# Target Scanpath-Guided 360-Degree Image Enhancement

**Yujia Wang, Fang-Lue Zhang*, Neil A. Dodgson**

Victoria University of Wellington, New Zealand
{yujia.wang, fanglue.zhang, neil.dodgson}@vuw.ac.nz

## Abstract

360° images have wide applications in fields such as virtual reality and user experience design. Our goal is to adjust these images to guide users' visual attention. To achieve this, we present a novel task: target scanpath-guided 360° image enhancement, which aims to enhance 360° images based on user-specified target scanpaths. We develop a Progressive Scanpath-Guided Enhancement Method (PSEM) to address this problem through three stages. In the first stage, we propose a Time-Alignment and Spatial Similarity Clustering (TASSC) algorithm that accounts for the spherical nature of 360-degree images and the temporal dependency of scanpaths to generate representative scanpaths. In the second stage, we learn the differences between the source and the target scanpaths and select the objects to be edited based on these differences. Particularly, we propose a Dual-Stream Scanpath Difference Encoder (DSDE) embedded into the Segment Anything Model (SAM) network for object mask generation. Finally, we employ a Stable Diffusion network fine-tuned with LoRA technology to produce the final enhanced image. Additionally, we design special loss functions to supervise the training of the second and third stages. Experimental results have demonstrated the effectiveness of our approach for scanpath-guided 360-degree image enhancement.

## Introduction

When users view 360° images in an immersive environment, their gaze movements form trajectories known as scanpaths. These scanpaths reflect the dynamic changes in users' visual attention over time, capturing both the spatial distribution of gaze points and the temporal sequence of fixations (Sui et al. 2023). In 360° environments, viewers often get lost in information-rich scenes captured from the real world. This may cause them to miss critical content and fail to capture important information promptly, thereby affecting scene comprehension and degrading the quality of the user experience. Therefore, we propose a novel task: target scanpath-guided 360° image enhancement as shown in Figure 1, aiming to edit the content of 360° images to guide users' visual attention and enhance scene perceptual quality.


Figure 1: Target Scanpath-guided 360° Image Enhancement

Compared to visual attention models that focus solely on the spatial positions of gaze points, such as salient regions (Hu et al. 2023; Zhang et al. 2018; Zhou et al. 2023) and fixation points (Deng et al. 2024; Qiao et al. 2020), scanpaths represent a more complex and precise visual pattern, as they also reflect the temporal order of gaze points. Therefore, scanpath-guided image enhancement introduces new challenges that existing saliency-guided image enhancement methods do not encounter (Jiang et al. 2021; Miangoleh et al. 2023; Mejjati et al. 2020). Firstly, since different individuals produce diverse scanpaths in the same 360° scene (Martin et al. 2022), we need to explore how to describe the source scene's visual attention pattern to compare it with the target scanpath. Secondly, modeling the relationship between the 360° content and the spatio-temporal differences of the source and the target scanpaths poses a new challenge that cannot be trivially addressed by any existing models. Finally, despite the impressive recent progress in 2D image generation achieved by diffusion models and their variants (Rombach et al. 2022; Shen et al. 2023; Zhang, Rao, and Agrawala 2023), these models cannot be directly applied to generate enhanced, realistic 360° images.

In this research, we propose a **P**rogressive **S**canpath-Guided **E**nhancement **M**ethod (PSEM) to address the above challenges. Our method consists of three stages, each focusing on specific sub-tasks to improve model interpretability and robustness (Shen et al. 2023). In the first stage, we estimate multiple diverse scanpaths from the source image and develop a clustering method to identify the most representative scanpath, which describes the visual attention pattern of the source image. Considering the temporal dependency of scanpaths and the spherical nature of 360° images, we combine Fast Dynamic Time Warping (FastDTW) (Froese et al. 2023) and Great Circle Distance (GCD) (Gava et al. 2023) to cluster scanpaths of 360° images. In the second

*Corresponding author

stage, we design a Dual-Stream Difference Encoder (DSDE) to capture the temporal and spatial differences between the given target scanpath and the source scanpath thereby identifying objects that need to be edited. DSDE employs relative time and spatial position encoding and is integrated into SAM (Kirillov et al. 2023). Finally, we leverage LoRA technology (Hu et al. 2021) and our own dataset to fine-tune Stable Diffusion models (Rombach et al. 2022), generating the final enhanced 360° image. We design specialized loss functions to guide the training process of our networks. Experimental results demonstrate that our approach effectively generates 360° images that align with the target scanpaths.

