SalDiff-DTM: A Novel Dual-Temporal Modulated Diffusion Model for Omnidirectional Images Scanpath Prediction

Xiaohui Kong[1], Qian Liu[2], Dandan Zhu[3*], Kaiwei Zhang[4*], Xiongkuo Min[5]

[1]Shanghai Institute of AI Education, East China Normal University, Shanghai 200062, China
[2]School of Computer Science and Technology, DongHua University, Shanghai 201620, China
[3]School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
[4]Shanghai AI Laboratory, Shanghai 200232, China
[5]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
xiaohui.kong@stu.ecnu.edu.cn, 2222845@mail.dhu.edu.cn, ddzhu@mail.ecnu.edu.cn,
{zhangkaiwei,minxiongkuo}@sjtu.edu.cn

Abstract

Scanpath prediction in omnidirectional images (ODIs) serves as a critical component for optimizing foveated rendering efficiency and enhancing interactive quality in virtual reality systems. However, existing scanpath prediction methods for ODIs still suffer from fundamental limitations: (1) inadequate modeling and capturing of long-range temporal dependencies in fixation regions, and (2) suboptimal integration of spatial and temporal visual features, ultimately compromising prediction performance. To address these limitations, we propose a novel Dual-Temporal Modulated Diffusion model for Omnidirectional Images Scanpath Prediction, named SalDiff-DTM model, to effectively generate realistic scanpaths. Specifically, to effectively model spatial relationships, we propose a novel Dual-Graph Convolutional Network (Dual-GCN) module that simultaneously captures semantic-level and image-level correlations. By integrating both local spatial details and global contextual information across the internal temporal dimension, this module achieves comprehensive and robust modeling of spatial relationships. To further enhance the modeling of temporal dependencies inherent in diverse fixation patterns, we introduce TABiMamba (Temporal-Aware BiLSTM-Mamba), a dedicated module that synergistically combines the contextual sensitivity of BiLSTM with the long-range sequence modeling capabilities of Mamba. This design facilitates deep information flow and context-aware sequential reasoning, thereby enabling high-fidelity capture of intricate temporal correlations. Inspired by the progressive refinement mechanism of diffusion models in various generative tasks, we propose a saliency-guided diffusion module that formulates the prediction problem as a conditional generative process, iteratively yielding accurate and perceptually plausible scanpaths. Extensive experiments demonstrate that SalDiff-DTM significantly outperforms state-of-the-art models, paving the way for future advancements in eye-tracking technologies and cognitive modeling.

Introduction

Virtual Reality (VR) has received increasing attention in recent years due to its broad applicability across various domains. Notably, VR has demonstrated significant benefits in education, rehabilitative medicine, and psychiatry, enabling immersive learning experiences, therapeutic interventions, and cognitive training (Al Farsi et al. 2021; Alhalabi 2016; Dhawan 2020; Mazurek et al. 2019; Satava and Jones 1998). At its core, VR provides a simulated 3D environment where users can interact, collaborate, and generate content in real time (Pensieri and Pennacchini 2016). In VR environment, omnidirectional images (ODIs) play a vital role in delivering immersive and interactive experiences, especially when viewed through VR headsets. During such

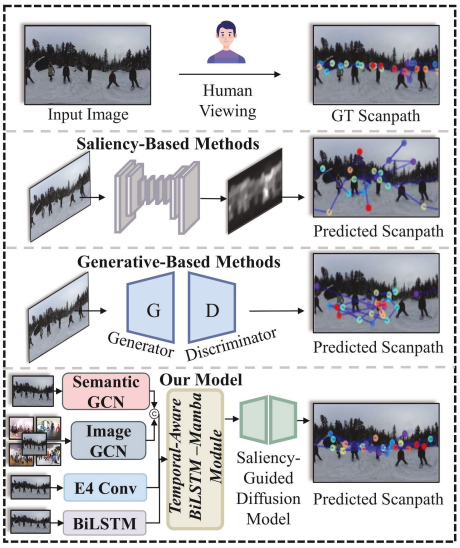

Figure 1: A clear comparison between existing scanpath prediction methods and our proposed SalDiff-DTM model. Saliency-based methods often produce scanpaths unstable gaze transitions, abrupt shifts, and insufficient focus on salient regions. Generative-based methods, while capable of producing plausible trajectories, tend to overlook important regions of interest. In contrast, our proposed SalDiff-DTM model effectively integrates spatial and temporal information, enabling the generation of more realistic scanpaths and human-aligned scanpaths.

*Corresponding Authors.

experiences, tracking users' gaze trajectories, referred to as scanpaths, enables the identification of areas of interest and user-specific attention patterns. These scanpaths can be further analyzed to uncover individual cognitive or perceptual traits, thereby facilitating the provision of personalized services and adaptive user experiences.

Recent progress in scanpath prediction has been primarily driven by four methodological paradigms: saliency-based methods (Assens Reina et al. 2017; Zhu, Zhai, and Min 2018; Zhu et al. 2019), generative-based methods (Assens et al. 2018; Martin et al. 2022), cognitive-inspired methods (Liu et al. 2013; Sun, Chen, and Wu 2019), and statistically-inspired methods (Boccignone and Ferraro 2004; Coutrot, Hsiao, and Chan 2018; Sui et al. 2023). They often still suffer from a universal drawback. A major shortcoming lies in their insufficient ability to model and capture long-range temporal dependencies across fixation regions—an aspect that is critical for predicting coherent and human-like scanpaths. To address this, transformer-based architectures (Quan et al. 2024; Wang, Zhang, and Dodgson 2024) have recently been introduced. While promising, their reliance on the self-attention mechanism leads to quadratic time and space complexity, resulting in high computational costs and memory overhead, thereby limiting their scalability in real-world applications.

