

Supplementary Materials for ScanTD: 360° Scanpath Prediction based on Time-Series Diffusion

Anonymous Authors

1 DATASET PROCESSING.

Coordinate Transformation of Gaze Points. A scanpath consists of multiple gaze points, and the original datasets we obtained provide the longitude and latitude coordinates of these gaze points: $Coordi(g) = (lon, lat)$. During dataset processing, the coordinate transformation mainly involves two processes: Initially, the standard transformation formula from spherical coordinates to Cartesian coordinates is used to represent gaze points in three-dimensional space. Specifically: $x = \cos(lat) \cdot \cos(lon)$, $y = \cos(lat) \cdot \sin(lon)$, $z = \sin(lat)$. At this point, the coordinates of gaze points have been transformed into $Coordi(g) = (x, y, z)$. Then, the three-dimensional coordinates are mapped back to the two-dimensional image coordinate system. Specifically: recalculating new longitude and latitude coordinates from the three-dimensional coordinates (x, y, z) :

$$lon = \arctan 2(y, x)$$

$$lat = \arctan 2\left(z, \sqrt{x^2 + y^2}\right)$$

. These longitude and latitude coordinates are converted to normalized two-dimensional image coordinates:

$$x' = \frac{lon + \pi}{2\pi}$$

$$y' = \left(\frac{lat}{\pi/2} + 1\right)/2$$

. Finally, the normalized coordinates (x', y') are converted to pixel coordinates by multiplying them by the width and height of the image:

$$x'' = x' \cdot \text{width}$$

$$y'' = y' \cdot \text{height}$$

. Now, the coordinates of gaze points have been transformed into:

$$Coordi = (x'', y'')$$

Sampling of Gaze Points. For the Sitzmann dataset [7], special dataset augmentation and sampling methods are described in the paper. For the AOI dataset [9], the length of each ground truth scanpath is different, so we uniformly sample 20 gaze points for training our model. Scanpaths with fewer than 20 gaze points are filled with linear interpolation to reach 20 and scanpaths with more than 20 gaze points, the first 20 points are sampled according to the temporal order. For the Salient360! dataset [6], we employ the same sampling method as AOI dataset [9].

2 SPHERICAL CNN

Spherical CNN. We employ Spherical CNNs to extract local convolutional features. The Spherical CNN layers are designed to account for the geometric properties of a sphere. Specifically, there are two operations that differ from traditional CNNs: for spherical convolution (*SphereConv2D*), utilizing neighborhood information on the sphere instead of local neighborhoods in a flat image. The

convolution kernels are applied in a spherical coordinate system, not in the traditional Cartesian coordinate system. Similarly, for spherical pooling (*SphereMaxPool2D*), this operation is adjusted to accommodate the geometry of the sphere.

ViT Embedding with Spherical CNN. The convolutional kernel sizes in the x and y directions are in a ratio of 2 : 1. The image size is (448, 224), and the patch size is (32, 16).

3 QUANTITATIVE COMPARISON IN SALIENCY DETECTION

Evaluation Metrics. Following [8], we also evaluate saliency detection with four metrics: the Judd variant of the area under curve (AUC) [1], normalized scanpath saliency (NSS) [5], correlation coefficient (CC) [4], and Kullback–Leibler divergence (KLD) [2].

When calculating the Judd variant of the area under curve (AUC), ground truth gaze points are used as positive samples, and random sampling of other points serves as negative samples. Different thresholds are then applied to the saliency map generated by the model, each producing a pair of True Positive Rate (TPR) and False Positive Rate (FPR) values. After plotting TPR against FPR, the area under the resulting curve represents the AUC [1]. The Normalized Scanpath Saliency (NSS) measures the degree of match between the saliency map generated by a model and ground truth gaze points. It is calculated by analyzing the values at the real gaze points on the model’s saliency map [5]. The Correlation Coefficient (CC) measures the linear correlation between the saliency map generated by a saliency detection model and the ground truth gaze points. We use the Pearson correlation coefficient to calculate the CC [4]. The Kullback-Leibler Divergence (KLD) is used to quantify the difference between the predicted saliency map and the ground truth. The calculation involves summing the product of the probabilities in the predicted distribution and the logarithmic difference between the predicted and ground truth distributions [2].