## Related Works

### Saliency-Guided Image Enhancement

In recent years, saliency-guided image enhancement has gained widespread attention for its potential to manipulate visual attention in images. These studies primarily focus on traditional 2D images, where target saliency maps serve as guidance. For example, GazeShiftNet (Mejjati et al. 2020) model redirects visual attention by applying global parameter transformations to both foreground and background regions, representing a significant advancement in visual focus manipulation. Subsequently, SalG-GAN (Jiang et al. 2021) introduced a saliency-based attention module and a disentangled representation framework, generating images that align with target saliency maps. Building on this, (Aberman et al. 2022) proposed a method to edit salient regions in 2D images through backpropagation using a pre-trained visual saliency prediction model, enabling more precise and targeted modifications. More recently, (Miangoleh et al. 2023) achieved a balance between reducing inference time and enhancing target objects by combining saliency loss with realism loss, producing high-quality images that maintain realism while satisfying target saliency requirements.

However, unlike saliency maps, scanpaths include both the spatial and temporal information of gaze points, reflecting more complex visual attention patterns. Effectively enabling models to learn and utilize the temporal information in scanpaths remains a critical challenge. Moreover, most current methods for using saliency maps to determine which objects to mask involve either overlapping the original object saliency with the target object saliency (Jiang et al. 2021; Miangoleh et al. 2023) or using source images rely on source images where users have manually selected specific objects (Mejjati et al. 2020). In contrast, scanpaths do not necessarily focus on specific objects, which poses new challenges for object-level image enhancement. Additionally, 360° images depict panoramic environments with richer objects and more complex scene information compared to 2D images. This added complexity makes it more difficult to analyze and utilize scanpaths. These factors underscore the need for innovative approaches to scanpath-guided 360° image enhancement.

### Pre-trained Segmentation and Generation Models

We adopt several pre-trained image segmentation and generation models in our approach. The Segment Anything Model (SAM) (Kirillov et al. 2023) is pre-trained on the large-scale SA-1B dataset and can segment any objects using various prompts, demonstrating exceptional zero-shot generalization capabilities for the rapid generation of high-quality segmentation masks. It also adapts well to distortions in 360° images without requiring specific labels. Therefore, we use a pre-trained SAM model for the second stage, leveraging its pre-trained knowledge of diverse scenes and objects. Stable Diffusion (SD) (Rombach et al. 2022) is a pre-trained model based on latent diffusion and exhibits excellent image generative capabilities. Numerous studies and applications have focused on fine-tuning SD (Shen et al. 2023; Brooks, Holynski, and Efros 2023) for image inpainting. These works employ various fine-tuning strategies, such as ControlNet (Zhang, Rao, and Agrawala 2023), which incorporates an additional conditional control network to handle multiple input modalities, and lightweight methods like LoRA (Hu et al. 2021) for parameter-efficient adaptive training. We found that running the vanilla SD (Rombach et al. 2022) model demonstrated its potential for 360° image enhancement, although the output results were not always stable. Given the limitations of our training dataset size and computational resources, we chose to fine-tune SD using LoRA, which enables efficient parameter adaptation in resource-constrained settings.

## Methods

### Overview of PSEM

The framework of our target scanpath-guided 360° image enhancement method, PSEM, is illustrated in Figure 2. PSEM consists of three stages: (1) Given the predicted scanpaths of a source 360° image, we apply a Temporal Alignment and Spatial Similarity Clustering (TASSC) method to generate a representative scanpath as the source scanpath. (2) The Dual-Stream Difference Encoder (DSDE) analyzes the spatiotemporal differences between the source and the target scanpaths. These differences are then combined with image features and fed into an attention module to generate the final mask, which identifies the objects to be edited to match the target scanpath. (3) The source image and the mask are fed into the SD network, which is fine-tuned with LoRA, to produce the final enhanced 360° image.