Another key limitation lies in the suboptimal integration of spatial and temporal visual features, which are inherently complementary and, when effectively fused, can provide a more comprehensive understanding of visual scenes. Spatial features capture the structural and contextual attributes of an image, whereas temporal features reflect dynamic transitions and sequential dependencies. Suboptimal integration results in an incomplete representation, reducing the overall accuracy and effectiveness.

To tackle the aforementioned challenges, we propose a novel Dual-Temporal Modulated Diffusion model for Omnidirectional Images Scanpath Prediction, named SalDiff-DTM, to effectively generate realistic scanpaths. As a key component, we propose a novel Dual-Graph Convolutional Network (Dual-GCN) module, inspired by the need for joint semantic and structural understanding of visual scenes. This module simultaneously captures semantic-level and image-level correlations, enabling a comprehensive modeling of spatial relationships within omnidirectional content.

Building upon the constructed spatial relationships, we further enhance the modeling of temporal dependencies in diverse fixation patterns by introducing TABiMamba (Temporal-Aware BiLSTM-Mamba), a dedicated module that synergistically integrates the contextual sensitivity of BiLSTM with the long-range sequence modeling capabilities of Mamba. This module enables the extraction of temporally coherent and robust feature embeddings. To translate these enhanced spatiotemporal feature representations into realistic scanpath, we introduce a saliency-guided diffusion module, inspired by the progressive refinement mechanism of diffusion models in various generative tasks. By formulating the prediction issue as a conditional generative process, this module iteratively yields accurate and perceptually plausible eye movement trajectories. Figure 1 presents an intuitive comparison between the general pipeline of existing scanpath prediction methods and our proposed SalDiff-DTM model. In brief, the main contributions of this work are as follows:

- We propose a novel Dual-Temporal Modulated Diffusion model for Omnidirectional Images Scanpath Prediction, named SalDiff-DTM, to effectively produce genuine scanpaths, which helps better understand and replicate human visual perception.
- We propose a novel Dual-Graph Convolutional Network (Dual-GCN) module that effectively captures both local spatial details and global contextual cues across the internal temporal dimension, enabling more comprehensive and robust modeling of spatial relationships.
- We develop a saliency-guided diffusion model that formulates scanpath prediction as a conditional generative process, iteratively producing perceptual eye trajectories that closely resemble human viewing patterns.

Related Work

Scanpath Prediction

The task of scanpath prediction on ODIs has achieved great progress in recent years. The ways for modeling scanpaths of ODIs can be categorized to different kinds: saliency-based methods, generative-based methods, statistically-inspired methods, and Transformer-based methods. Based on the diverse applications in visual perception (Zhu et al. 2023a; Zhang et al. 2024a, 2025a,b,c), saliency-based methods generate scanpath by firstly producing the saliency map then sampling gaze points through probabilistic strategy (Assens Reina et al. 2017), which often incur temporal inconsistency. Generative-based methods (Assens et al. 2018; Martin et al. 2022) construct scanpaths by learning from human data. Generative Adversarial Networks (GANs) are mostly chosen to generate scanpaths. which are modeled sequentially by training discriminator and the generator, generating gaze points step by step from prior fixations and image features using models like RNNs, which excel at capturing spatio-temporal patterns but struggle with dynamic ROIs. Meanwhile, they are also less flexible in determining the length of scanpaths and commonly suffer from unstable training. As for statistically-inspired methods, ScanDMM (Sui et al. 2023) models dynamic gaze shift through Deep Markov Model (DMM), which employs state initialization, transition function and emission function for predicting scanpaths. However, it causes early fixations to have negligible influence on later predictions, thus failing in long-sequence modeling. Furthermore, their fixed transition probabilities lack flexibility across varied scenes. For Transformer-based methods (Quan et al. 2024; Wang, Zhang, and Dodgson 2024) demand high computation and memory usage because of the quadratic complexity of Transformer, limiting its scalability.

Diffusion Models

Compared with GANs, diffusion models inspired by nonequilibrium thermodynamics, which can generate high-

quality data with better stability and robustness, have outperformed many existing methods in many domains, such as image generation (Yoon et al. 2023; Saharia et al. 2022), audio generation (Kong et al. 2020), motion generation (Zhang et al. 2024b), symbolic music generation (Mittal et al. 2021) and so on. Score-based diffusion models (Song et al. 2020) utilize SDEs to transform complex data distributions into a known prior distribution by gradually adding noise. The application of diffusion models are diverse in various domains. As for image generation, diffusion models adopt U-Net architecture (Kazerouni et al. 2022). CONPREDIFF (Yang et al. 2023b) improves diffusion-based image synthesis with context prediction. For condition-guided tasks, diffusion models also show superiority in text-to-sound (Yang et al. 2023a) and image-to-text (Zhu et al. 2023b). In sequence generation tasks, natural language is also explored by using DiffuSeq model (Gong et al. 2022), achieving comparable or even better performance. TimeGrad (Rasul et al. 2021) employs diffusion probabilistic models for probabilistic time series forecasting. Another work is conditional score-based diffusion models (CSDI) (Tashiro et al. 2021), which proposes a conditional score-based diffusion models for time series imputation. To the best of our knowledge, there are few works that employ diffusion models to the task of scanpath prediction. Inspired by CSDI, we propose SalDiff-DTM to apply the generation capability of diffusion models to the scanpath prediction task.

Proposed Model

We propose an efficient model named SalDiff-DTM for scanpath prediction in ODIs with consistent spatiotemporal correlations. We present the overall architecture in the following section.