Quantitative Performance Comparison. We apply different methods (ScanDMM [8], ScanGAN360 [3], and ScanTD) in saliency detection on three datasets: AOI [9], Salient360! [6] and Sitzmann [7]. The quantitative performance comparison results are illustrated in Tab. 1. Tab. 1 shows that compared with ScanGAN360 [3] and ScanDMM [8], our approach ScanTD can achieve better performance when applying to saliency detection on these three datasets.

4 MORE EXPERIMENTAL RESULTS

4.1 Diverse Results of ScanTD in Scanpaths Generation

As shown in Fig. 1, our ScanTD can generate multiple scanpaths for the same scene to satisfy the diverse requirements of viewers, which is critical for practical applications. Moreover, the different scanpaths are plausible and capable of focusing on meaningful and relevant areas within the scene. The street and Monument scenes

Table 1: Quantitative comparison in saliency detection on three datasets

Database	Method	AUC \uparrow	NSS \uparrow	CC \uparrow	KLD \downarrow
AOI [9]	ScanGAN360 [3]	0.69	0.81	0.38	0.79
	ScanDMM [8]	0.75	0.88	0.41	0.62
	ScanTD	0.81	0.92	0.46	0.40
	Human	0.89	1.94	1.00	0.15
Salient360! [6]	ScanGAN360 [3]	0.70	0.72	0.39	0.60
	ScanDMM [8]	0.75	0.92	0.57	0.41
	ScanTD	0.79	1.13	0.64	0.38
	Human	0.91	2.07	1.00	0.18
Sitzmann [7]	ScanGAN360 [3]	0.74	0.86	0.45	0.66
	ScanDMM [8]	0.73	1.04	0.52	0.49
	ScanTD	0.84	1.37	0.59	0.38
	Human	0.88	2.46	1.00	0.10