**Problem Definition** A scanpath $S$ is defined as a temporal sequence of $N$ gaze points $\{s_1, ..., s_N\}$. Given a source image $I_s$ and a target scanpath $S_t$ containing a sequence of gaze points, our goal is to generate an enhanced image $I_e$ with a scanpath that matches $S_t$. In the dataset preprocessing stage, we use an external scanpath prediction model (Wang, Zhang, and Dodgson 2024) to estimate multiple scanpaths $\{S_p^i\}_{i=1,...,M}$ for the source image $I_s$.

### Stage 1: Source Scanpath Generation

We propose a Temporal Alignment and Spatial Similarity Clustering (TASSC) algorithm to cluster multiple predicted scanpaths of a 360° image into a single representative source scanpath. TASSC effectively addresses two primary issues in scanpath analysis: the spherical nature of 360° images (Zhang et al. 2023; Li et al. 2022) and the temporal

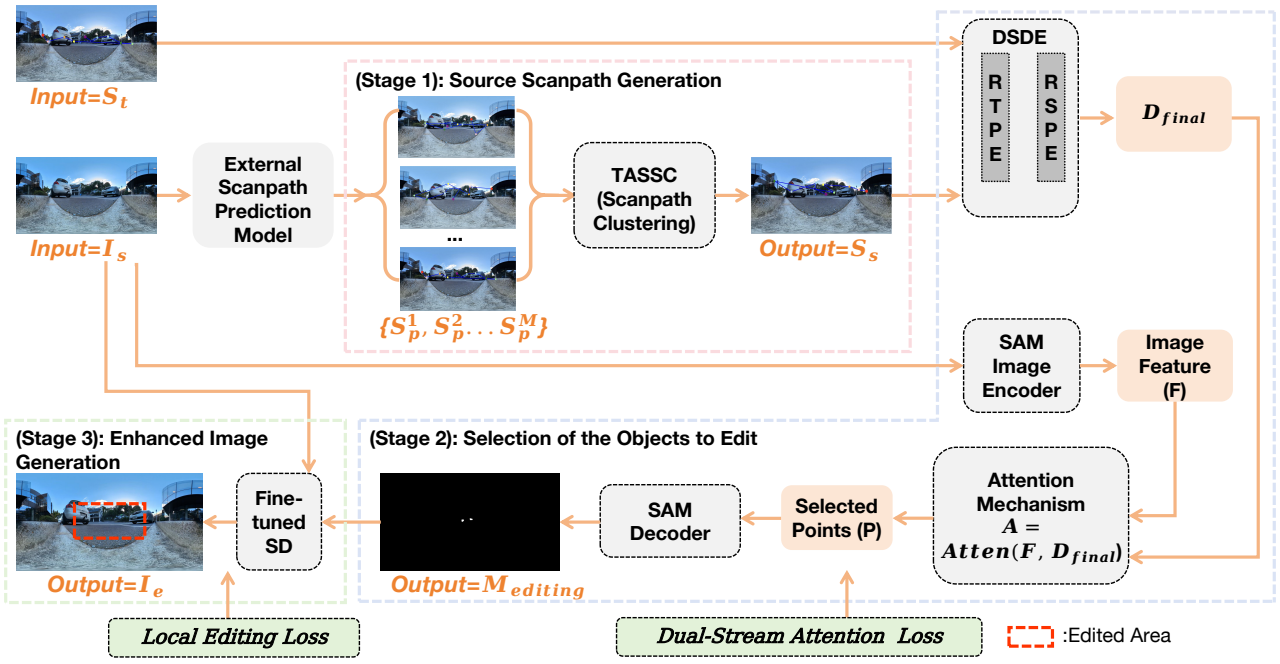


Figure 2: The framework of PSEM. First, a scanpath prediction model generates multiple scanpaths for the source image ($I_s$). From these, a representative source scanpath ($S_s$) is extracted. Next, the Dual-Stream Scanpath Difference Encoder (DSDE) analyzes the spatiotemporal differences between $S_s$ and the target scanpath, utilizing the SAM Encoder to create an editing mask. Finally, a fine-tuned Stable Diffusion model produces the enhanced image ($I_e$).

dependency of scanpaths.