Overall Architecture

Unlike existing scanpath prediction methods, the proposed SalDiff-DTM model seeks to achieve precise scanpath prediction by integrating the modulation of dual-temporal constraints with the powerful generalization capabilities of the diffusion model. Specifically, our SalDiff-DTM model consists of four components: a Dual-GCN module, a TABi-Mamba module, a saliency-guided diffusion module and a scanpath predictor module. Our model architecture is shown in Figure 2.

Firstly, we devise a Dual-GCN module which includes a semantic-level GCN and an image-level GCN to capture innate spatial relationships. To further model temporal dependencies in various fixation patterns, we introduce an advanced TABiMamba module, utilizing the BiLSTM's contextual sensitivity and the Mamba's long-range sequence modeling abilities to capture intricate temporal correlations. In this module, we will gain the spatiotemporal-aware feature embeddings through two-phase feature processing. On the basis of that, we propose a saliency-guided diffusion module, employing the fused feature embeddings as conditions, along with outside diffusion timestamp constraint to capture the underlying spatiotemporal dependencies of embedded in the continuous human gaze sequences prior to accurate scanpath generation. Finally, in the scanpath predictor module, we apply Deep Convolutional Neural Networks (DCNNs) to acquire scanpaths which have powerful temporal correlation.

Dual-GCN Module. In order to learn spatial relationship of objects in ODIs, we design a Dual-GCN module which includes a semantic-level GCN and an image-level GCN inspired by (Dong et al. 2021). By leveraging region features extracted from the Faster R-CNN (Ren et al. 2016) object detector, our Dual-GCN module enhances spatial feature representation by capturing spatial relationships between objects in the same image and identifying the similarity of this image to other different images in the whole dataset.

Semantic-Level GCN. We adopt the semantic-level GCN to establish spatial relationships in various semantic objects within the same image. We treat each feature extracted from an individual semantic object as a node and regard the relative positions between different semantic objects as directed line segments. Specifically, we construct a spatial graph $G_{semantic} = (N_{semantic}, E_{semantic})$, where $E_{semantic} = \{(n_i, n_j)\}$ is the set of spatial relation edges between nodes, and the edge (n_i, n_j) refers to the relative geometry relationship between the i-th node and the j-th node. Here, we adopt the relative spatial distance between two nodes as their relative geometric relationship. It should be noted that if the relative spatial distance between two nodes exceeds half of the image width, they are considered unrelated. Therefore, each node n_i is encoded by a modified GCN as follows:

$$n_i^{semantic} = \sigma\left(\sum_{n_j \in S(n_i)} W_{n_j \to n_i} n_j + b_{n_j \to n_i}\right), \quad (1)$$

where each n_i is a real-valued vector representing the information of a specific semantic object in the graph. σ refers to the activation function, and $S(n_i)$ denotes the set of neighbor nodes of n_i. In addition, the notation $n_j \to n_i$ denotes the direction from n_j to n_i, and $W_{n_j \to n_i}$ and $b_{n_j \to n_i}$ are the transformation matrix and bias vector, respectively.

Image-Level GCN. It is observed that images sharing a certain degree of similarity with the input image can serve as valuable auxiliary information sources. On the basis of that, we propose a novel image-level GCN as a supplementary component. This module augments the limited information of a single image by incorporating knowledge from similar images, thereby improving the information flow within the graph structure.

Let $\overline{n}_j$ denotes the encapsulation of semantic content in the j-th image, which encompasses all the semantic objects denoted by $N^j_{semantic} = \{n_j^{semantic}\}_i^n$ within the image. Mathematically, we define it in the form of Eq. 2:

$$\bar{n}_j = \frac{1}{k}\sum_{i=1}^{k} n_i^{semantic}, \quad (2)$$

where k means the number of semantic objects in the j-th image. We obtain similar images of the input image by $\mathcal{L}_2$ normalization. Thus, the process can be formulated as follows:

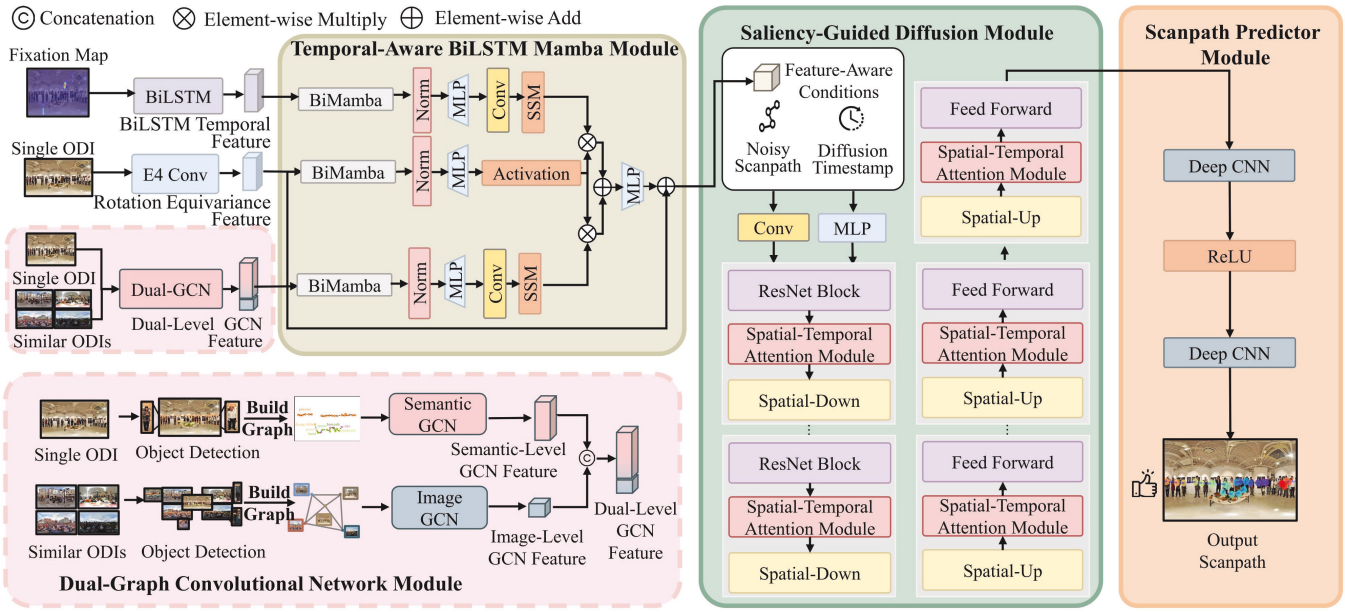

Figure 2: The overall SalDiff-DTM architecture. It includes several main components: a Dual-GCN module, a TABiMamba module, a saliency-guided diffusion module and a scanpath predictor module. TABiMamba module refers to Temporal-Aware BiLSTM Mamba module.

$$\bar{n}_s \in S(\bar{n}_j) := \left[\sum_{d=1}^{D} (\bar{n}_s^d - \bar{n}_j^d)^2 \right]_K , \tag{3}$$

where D refers to the dimension of the feature, and $[\cdot]_K$ means the selected K images with the lowest distance values. After that, the selected images are manipulated to construct an image-level graph G_{image}=(N_{image}, E_{image}), which comprises a set of nodes N_{image} and a set of edges E_{image}. Each node in N_{image} stands for an image feature, and each edge in E_{image} denotes the similarities between a specific image and pairs of other images. Then we encode the image features by a GCN:

$$m_j^{image} = \sigma \left(\sum_{\bar{n}_s \in S(\bar{n}_j)} W_m \bar{n}_s + b_m \right), \tag{4}$$

where $S(\bar{n}_j)$ stands for the set of similar images of $\bar{n}_j$ and m_j^{image} denotes the image feature of the j-th image along with image features of its K nearest neighbors.

Dual-GCN. To obtain local and global spatial semantic information comprehensively, we concatenate semantic-level features with image-level features. Mathematically, it conforms to the form of:

$$D_j = Concat(N_{semantic}^j, M_j^{image}), \tag{5}$$

where D_j is the Dual-GCN feature and $Concat$ means the concatenation operation.

Temporal-Aware BiLSTM-Mamba Module. To further capture temporal-aware feature embeddings, we propose a TABiMamba module, which incorporates different kinds of feature sources to gain spatiotemporal-aware feature embeddings, which will serve as conditions in the diffusion model. These features contains: (1) the rotation equivariance feature, (2) the BiLSTM feature containing temporal information, and (3) the Dual-GCN feature involving spatial relationships between semantic objects.

For the rotation equivariance feature, we employ E4 convolution network (He et al. 2021) to extract. Its rotation equivariance ensures consistent feature extraction from any orientation, which is essential for capturing a distortion-free 360-degree view. For temporal information, we utilize BiLSTM to capture temporal relationships, leveraging its contextual sensitivity drawn inspiration from (Schuster and Paliwal 1997). Firstly, we generate the fixation map based on the training scanpaths. Next, we sample the most salient points, along with the original image features around each collected point, which will be fed into the BiLSTM block. Specifically, we obtain the original image features by utilizing a bounding box to capture the image features around the most salient point, and then employing a Convolutional Neural Network (CNN) to extract the image features further. Subsequently, these image features are concatenated with the embeddings of the sampled points, which will act as the input of the BiLSTM block to learn the temporal dependencies between fixation points better.

After acquiring the extracted the Dual-GCN feature, the BiLSTM feature and the rotation equivariance feature, we design the TABiMamba module to gain the refined spatiotemporal-aware embeddings. It includes two phases: the first phase is for single modality feature extraction, and the second phase is for feature fusion of different modalities. In the first phase, the extracted features will undergo a

respective BiMamba block. Here, the BiMamba block deals with the aforementioned feature in a forward and a backward way so as to learn the long-range modality-specific correlations exhaustively. In the second phase, these features processed by BiMamba blocks will undergo a subtly designed M3 block to acquire more effective spatiotemporal correlated feature embeddings. It utilizes modality-specific features, i.e. the BiLSTM feature and the Dual-GCN feature to guide the generation of modality-correlated spatiotemporal feature embeddings, combining with the rotation equivariance feature with strong robustness. The integration of multiple modalities leads to richer and more complementary information, yielding a final feature embedding with stronger spatiotemporal correlations.

Saliency-Guided Diffusion Module. Diffusion models, known for superior generation ability in sequential data generation, provide a natural framework for modeling the complex distributions of eye movements owing to their gradual denoising mechanism. Leveraging the diffusion models, we design a saliency-guided diffusion module which is tailored to scanpath prediction. In this work, the forward process is defined as follows:

$$q(s_t|s_0) := \mathcal{N}(\sqrt{1-\beta_t}s_0, \beta_t I), \tag{6}$$

where $\beta_t \in [0, 1]$ indicates the noise level in time t, and s_0 is the original scanpath at the beginning time of the model. If we formulate $\hat{\alpha}_t := 1 - \beta_t$ and $\alpha_t := \prod_{i=0}^{t} \hat{\alpha_t}$, the pertubed scanpath data distribution can be gained by the following formula:

$$q(s_t \mid s_0) := \mathcal{N}(s_t; \sqrt{\alpha_t}s_0, (1-\alpha_t)I). \tag{7}$$

Therefore, the noised scanpath s_t is expressed as the form of:

$$s_t := \sqrt{\alpha_t}s_0, (1-\alpha_t)\epsilon, \tag{8}$$

where ϵ denotes the added noise to the original scanpath and $\epsilon \in \mathcal{N}(0, I)$, and I stands for idendity matrix.