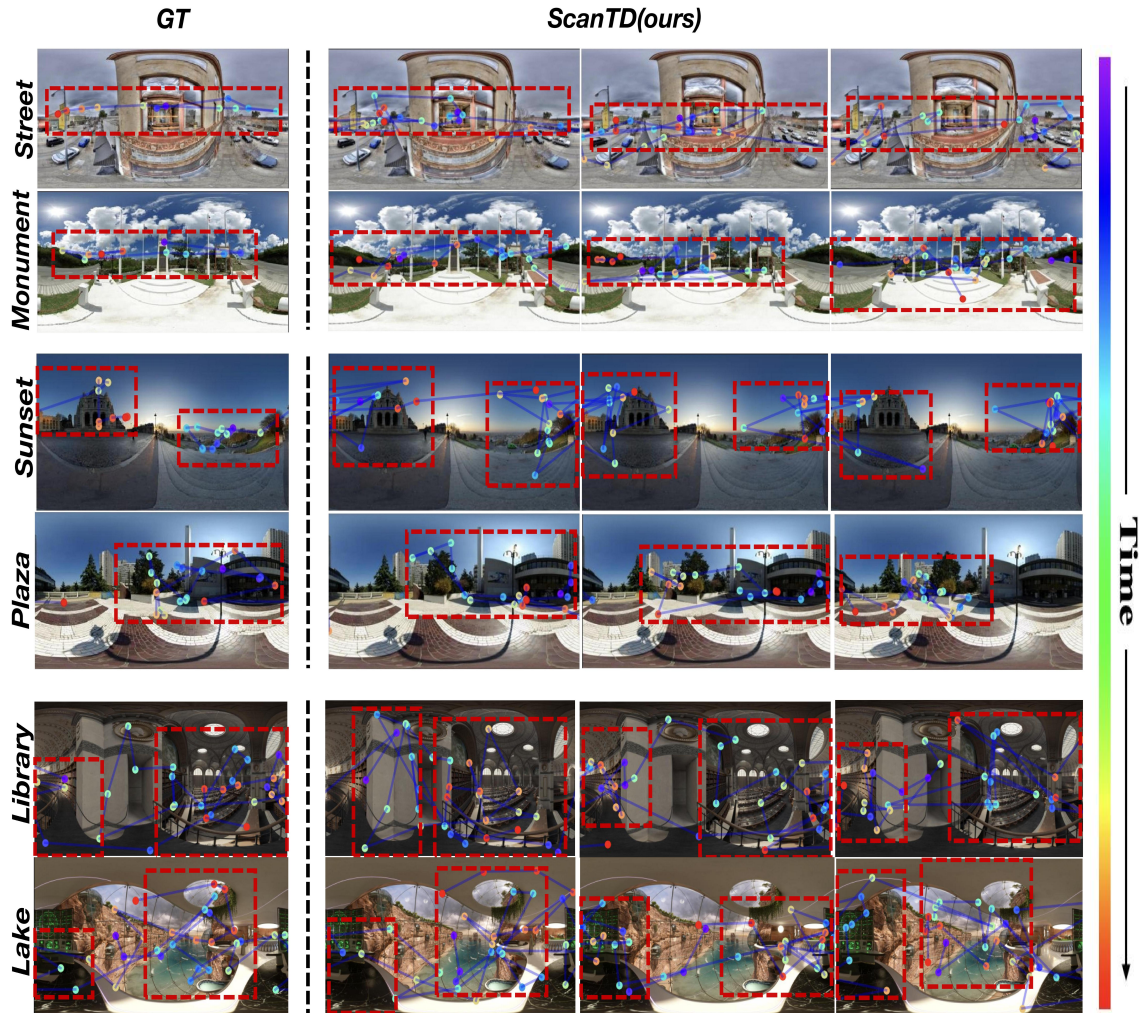


Figure 1: Diverse generation results of ScanTD on three datasets

are from AOI dataset [9], and the sunset and plaza scenes are from

Salient360! [6], and the library and lake scenes are from Sitzmann dataset [7].

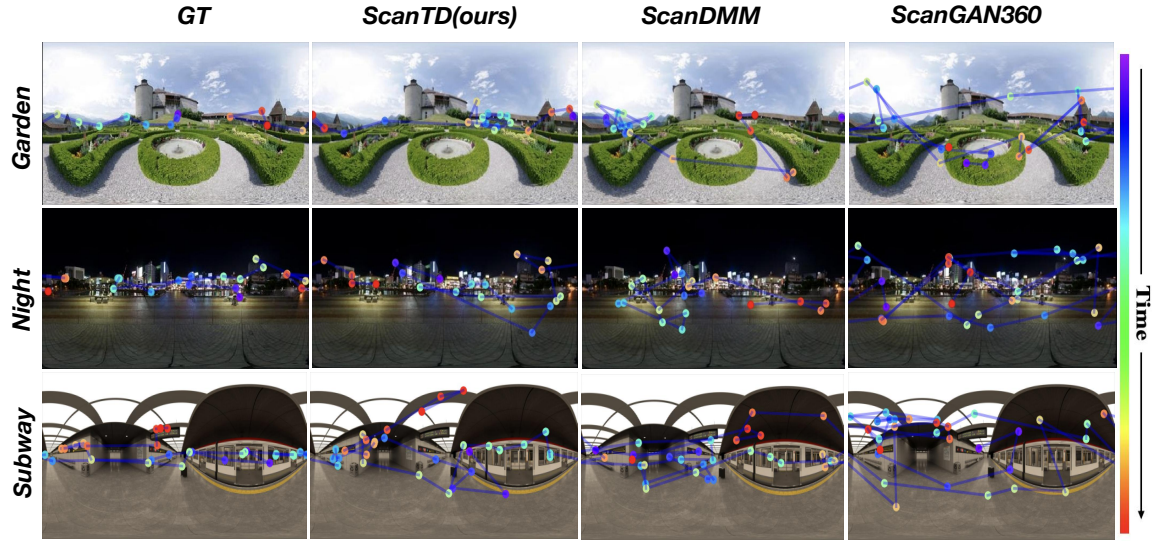


Figure 2: Qualitative comparison to different scanpath prediction models on three datasets