The predicted scanpaths for a single image are represented as $\{S_p^1, S_p^2, \ldots, S_p^M\}$, where each path $S_p^k = \{s_{p,k}^1, s_{p,k}^2, \ldots, s_{p,k}^{N_k}\}$. Each gaze point $s_{p,k}^i$ is represented by its spherical coordinates $s_{p,k}^i = (\phi_{p,k}^i, \lambda_{p,k}^i)$, where $\phi_{p,k}^i$ denotes the latitude and $\lambda_{p,k}^i$ is the longitude. To improve the stability of our method, we preprocess the predicted scanpaths by removing outlier gaze points based on the distances between consecutive gaze points. If the distance between a point and its neighbors exceeds the mean by more than 2.5 standard deviations, it is identified as a potential spatial outlier (Holmqvist et al. 2011). Gaze points with significantly longer time intervals are identified as potential temporal outliers (Salvucci and Goldberg 2000). A gaze point is confirmed as an outlier only if it is classified as both a spatial and temporal outlier. Outliers are then removed and replaced using spherical linear interpolation to maintain continuity (Hessels et al. 2017). Particularly, the spatial distance between two gaze points on the spherical surface is measured by the Great Circle Distance (GCD) (Gava et al. 2023) as:

$$\begin{aligned} \mathrm{GCD}(s_1, s_2) = r \cdot \arccos\Big(&\sin(\phi_1)\sin(\phi_2) \\ &+ \cos(\phi_1)\cos(\phi_2)\cos(\Delta\lambda)\Big) \end{aligned} \tag{1}$$

where $(\phi_1, \lambda_1)$ and $(\phi_2, \lambda_2)$ are the spherical coordinates of $s_1$ and $s_2$, $\Delta\lambda = |\lambda_1 - \lambda_2|$, and $r$ is the radius of the sphere (assumed to be 1 for a unit sphere). We also use GCD to obtain a distance matrix $D$, which represents the differences between all the scanpaths. For each entry, $D(i,j) = GCD(s_{p,k}^i, s_{p,l}^j)$. Then we apply the FastDTW (Froese et al. 2023) algorithm to the distance matrix $D$, where we set the radius $r$ as 20. We employ the Complete Linkage method (Murtagh and Contreras 2012) for hierarchical clustering, which does not assume any specific cluster shape and works well with non-Euclidean distance measures such as the Great Circle Distance. We determine the number of clusters ($k = 30$) experimentally. Let $\{C_1, C_2, \ldots, C_k\}$ denote the set of all clusters. For each cluster $C_i$, we generate an intermediate representative scanpath using the FastDTW Barycenter Averaging (FDBA) method (Petitjean, Ketterlin, and Gancarski 2011) and select the scanpath closest to the cluster center as the initial representative scanpath. We iteratively align all scanpaths in the cluster to the intermediate representative sequence and update it by averaging the aligned scanpaths at each time step. The barycenter $B_i(t)$ at each time step $t$ for cluster $C_i$ is calculated as follows:

$$\mathrm{B}_i(t) = \frac{1}{|C_i|} \sum_{S \in C_i} S(t) \tag{2}$$

where $|C_i|$ is the number of scanpaths in cluster $C_i$, and $S(t)$ is the value of each scanpath in $C_i$ at time $t$. For each cluster $C_i$, we iteratively apply FDBA until one of the following conditions is met: (i) the maximum number of iterations, set to 300, is reached, or (ii) the change in the central sequence between consecutive iterations is less than a predefined threshold of 0.001. This ensures convergence of the

Algorithm 1: Temporal Alignment and Spatial Similarity Clustering (TASSC)