In the reverse process, the denoising function recover the original scanpath s_0 from the noised scanpath s_t by the following Markov chain:

$$p_\theta(s_0|s_t) := \mathcal{N}(s_{t-1}; \mu_\theta(s_t, t), \sigma_\theta(s_t, t)I), \tag{9}$$

where $\sigma_\theta(s_t, t)$ is fixed to a constant α_t for easier optimization. $\mu_\theta(s_t, t)$ is divided into the linear combination of s_t and a denoising function $\epsilon_\theta(s_t, t)$, which can be optimized by the following loss function:

$$\min_\theta \mathcal{L}(\theta) := \min_\theta E_{s_0\sim q(s_0),\, \epsilon\sim\mathcal{N}(0,I),\, t} \|\epsilon - \epsilon_\theta(s_t, t)\|_2^2, \tag{10}$$

where $\mathcal{L}$ stands for Mean Absolute Error (MAE) loss.

During the generation process of SalDiff-DTM, we incorporate a dual-temporal modulated mechanism to regulate time-related behavior. The external modulation is driven by the diffusion timestamp feature, while the internal modulation comes from the scanpath sequence timestamp feature. Together, the dual-temporal modulation strategy comprehensively stimulates the compelling scanpath prediction, ensuring both global temporal coherence and local temporal consistency. In general, the pseudo algorithm of SalDiff-DTM training can be found in the supplementary material.

Scanpath Predictor Module. After generating preliminary scanpaths using the saliency-guided diffusion module, we introduce a scanpath predictor module designed to refine these trajectories, aligning them more closely with human visual gaze patterns. This module consists of two DCNNs followed by a ReLU non-linear activation function. Through this refinement process, the resulting scanpaths exhibit enhanced temporal coherence and more realistic sequential dependencies.

Experiment

Dataset

Four datasets are used to conduct our experiments: Sitzmann (Sitzmann et al. 2018), Salient360! (Rai, Gutiérrez, and Le Callet 2017), AOI (Xu et al. 2021), and JUFE (Fang et al. 2022), which are highly recognized datasets in the task of scanpath prediction. The Sitzmann (Sitzmann et al. 2018) dataset includes 22 high-resolution ODIs and near 2,000 scanpaths from 169 viewers. The Salient360! (Rai, Gutiérrez, and Le Callet 2017) dataset contains 85 images and 3,036 scanpaths from 42 observers. The AOI (Xu et al. 2021) dataset consists of 600 images with four diverse categories of ODIs: cityscapes, natural landscapes, indoor scenes, and human scenes, which contains 18,000 scanpaths. The JUFE (Fang et al. 2022) dataset consists of 1,032 images and 30,960 scanpaths, which is used for quality evaluation. We train our model on the Sitzmann datasets, sampling the raw scanpaths at 1 Hz on each dataset. All the Salient360! test dataset are used for testing. For the latter two datasets, we randomly select 20% images for testing. By leveraging these datasets, we can thoroughly assess our model's performance across various scenarios.

Performance Comparison

Quantitative Comparison. Table 1 displays the results by employing three selected metrics to quantitatively compare the performance of different representative SOTA models. Experimental setup is detailed in the supplementary material. Our model exhibits superior overall performance, outperforming other SOTA methods. In particular, it achieves a lower LEV score and a higher REC score, which can be attributed to the TABiMamba module, generating spatiotemporal-aware features, thereby improving prediction accuracy. Besides, BiLSTM feature captures the intricate temporal correlation between fixation points, building a strong temporal connection.

Qualitative Comparison. We demonstrate the visualized results of generated scanpaths across different datasets in Figure 3. The color transition from purple to red represents the progression of the scanpath, with purple indicating the starting point and red marking the endpoint. For CLE (Boccignone, Cuculo, and D'Amelio 2020) and DeepGazeIII (Kümmerer, Bethge, and Wallis 2022), they exhibit considerable positional deviation due to their design for 2D images. SaltiNet (Assens Reina et al. 2017) suffers from significant displacements caused by the instability of saliency sampling strategy. Even though methods like ScanGAN (Martin

Dataset	Model	LEV↓	DTW↓	REC↑
Salient360!	Random walk	40.802	2231.681	2.774
	CLE	39.985	1747.164	1.282
	DeepGazeIII	42.337	1872.488	1.821
	SaltiNet	43.999	2022.041	1.287
	ScanGAN	43.597	1837.925	1.461
	ScanDMM	38.251	1672.968	3.403
	Pathformer3D	39.066	1645.650	3.114
	Ours (SalDiff-DTM)	**36.263**	**1552.055**	**4.384**
	Human	35.084	1382.590	5.202
Sitzmann	Random walk	48.942	2232.987	2.669
	CLE	48.449	2106.828	1.072
	DeepGazeIII	48.638	2167.110	1.778
	SaltiNet	52.384	2376.972	1.266
	ScanGAN	52.058	2217.107	1.407
	ScanDMM	45.175	**1953.319**	3.364
	Pathformer3D	47.441	1974.809	2.819
	Ours (SalDiff-DTM)	**43.996**	1984.446	**3.569**
	Human	41.188	1836.986	6.345
AOI	Random walk	13.696	711.516	2.993
	CLE	13.420	531.374	1.924
	DeepGazeIII	14.098	608.326	2.056
	SaltiNet	14.711	647.066	1.363
	ScanGAN	14.366	553.722	2.329
	ScanDMM	13.056	591.885	3.409
	Pathformer3D	13.405	498.546	3.263
	Ours (SalDiff-DTM)	**12.593**	**471.671**	**4.226**
	Human	9.243	389.477	6.228
JUFE	Random walk	24.039	1193.725	3.109
	CLE	25.126	1139.619	0.984
	DeepGazeIII	25.059	1181.411	2.126
	SaltiNet	26.617	1277.556	1.281
	ScanGAN	26.090	1149.259	1.657
	ScanDMM	23.448	1105.248	3.369
	Pathformer3D	24.655	1068.393	2.847
	Ours (SalDiff-DTM)	**22.700**	**1026.104**	**4.118**
	Human	18.306	1038.880	7.745

