, our method can better simulate human gaze.

Our proposed PDS is the most reasonable measurement and has the best discriminative power among the four metrics. In terms of the metrics ED, DTW, and ScanMatch, our model and some baselines even outperform Human result, which is theoretically the upper bound of model performance. However, in our proposed PDS, there is still more than ten points room to be onpar with Human. This indicates that those point-based metrics are ineffective in distinguishing between natural and artificial gaze trajectories. Furthermore, all the metrics except for PDS yield close results for different models. In contrast, the evaluation results of PDS clearly differentiate the performance of various models.

Model	ED ↓	DTW ↓	SMatch ↑	PDS ↑
Pre-trained image-text models				
CLIP	457.8	432.9	0.200	0.1689
BriVL	437.6	402.6	0.239	0.1758
Visual grounding models				
G-DINO	267.0	229.3	<u>0.443</u>	0.4628
MITR	<u>239.3</u>	234.8	0.403	0.4792
Gaze trajectory prediction models				
Chen et al (VS)	443.3	407.8	0.253	0.1586
Gazeformer	250.5	244.9	0.391	0.4807
Chen et al (VQA)	249.4	<u>228.6</u>	0.431	<u>0.5325</u>
EyEar (Ours)	**221.6 (+8.0%)**	**201.3 (+11.9%)**	**0.464 (+4.7%)**	**0.6138 (+15.3%)**
Human	272.9	238.9	0.438	0.7243

Table 2: Performance of different models on our test set. The best and runner-up are in **bold** and <u>underlined</u>. Improvements are calculated between the best to the runner-up.

Among the three types of baselines, gaze trajectory prediction models perform the best, while pre-trained image-text models perform the worst. This is because the scenario of gaze trajectory prediction models is closer to our task. Chen et al for VQA performs the best among the baselines, which we attribute to the fact that the VQA gaze trajectory prediction task is structurally the closest to ours, as both involve images and sentences as input. However, EyEar significantly outperforms Chen et al for VQA by 15% in terms of PDS. The improvements are attributed to the proposed three modules, which is indicated by ablation studies.

Model	ED ↓	DTW ↓	SMatch ↑	PDS ↑
EyEar	**221.6**	**201.3**	**0.464**	**0.614**
w/o Salience	224.4	202.5	0.460	0.602
w/o DynS	231.3	207.3	0.460	0.568
w/o GRU	230.4	210.1	0.452	0.555
w/o PD loss	237.3	237.3	0.397	0.515

Table 3: Overall ablation results. The best results are in **bold**.

5.3 Ablation Study

Overall Results. We verify the effectiveness of dynamical system (DynS), GRU in Module 2, and PD loss by ablating one of them at a time. When ablating the dynamical system, we simply leverage feedforward neural networks to integrate audio semantic attraction point and vision salient attraction point: $\hat{s}_i = \alpha \cdot MLP_1(s^{Sal}) + (1-\alpha) \cdot MLP_2(s_i^{Sem})$. We also remove vision salient attraction point, denoted by Salience, from DynS inputs to observe its impact. Table 3 shows the results. We have some findings from the table. (1) Ablating any component results in a performance drop over all the metrics. This indicates that all the proposed components contribute to the superior performance of EyEar. (2) Removing the PD loss incurs the largest performance drop consistently over all the metrics. It is expected, because a good loss function is crucial for model optimization. The probability density can better capture the loss between artificial and natural gaze trajectories from the perspective of statistics, making it better fit our data with high individual variability. (3) Ablating DynS and GRU results in similar performance drops, suggesting that capturing the motion patterns of eyes and achieving fine-grained audio-visual alignment in calculating the audio semantic attraction force hold equal importance in our task. (4) The removal of Salience has the least effect, indicating that the gaze trajectory is mainly influenced by the auditory stimuli, rather than the image itself, in our audio-visual scenarios.

The effect of DynS. We further provide a visualization result to observe the effect of DynS. Similar to (Dewhurst et al. 2012), we first decompose the gaze trajectories into saccade vectors pointing from the previous gaze point to the next. Then, we statistically visualize the angle, length, and speed of these vectors, as shown in Figure 5. In these radar charts, each coordinate corresponds to a kind of saccade vectors, e.g., the coordinate (400, 45°) indicates the group of vectors with a length of 400 pixels and an angle of 45 degrees counterclockwise from the horizontal direction. The color intensity in charts indicates their average speed, where the darker red color represents higher speed. In human data, we can observe that the speed of saccade vectors remains stable as the length increases. When the length is the same, the speeds in different directions are similar to uniform. Our model with DynS also exhibits these phenomena; whereas the model without DynS exhibits much higher speeds (in darker red) for longer lengths. This indicates that DynS can learn the motion patterns of eyes and does not generate saccades with abnormal speed.