**Input:**
$\{S_p^1, S_p^2, \ldots, S_p^M\}$ - Predicted scanpaths
Each $S_p^k = \{s_{p,k}^1, \ldots, s_{p,k}^{N_k}\}$, each $s_{p,k}^i = (\phi_{p,k}^i, \lambda_{p,k}^i)$
**Output:** Source scanpath $S_s$
Remove outliers
**for** each pair $S_p^k$ and $S_p^l$ **do**
    Calculate GCD matrix $D(i,j)$ using Equation (2)
    Apply FastDTW($D$, radius=20)
**end for**
Get clusters $\{C_1, C_2, \ldots, C_k\}$ by Complete Linkage
**for** each cluster $C_i$ **do**
    Select scanpath closest to the cluster center as initial central scanpath $B_i$
    **while** iteration $< 300$ **and** $\Delta B_i > 0.001$ **do**
        Align all scanpaths in $C_i$ to $B_i$
        Update $B_i(t)$ for all time steps by:
        $B_i(t) \leftarrow \frac{1}{|C_i|} \sum_{S \in C_i} S(t)$
        Record the change $\Delta B_i$
    **end while**
**end for**
Apply FDBA on $\{B_k\}$
**return** the final source scanpath $S_s$

algorithm to a stable result within a reasonable time. To obtain a single representative scanpath for the source image, referred to as the 'source scanpath', we apply the FDBA method once more to the $k$ intermediate representative scanpaths. This final step integrates the characteristics of all clusters into a unified source scanpath $S_s$, providing a comprehensive representation of the viewing behavior for the given $360^\circ$ image.

## Stage 2: Selection of the Objects to Edit

The second stage of our method generates a mask indicating the objects to be edited. We propose a Dual-Stream Scanpath Difference Encoder (DSDE) integrated with the pre-trained SAM model for the mask prediction. The DSDE is designed to capture the temporal and spatial differences between the source scanpath $S_s = \{s_s^1, s_s^2, \ldots, s_s^N\}$ and a given target scanpath $S_t = \{s_t^1, s_t^2, \ldots, s_t^N\}$. Before inputting these two scanpaths into the DSDE module, we convert the spherical coordinates $s_{\mathrm{s},k}^i = (\phi_{\mathrm{s},k}^i, \lambda_{\mathrm{s},k}^i)$ into normalized 2D planar coordinates $(x_i, y_i)$ to facilitate processing by neural networks. Additionally, we use $t_i$ to represent the time stamp of these gaze points for temporal sequence encoding.

First, we use the pre-trained SAM encoder to process $I_{\mathrm{s}}$ and extract features $F$, with a feature dimension of 256. Then, we apply Relative Temporal Position Encoding (RTPE) and Relative Spatial Position Encoding (RSPE) to encode the temporal and spatial information of the scanpaths, respectively. RTPE is conducted based on the temporal order difference $\Delta t = |t_i - t_j|$, using the following formulas:

$$\mathrm{RTPE}(2i) = \sin\left(\frac{\Delta t}{10000^{2i/d_{\text{model}}}}\right) \tag{3}$$

$$\mathrm{RTPE}(2i+1) = \cos\left(\frac{\Delta t}{10000^{2i/d_{\text{model}}}}\right) \tag{4}$$

where $d_{\text{model}}$ is set to 128. RSPE is encoded based on the spatial coordinate differences $\Delta x = |x_i - x_j|$ and $\Delta y = |y_i - y_j|$, with the following formula:

$$\mathrm{RSPE}(\Delta x, \Delta y) = \left[\sin\frac{\Delta x}{\sigma}, \cos\frac{\Delta x}{\sigma}, \sin\frac{\Delta y}{\sigma}, \cos\frac{\Delta y}{\sigma}\right] \tag{5}$$

where $\sigma$ is set to 100.