Table 1: Quantitative comparison of different scanpath prediction models on four benchmark datasets. **Bolded** indicates the better values.

et al. 2022) and ScanDMM (Sui et al. 2023) show some improvements, the former lacks emphasis on salient regions, resulting in more scattered prediction results, while the latter fails to capture the relationship between global and local fixations due to its reliance on a single level of temporal constraint. Pathformer3D (Quan et al. 2024) shows a significant improvement over ScanGAN and ScanDMM by utilizing Transformer architecture to capture salient information. However, it lacks spatial semantic information to guide meaningful scanpath prediction. In contrast to these approaches, SalDiff-DTM has the ability to generate scanpaths that are closer to the realistic scanpaths by introducing dual-temporal modulation through inner temporal constraint of Dual-GCN and outer diffusion temporal constraint to generate faithful scanpaths.

Ablation Study

We conduct ablation experiments to evaluate the effectiveness of each core module. Specifically, the Dual-GCN module, the BiLSTM module, and TABiMamba module are individually ablated to validate their contributions to the overall performance.

Effectiveness of Dual-GCN Module. To verify the influence of Dual-GCN module, we analyze the performance under four different configurations: (1) the model incorporating Dual-GCN feature GCN_{dual}, (2) the model using only semantic-level GCN feature $GCN_{semantic}$, (3) the model utilizing only image-level GCN feature GCN_{image}, and (4) the model excluding Dual-GCN feature GCN_{no}. The results are presented in Table 2. We can clearly observe the effectiveness of the Dual-GCN module. Both semantic-level GCN features and image-level GCN features contributes to achieving superior performance, highlighting the importance of the Dual-GCN module in enhancing overall model performance.

Dataset	GCN	LEV↓	DTW↓	REC↑
Salient360!	GCN$_{no}$	36.275	1552.069	4.334
	GCN$_{semantic}$	36.277	1552.943	4.317
	GCN$_{image}$	36.296	1552.543	4.331
	GCN$_{dual}$	**36.263**	**1552.055**	**4.384**
AOI	GCN$_{no}$	12.599	472.286	4.216
	GCN$_{semantic}$	12.596	471.735	4.188
	GCN$_{image}$	12.601	472.061	4.183
	GCN$_{dual}$	**12.593**	**471.671**	**4.226**

Table 2: Ablation experiments of different GCN configurations in our model. GCN_{no}, $GCN_{semantic}$, GCN_{image} and GCN_{dual} stands for no Dual-GCN feature, only semantic-level GCN feature, only image-level GCN feature and Dual-GCN feature in the model, respectively.

Effectiveness of BiLSTM. To validate the effectiveness of BiLSTM, we run ablated experiments in our model. $Model_{noBiLSTM}$ refers to the model without BiLSTM and $Model_{withBiLSTM}$ stands for model with BiLSTM, respectively. The results are shown in Table 3. Performance degradation after removing BiLSTM highlights its importance in capturing contextual temporal information. BiLSTM is particularly good at modeling complex temporal dependencies by combining image features and fixation embeddings.

Dataset	BiLSTM	LEV↓	DTW↓	REC↑
Salient360!	Model$_{noBiLSTM}$	36.264	1552.832	4.340
	Model$_{withBiLSTM}$	**36.263**	**1552.055**	**4.384**
AOI	Model$_{noBiLSTM}$	12.594	471.911	4.203
	Model$_{withBiLSTM}$	**12.593**	**471.671**	**4.226**

Table 3: Ablation experiments of different BiLSTM configurations in our model. $Model_{noBiLSTM}$ means that no BiLSTM exists in the model, and $Model_{withBiLSTM}$ means there is a BiLSTM in the model.