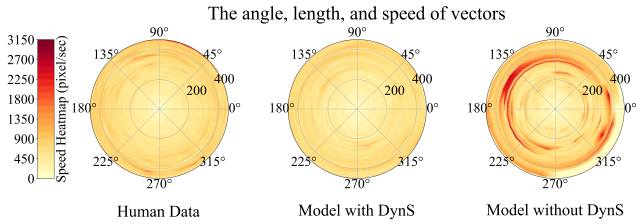


Figure 5: A radar chart showing the effect of DynS. Gaze trajectories are decomposed into saccade vectors pointing from the previous gaze point to the next. The degree of the polar coordinate represents the angle between the vectors and the horizontal direction. The radius of the polar coordinate is the length of the vectors. The heat map refers to the speed of vectors, calculated by length/duration.

5.4 Qualitative Analysis

As shown in Figure 4, we qualitatively compare the predicted gaze trajectories of different models to the ground-truth human gaze trajectories. It can be observed that EyEar predicts the most human-like gaze trajectories among all models in terms of both the fixation locations and the order of the fixations. EyEar is able to not only gaze at the region when its corresponding object is mentioned in the audio, but also exhibit motion patterns similar to human eyes. For example, as shown in the last line of Figure 4, the gaze trajectories of EyEar and human both make a loop and then move to the "orange roof house", while other models either fail to gaze at the correct locations or fail to simulate the motion patterns of eyes. Moreover, the visual grounding models like Grounding DINO only predict separate points without movement process. However, we still have big room to improve because even our model cannot predict as large range motions as Human.

6 Conclusion

Aiming at enabling virtual characters to better mimic human eyes, we introduce a new task that aims to predict a gaze trajectory when a person views an image while hearing a narration. We collect a dataset with 20k gaze points to support related research. To address the challenges in our task, we propose a framework EyEar. EyEar uses a physics-informed dynamical system (DynS) to simulate the motion of eyes. In DynS, three potential forces affecting the eyes' motion are considered. The most important force comes from audio semantic attraction points, which we design a multimodal attention mechanism to predict. We also propose the probability density loss and score for better optimization and evaluation. EyEar shows a notable performance gain, i.e., 15% in probability density score (PDS), compared to the baselines, indicating the uniqueness and challenge of our task. However, there is still a gap between EyEar and human, leaving room for future improvements. In the future, we plan to transition from static images to more complex videos, aiming to restore the continuous changes in visual stimuli as experienced in the real world. We are also interested to experiment more open audio stimuli that is not narrative of an image.

Ethical Statement

Due to page limitation, the supplementary material and reproducing details are publically available at https://github.com/XiaochuanLiu-ruc/EyEar/blob/main/AAAI2025_EyEar_Supplementary.pdf.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62276268, No. 92270118). We acknowledge Associate Professor Xiting Wang for providing valuable feedback and insightful suggestions that improved this paper. We would also like to express our gratitude to the Department of Psychology at Renmin University of China and Information Retrieval Lab at Tsinghua University for their support in providing the eye-tracking equipment used in this work.

References

Assens, M.; Giro-i-Nieto, X.; McGuinness, K.; and O'Connor, N. E. 2017. SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2331–2338.

Birkhoff, G. D. 1927. *Dynamical systems*, volume 9. American Mathematical Soc.

Borji, A.; and Itti, L. 2015. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. https://arxiv.org/abs/1505.03581v1.

Chen, S.; Jiang, M.; Yang, J.; and Zhao, Q. 2020. Air: Attention with reasoning capability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 91–107. Springer.

Chen, X.; Jiang, M.; and Zhao, Q. 2021. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10876–10885.

Chen, Y.; Yang, Z.; Ahn, S.; Samaras, D.; Hoai, M.; and Zelinsky, G. 2021. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1): 8776.

Chen, Y.; Yang, Z.; Chakraborty, S.; Mondal, S.; Ahn, S.; Samaras, D.; Hoai, M.; and Zelinsky, G. 2022. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5031–5040.

Cristino, F.; Mathôt, S.; Theeuwes, J.; and Gilchrist, I. D. 2010. ScanMatch: A Novel Method for Comparing Fixation Sequences. *Behavior Research Methods*, 42(3): 692–700.

Dewhurst, R.; Nyström, M.; Jarodzka, H.; Foulsham, T.; Johansson, R.; and Holmqvist, K. 2012. It Depends on How You Look at It: Scanpath Comparison in Multiple Dimensions with MultiMatch, a Vector-Based Approach. *Behavior Research Methods*, 44(4): 1079–1100.

Engbert, R.; Trukenbrod, H. A.; Barthelmé, S.; and Wichmann, F. A. 2015. Spatial Statistics and Attentional Dynamics in Scene Viewing. *Journal of Vision*, 15(1): 15.1.14.