These encodings are passed through a 6-layer Transformer encoder to obtain the temporal sequence representations $T_s^{\text{temp}}$ and $T_t^{\text{temp}}$, as well as the spatial distribution representations $T_s^{\text{spat}}$ and $T_t^{\text{spat}}$ for the source and target scanpaths. Next, we calculate their temporal difference as $D_{\text{temp}} = \mathrm{MLP}(T_s^{\text{temp}} - T_t^{\text{temp}})$, and the spatial difference as $D_{\text{spat}} = \mathrm{MLP}(T_s^{\text{spat}} - T_t^{\text{spat}})$. These differences are integrated through a feature fusion layer to generate the final difference feature $D_{\text{final}} = \mathrm{MLP}(\mathrm{Concat}[D_{\text{temp}}, D_{\text{spat}}])$, where the MLP layer contains 512 hidden units. Using $D_{\text{final}}$ and $F$ from the SAM encoder, we create an attention map $A = \mathrm{Attention}(F, D_{\text{final}})$. The attention mechanism employs an 8-head multi-head attention, with each head having a dimension of 32. We then select two points with the highest attention values from $A$, denoted as $P = \{P_1, P_2\}$. This selection is motivated by our goal of editing one or two objects in the image, which is typically sufficient for personalized image enhancement tasks. This approach allows us to focus on the most salient areas that differ between the source and target scanpaths, while maintaining computational efficiency. Finally, the selected points $P$ and image features $F$ are input into the SAM decoder to generate the final segmentation mask: $M_{\text{editing}}$. The decoder consists of four convolutional blocks, each with a 3x3 kernel, where the number of channels decreasing from 256 to 64.

**Training** We designed a Dual-Stream Attention Loss, $L_{\text{DSA}}$, to guide the training of the second stage. We do not rely on an external scanpath prediction model because it could introduce additional errors and significantly increase system complexity, complicating training and optimization. Our approach offers a more direct and efficient method for utilizing scanpath information, specifically targeting the selection of two key points that best reflect the differences between scanpaths. Our loss function considers multiple factors, including attention focus, spatial diversity, difference consistency, and attention distribution. It is defined as follows:

$$\begin{aligned} L_{\text{DSA}} = {} & L_{\text{attn}} + \lambda_{\text{div}} \cdot L_{\text{div}} \\ & + \lambda_{\text{cons}} \cdot L_{\text{cons}} + \lambda_{\text{dist}} \cdot L_{\text{dist}} \end{aligned} \tag{6}$$

Specifically, the attention focus loss is defined as: $L_{\text{attn}} = 1 - 0.5 \cdot (A(P_1) + A(P_2))$. Here, $A(P_k)$ represents the attention value at point $P_k$, ranging from 0 to 1. This loss aims

to guide the model to focus more on the two points with the highest attention values. When their average attention value approaches 1 (maximum), $L_{\text{attn}}$ approaches 0, achieving the optimal state. This design effectively penalizes the selection of low-attention points, encouraging the model to focus on the most salient areas in the image, which typically correspond to the main differences between the source and target scanpaths.

The spatial diversity loss ensures that the selected points are adequately separated in space, which is given by:

$$L_{\text{div}} = \exp\left(-\frac{\|P_1 - P_2\|^2}{2 \cdot \sigma^2}\right). \tag{7}$$

The difference consistency loss is $L_{\text{cons}} = ||D_{\text{final}} - f(S_{\text{s}}, S_{\text{t}})||^2$. $D_{\text{final}}$ is the difference feature output by the DSDE, and $f(S_{\text{s}}, S_{\text{t}})$ is the expected difference calculated based on the source and target scanpaths. The attention distribution loss is $L_{\text{dist}} = \text{KL}(A||U)$, which is the Kullback-Leibler divergence between the attention distribution $A$ and a uniform distribution $U$. $L_{\text{cons}}$ ensures that the difference feature generated by the DSDE aligns with the actual scanpath differences. $L_{\text{dist}}$ prevents the over-concentration of attention in certain areas of the image. We experimentally set the parameters as: $\lambda_{\text{div}} = 0.2$, $\lambda_{\text{cons}} = 0.1$, $\lambda_{\text{dist}} = 0.05$, and $\sigma = 50$ pixels, ensuring the model selects the most relevant points while maintaining diversity.

### Stage 3: Enhanced Image Generation

This stage learns to generate an enhanced image based on the source image and the mask containing the objects to be edited. We employed the Low-Rank Adaptation (LoRA) technique to fine-tune the Stable Diffusion v1-5 model for our 360° image enhancement task. Specifically, we applied LoRA to all attention modules within the U-Net, including each self-attention and cross-attention layer in the 12 ResNet blocks, totaling 48 attention layers. For each attention layer, LoRA was applied to its query, key, and value projection matrices. The weight matrix decomposition was performed using the formula:

$$W' = W + \alpha BA, \tag{8}$$

where $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the original weight matrix, $A \in \mathbb{R}^{r \times d_{\text{in}}}$, and $B \in \mathbb{R}^{d_{\text{out}} \times r}$, with $r$ being the low-rank parameter. We chose $r = 4$ and set the adjustable scaling factor $\alpha$ to 0.1. This approach significantly reduced the number of trainable parameters while maintaining the model's adaptability. By leveraging LoRA, we effectively adapted the model to the unique requirements of 360° image editing, ensuring computational efficiency and performance preservation.