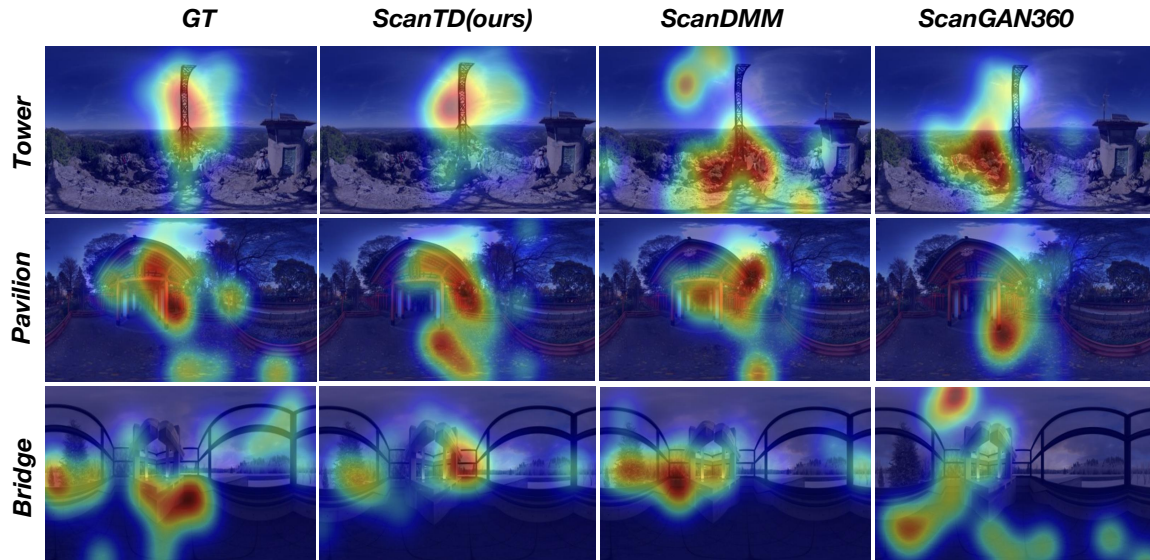


Figure 3: Qualitative comparison to different saliency detection models on three datasets

4.2 Qualitative Comparison in Scanpath Prediction

As shown in Fig. 2, for the first garden scene and the second night scene, our ScanTD is closer to the ground truth and the predicted gaze points do not exhibit unreasonable large-span displacements in the vertical direction. Particularly, in the second night scene, the majority of the gaze points predicted by ScanTD are concentrated on meaningful buildings. For the third subway scene, according to the color distribution of the predicted gaze points, it can be demonstrated that our ScanTD is better able to capture the temporal sequence of generated gaze points which is a crucial aspect in the task of scanpath prediction. The garden scene is from AOI dataset

[9], the night scene is from Salient360! dataset [6], and the subway scene is from Sitzmann dataset [7].

4.3 Qualitative Comparison in Saliency Detection

As illustrated in Fig. 3, for these three distinct scenarios from disparate datasets, our approach ScanTD can more accurately highlight the spatial locations and distribution of salient regions within the ground truth. This capability of ScanTD has wide-ranging practical applications, including image recognition systems, environmental monitoring, and surveillance technologies, where pinpointing salient regions is critical in decision-making and analysis. The tower

scene is from AOI dataset [9], the pavilion scene is from Salient360! dataset [6], and the bridge scene is from Sitzmann dataset [7].

REFERENCES

[1] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Fredo Durand, Aude Oliva, and Antonio Torralba. 2015. MIT saliency benchmark. *MIT* (2015).

[2] Matthias Kummerer, Thomas S. A. Wallis, and Matthias Bethge. 2015. Information-theoretic model comparison unifies saliency metrics. *National Academy of Sciences* 112, 52 (2015), 16054–16059.

[3] Daniel Martin, Ana Serrano, Alexander W. Bergman, Gordon Wetzstein, and Belen Masia. 2022. ScanGAN360: A generative model of realistic scanpaths for 360° images. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2003–2013. <https://doi.org/10.1109/TVCG.2022.3150502>

[4] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. 2007. Predicting visual fixations on video based on low-level visual features. *Vision Research* 47, 19 (2007), 2483–2498.

[5] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research* 45, 18 (2005), 2397–2416.

[6] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. A Dataset of Head and Eye Movements for 360 Degree Images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 205–210.

[7] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1633–1642. <https://github.com/vsitzmann/vr-saliency>

[8] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. 2023. ScanDMM: A Deep Markov Model of Scanpath Prediction for 360° Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 6989–6999.

[9] Mai Xu, Li Yang, Xiaoming Tao, Yiping Duan, and Zulin Wang. 2021. Saliency Prediction on Omnidirectional Image With Generative Adversarial Imitation Learning. *IEEE Transactions on Image Processing* 30 (2021), 2087–2102. <https://doi.org/10.1109/TIP.2021.3050861>