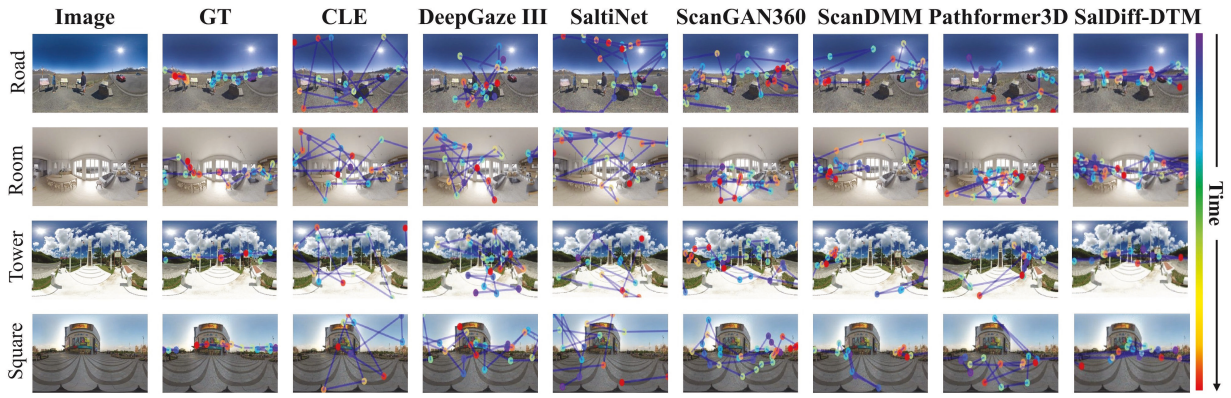

Figure 3: Qualitative comparison of various scanpath prediction models across different scenes. These scenes are selected from the *Road* in Salient360!, the *Room* in Sitzmann, the *Tower* in AOI, and the *Square* in JUFE. From left to right: the original image, scanpath prediction results obtained by humans, predicted scanpath yielded by CLE (Boccignone, Cuculo, and D'Amelio 2020), DeepGaze III (Kümmerer, Bethge, and Wallis 2022), SaltiNet (Assens Reina et al. 2017), ScanGAN360 (Martin et al. 2022), ScanDMM (Sui et al. 2023), Pathformer3D (Quan et al. 2024), and our proposed SalDiff-DTM model.

Effectiveness of TABiMamba Module. We conduct comparative experiments to verify the performance of TABiMamba module. Four configurations are included: (1) the model without BiMamba and M3 block, (2) the model with BiMamba block, (3) the model with M3 block, and (4) the model with TABiMamba, respectively. To verify the synergy function of BiLSTM feature and Dual-GCN feature, the model without BiLSTM feature and Dual-GCN feature is also incorporated. The results are shown in Table 4. The results indicate that the two-stage TABiMamba module shows an outstanding ability to capture spatiotemporal-aware information.

Dataset	Configurations	LEV↓	DTW↓	REC↑
Salient360!	$\text{Model}_{noBiLSTMGCN}$	36.281	1552.917	4.323
	$\text{Model}_{noBiMambaM3}$	36.283	1552.480	4.331
	$\text{Model}_{BiMamba}$	36.287	1555.292	4.342
	Model_{M3}	36.283	1552.090	4.333
	$\text{Model}_{TABiMamba}$	**36.263**	**1552.055**	**4.384**
AOI	$\text{Model}_{noBiLSTMGCN}$	12.595	471.904	4.161
	$\text{Model}_{noBiMambaM3}$	12.600	472.062	4.177
	$\text{Model}_{BiMamba}$	12.595	473.805	4.205
	Model_{M3}	12.603	472.613	4.211
	$\text{Model}_{TABiMamba}$	**12.593**	**471.671**	**4.226**

Table 4: Ablation experiments of different Mamba blocks in our model. $Model_{noBiMambaM3}$, $Model_{BiMamba}$, $Model_{M3}$, and $Model_{TABiMamba}$ stands for neither BiMamba block nor M3 block, only BiMamba block, only M3 block, and TABiMamba exists in our model, respectively. $Model_{noBiLSTMGCN}$ denotes neither BiLSTM feature and Dual-GCN feature exists.

Efficiency Analysis

We conduct a comparative efficiency analysis across methods of parameter size, inference latency of generating a single scanpath, and memory usage for predicting 100 scanpaths on Sitzmann test dataset. As shown in Table 5, SalDiff-DTM offers favorable trade-offs: it has the smallest parameter size and inference latency among all methods while maintaining a reasonable memory usage than many methods, making it ideal for real-time VR applications.

Model	Param. size	Inf. latency	Mem. Usage
CLE	-	30.8992 s	-
DeepGazeIII	78.9MB	0.7219 s	100.75MB
SaltiNet	98.7MB	348.2060 s	1438.06 MB
ScanGAN	31.4MB	0.4017 s	**29.63 MB**
ScanDMM	17.8MB	0.2887 s	1276.09 MB
Pathformer3D	19.3MB	2.2089 s	462.60MB
Ours	**10.1MB**	**0.2537 s**	232.87MB

Table 5: Comparison of parameter size (Param. size), inference latency (Inf. latency), and memory usage (Mem. usage) on different models.

Conclusion

We propose SalDiff-DTM, a novel Dual-Temporal Modulated Diffusion model for ODIs scanpath prediction. We incorporate different spatial and temporal features to generate powerful conditions to increase the robustness and generalization ability of SalDiff-DTM. Lightweight Mamba structure enables lower inference latency. Extensive experiments demonstrate that SalDiff-DTM can generate accurate scanpaths surpassing SOTA models on different benchmark datasets. Our study underscores the model's exceptional capacity to learn and represent the complex spatiotemporal dynamics that characterize human visual attention behavior.

Acknowledgments

This work was supported by the Shanghai Artificial Intelligence Laboratory, and by the National Natural Science Foundation of China under Grant 62377011.

References

Al Farsi, G.; Yusof, A. b. M.; Romli, A.; Tawafak, R. M.; Malik, S. I.; Jabbar, J.; and Bin Rsuli, M. E. 2021. A Review of Virtual Reality Applications in an Educational Domain. *International Journal of Interactive Mobile Technologies*, 15(22).

Alhalabi, W. 2016. Virtual reality systems enhance students' achievements in engineering education. *Behaviour & Information Technology*, 35(11): 919–925.

Assens, M.; Giro-i Nieto, X.; McGuinness, K.; and O'Connor, N. E. 2018. PathGAN: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Assens Reina, 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 *Proceedings of the IEEE international conference on computer vision workshops*, 2331–2338.

Boccignone, G.; Cuculo, V.; and D'Amelio, A. 2020. How to look next? A data-driven approach for scanpath prediction. In *Formal Methods. FM 2019 International Workshops: Porto, Portugal, October 7–11, 2019, Revised Selected Papers, Part I 3*, 131–145. Springer.

Boccignone, G.; and Ferraro, M. 2004. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2): 207–218.

Coutrot, A.; Hsiao, J. H.; and Chan, A. B. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior research methods*, 50(1): 362–379.