Although the vanilla Stable Diffusion model (Rombach et al. 2022) is capable of editing specific objects within a mask while preserving the background unchanged, we observed suboptimal performance in certain scenes when removing objects. To address this issue, we designed a localized sensitivity reconstruction loss, $L_{\text{local}}$, and fine-tuned the model using our dataset. $L_{\text{local}}$ is defined as:

$$L_{\text{local}} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot |P_i(I_{\text{gen}}) - P_i(I_{\text{target}})| + \lambda \cdot R(w) \tag{9}$$

where $N = 256$ represents the block number when dividing the image, $P_i$ is the operation extracting a $32 \times 32$ pixel patch from the image, $I_{\text{gen}}$ and $I_{\text{target}}$ are the generated and target images respectively, $w_i$ is the weight assigned to each patch, and $R(w)$ is the entropy regularization of the weights with $\lambda = 0.01$ as the regularization strength. The weights $w_i$ are determined by calculating the Structural Similarity Index (SSIM) for each patch. We designed this localized sensitive loss function mainly because only the objects within the mask area are edited. By focusing on localized regions, we can perform finer adjustments within the mask-specified area, improving editing outcomes such as more thorough object removal and avoiding unnecessary alterations to the background regions.

## Experiment

### Datasets Establishment

**Training Datasets** Currently, there is no existing dataset for the scanpath-guided 360° image enhancement task. Inspired by SalG-GAN (Jiang et al. 2021), we constructed a dataset of 1000 360° image pairs to train our model. We first collected 1000 real-life 360° images covering various scenes as source images. For each source image, we used a scanpath prediction model to generate 1000 scanpaths and find a representative source scanpath using our proposed TASSC method. We then randomly selected two objects in each source image and manually created masks for them using Labelme. To create the corresponding target images, we removed these two masked objects from each source image. We then applied the same scanpath prediction model to generate 1000 scanpaths and a representative scanpath using TASSC for each target image. Our final training dataset consists of 1000 pairs, each containing a source 360° image with its representative source scanpath, and a target 360° image (created by removing two objects from the source image) with its representative target scanpath. This dataset enables our model to learn the relationship between object removal in 360° images and the resulting changes in scanpaths, facilitating scanpath-guided image enhancement.

Notably, we chose to use scanpath prediction model to generate possible scanpath rather than real data for training for two main reasons: First, acquiring a sufficiently large scale of real human scanpath data in VR environments would require extensive eye-tracking experiments, which is a significant challenge in terms of time and resources. Secondly, even if real experiments were conducted, the amount of data obtained might not provide sufficient diversity or volume to train a complex deep learning model. When enough real data becomes available in the future, our model architecture and design approach can be directly applied to those data.

**Test Set** To evaluate the performance and generalizability of our model, we constructed a test set of 200 360° images, which includes: (i) 200 source images captured by us, covering various scenes. (ii) Generated source scanpaths using the same method for the training set. (iii) Each image is randomly assigned a target scanpath provided by users.