Huo, Y.; Zhang, M.; Liu, G.; Lu, H.; Gao, Y.; Yang, G.; Wen, J.; Zhang, H.; Xu, B.; Zheng, W.; Xi, Z.; Yang, Y.; Hu, A.; Zhao, J.; Li, R.; Zhao, Y.; Zhang, L.; Song, Y.; Hong, X.; Cui, W.; Hou, D.; Li, Y.; Li, J.; Liu, P.; Gong, Z.; Jin, C.; Sun, Y.; Chen, S.; Lu, Z.; Dou, Z.; Jin, Q.; Lan, Y.; Zhao, W. X.; Song, R.; and Wen, J.-R. 2021. WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training. arxiv:2103.06561.

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. [DATASET] SALICON: Saliency in Context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1072–1080.

Judd, T.; Durand, F.; and Torralba, A. 2012. A Benchmark of Computational Models of Saliency to Predict Human Fixations.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR – Modulated Detection for End-to-End Multi-Modal Understanding. arxiv:2104.12763.

Krauzlis, R. J.; Lovejoy, L. P.; and Zénon, A. 2013. Superior colliculus and visual spatial attention. *Annual review of neuroscience*, 36: 165–182.

Kümmerer, M.; and Bethge, M. 2021. State-of-the-Art in Human Scanpath Prediction. arxiv:2102.12239.

Kümmerer, M.; Bethge, M.; and Wallis, T. S. A. 2022. DeepGaze III: Modeling Free-Viewing Human Scanpaths with Deep Learning. *Journal of Vision*, 22(5): 7.

Lan, G.; Scargill, T.; and Gorlatova, M. 2022. 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)*, 233–246.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022. Grounded Language-Image Pre-training. arxiv:2112.03857.

Linardos, A.; Kümmerer, M.; Press, O.; and Bethge, M. 2021. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. arXiv:2105.12441.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. arxiv:2304.08485.

Liu, H.; Xu, D.; Huang, Q.; Li, W.; Xu, M.; and Lin, S. 2013. Semantically-Based Human Scanpath Estimation with HMMs. In *2013 IEEE International Conference on Computer Vision*, 3232–3239.

Liu, S.; Liang, Y.; Li, F.; Huang, S.; Zhang, H.; Su, H.; Zhu, J.; and Zhang, L. 2022. DQ-DETR: Dual Query Detection Transformer for Phrase Extraction and Grounding. arxiv:2211.15516.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2023b. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arxiv:2303.05499.

Meng, Z.; Yu, L.; Zhang, N.; Berg, T.; Damavandi, B.; Singh, V.; and Bearman, A. 2021. Connecting What to Say With Where to Look by Modeling Human Attention Traces.

In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12674–12683.

Mondal, S.; Yang, Z.; Ahn, S.; Samaras, D.; Zelinsky, G.; and Hoai, M. 2023. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1441–1450.

Moschovakis, A.; Gregoriou, G.; and Savaki, H. 2001. Functional imaging of the primate superior colliculus during saccades to visual targets. *Nature neuroscience*, 4(10): 1026–1031.

Müller, M. 2007. Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Needleman, S. B.; and Wunsch, C. D. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48(3): 443–453.

Pont-Tuset, J.; Uijlings, J.; Changpinyo, S.; Soricut, R.; and Ferrari, V. 2020. Connecting Vision and Language with Localized Narratives. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, 647–664. Cham: Springer International Publishing. ISBN 978-3-030-58558-7.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arxiv:2103.00020.

Shen, C.; and Zhao, Q. 2014. Webpage Saliency. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, 33–46. Cham: Springer International Publishing. ISBN 978-3-319-10584-0.

Sun, W.; Chen, Z.; and Wu, F. 2021. Visual Scanpath Prediction Using IOR-ROI Recurrent Mixture Density Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 2101–2118.

Tavakoli, H. R.; Ahmed, F.; Borji, A.; and Laaksonen, J. 2017. [DATASET] Saliency Revisited: Analysis of Mouse Movements Versus Fixations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6354–6362.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xia, C.; Han, J.; Qi, F.; and Shi, G. 2019. Predicting Human Saccadic Scanpaths Based on Iterative Representation Learning. *IEEE Transactions on Image Processing*, 28(7): 3502–3515.

Xu, J.; Jiang, M.; Wang, S.; Kankanhalli, M. S.; and Zhao, Q. 2014. Predicting human gaze beyond pixels. *Journal of vision*, 14(1): 28–28.

Yang, Z.; Huang, L.; Chen, Y.; Wei, Z.; Ahn, S.; Zelinsky, G.; Samaras, D.; and Hoai, M. 2020. Predicting Goal-Directed Human Attention Using Inverse Reinforcement Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 190–199.

Zanca, D.; Melacci, S.; and Gori, M. 2020. Gravitational Laws of Focus of Attention. *IEEE transactions on pattern analysis and machine intelligence*, 42(12): 2983–2995.

Zaremba, W.; Sutskever, I.; and Vinyals, O. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.