Dhawan, S. 2020. Online learning: A panacea in the time of COVID-19 crisis. *Journal of educational technology systems*, 49(1): 5–22.

Dong, X.; Long, C.; Xu, W.; and Xiao, C. 2021. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the 29th ACM international conference on multimedia*, 2615–2624.

Fang, Y.; Huang, L.; Yan, J.; Liu, X.; and Liu, Y. 2022. Perceptual quality assessment of omnidirectional images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 580–588.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

He, L.; Chen, Y.; Dong, Y.; Wang, Y.; Lin, Z.; et al. 2021. Efficient equivariant network. *Advances in Neural Information Processing Systems*, 34: 5290–5302.

Kazerouni, A.; Aghdam, E. K.; Heidari, M.; Azad, R.; Fayyaz, M.; Hacihaliloglu, I.; and Merhof, D. 2022. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Kümmerer, M.; Bethge, M.; and Wallis, T. S. 2022. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5): 7–7.

Liu, H.; Xu, D.; Huang, Q.; Li, W.; Xu, M.; and Lin, S. 2013. Semantically-based human scanpath estimation with HMMs. In *Proceedings of the IEEE International Conference on Computer Vision*, 3232–3239.

Martin, D.; Serrano, A.; Bergman, A. W.; Wetzstein, G.; and Masia, B. 2022. Scangan360: A generative model of realistic scanpaths for 360 images. *IEEE Transactions on Visualization and Computer Graphics*, 28(5): 2003–2013.

Mazurek, J.; Kiper, P.; Cieślik, B.; Rutkowski, S.; Mehlich, K.; Turolla, A.; and Szczepańska-Gieracha, J. 2019. Virtual reality in medicine: a brief overview and future research directions. *Human Movement*, 20(3): 16–22.

Mittal, G.; Engel, J.; Hawthorne, C.; and Simon, I. 2021. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*.

Pensieri, C.; and Pennacchini, M. 2016. Virtual reality in medicine. In *Handbook on 3D3C Platforms: Applications and Tools for Three Dimensional Systems for Community, Creation and Commerce*, 353–401. Springer.

Quan, R.; Lai, Y.; Qiu, M.; and Liang, D. 2024. Pathformer3D: A 3D Scanpath Transformer for 360 Images. In *European Conference on Computer Vision*, 73–90. Springer.

Rai, Y.; Gutiérrez, J.; and Le Callet, P. 2017. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, 205–210.

Rasul, K.; Seward, C.; Schuster, I.; and Vollgraf, R. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*, 8857–8868. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Satava, R. M.; and Jones, S. B. 1998. Current and future applications of virtual reality for medicine. *Proceedings of the IEEE*, 86(3): 484–489.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11): 2673–2681.

Sitzmann, V.; Serrano, A.; Pavel, A.; Agrawala, M.; Gutierrez, D.; Masia, B.; and Wetzstein, G. 2018. Saliency in VR: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4): 1633–1642.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sui, X.; Fang, Y.; Zhu, H.; Wang, S.; and Wang, Z. 2023. Scandmm: A deep markov model of scanpath prediction for 360deg images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6989–6999.

Sun, W.; Chen, Z.; and Wu, F. 2019. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 2101–2118.

Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34: 24804–24816.

Wang, Y.; Zhang, F.-L.; and Dodgson, N. A. 2024. Scantd: 360° scanpath prediction based on time-series diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7764–7773.

Xu, M.; Yang, L.; Tao, X.; Duan, Y.; and Wang, Z. 2021. Saliency prediction on omnidirectional image with generative adversarial imitation learning. *IEEE Transactions on Image Processing*, 30: 2087–2102.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2023a. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1720–1733.

Yang, L.; Liu, J.; Hong, S.; Zhang, Z.; Huang, Z.; Cai, Z.; Zhang, W.; and Cui, B. 2023b. Improving diffusion-based image synthesis with context prediction. *Advances in Neural Information Processing Systems*, 36: 37636–37656.

Yoon, J. S.; Zhang, C.; Suk, H.-I.; Guo, J.; and Li, X. 2023. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, 388–400. Springer.

Zhang, K.; Zhu, D.; Min, X.; Duan, H.; and Zhai, G. 2024a. Explain Vision Focus: Blending Human Saliency Into Synthetic Face Images. *IEEE Transactions on Multimedia*.

Zhang, K.; Zhu, D.; Min, X.; and Zhai, G. 2025a. Mesh Mamba: A Unified State Space Model for Saliency Prediction in Non-Textured and Textured Meshes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16219–16228.

Zhang, K.; Zhu, D.; Min, X.; and Zhai, G. 2025b. Textured mesh saliency: Bridging geometry and texture for human perception in 3d graphics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9977–9984.

Zhang, K.; Zhu, D.; Min, X.; and Zhai, G. 2025c. Unified Approach to Mesh Saliency: Evaluating Textured and Non-Textured Meshes Through VR and Multifunctional Prediction. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024b. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6): 4115–4128.

Zhu, D.; Zhang, K.; Zhang, N.; Zhou, Q.; Min, X.; Zhai, G.; and Yang, X. 2023a. Unified audio-visual saliency model for omnidirectional videos with spatial audio. *IEEE Transactions on Multimedia*, 26: 764–775.

Zhu, Y.; Li, Z.; Wang, T.; He, M.; and Yao, C. 2023b. Conditional text image generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14235–14245.

Zhu, Y.; Zhai, G.; and Min, X. 2018. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69: 15–25.

Zhu, Y.; Zhai, G.; Min, X.; and Zhou, J. 2019. The prediction of saliency map for head and eye movements in 360 degree images. *IEEE Transactions on Multimedia*, 22(9): 2331–2344.