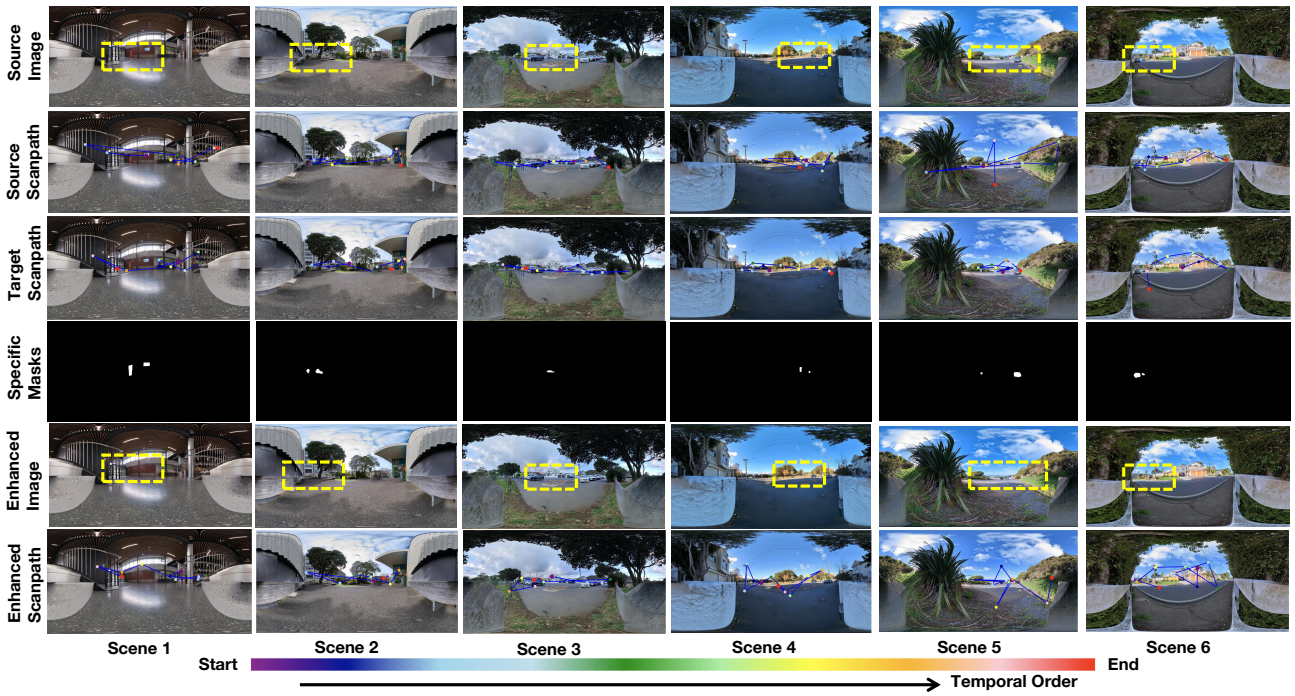


Figure 3: Experimental Results

## Evaluation Metrics

We quantitatively evaluate the performance of the clustering algorithm TASSC and the effectiveness of PSEM in guiding visual attention. To evaluate the performance of the TASSC clustering algorithm, we use three clustering quality metrics, including the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index (Arbelaitz et al. 2013). The Silhouette Coefficient ranges from -1 to 1, where higher values indicating better clustering quality. The Calinski-Harabasz Index measures the ratio of cluster separation to compactness, with higher values indicating better clustering. The Davies-Bouldin Index measures cluster separation, with lower values indicating better clustering quality.

To evaluate the effectiveness of our PSEM in guiding visual attention, we use the same external scanpath prediction model to generate 1000 scanpaths for both the source images and the enhanced images. We then measure the similarity of these scanpaths to the target scanpaths using three metrics: Levenshtein Distance (LEV), Dynamic Time Warping (DTW), and Recurrence Quantification Analysis (REC) (Wang, Zhang, and Dodgson 2024). Smaller LEV and DTW value indicates greater similarity, while a higher REC value suggests more structured and regular patterns within sequences (Martin et al. 2022). We calculate the average similarity of the 1000 predicted scanpaths from both the source and enhanced images to the target scanpaths, as well as the similarity of the clustered representative scanpaths to the target scanpaths. By comparing these metrics, if the scanpaths of the enhanced images show significantly higher similarity to the target scanpaths than those of the source images, it demonstrates that our method, PSEM, successfully guides users' visual attention. Additionally, we conducted qualitative evaluations by visually presenting the source scanpaths, the target scanpath, and the scanpaths of the enhanced images.

## Training Details

The second and third stages are separately trained. The entire training process was carried out on two NVIDIA A100 GPUs. When training the second stage, all parameters of SAM were frozen, and we only updated the parameters of the DSDE and Attention modules. We used the Adam optimizer with a gradually decreasing learning rate from 1e-4 to 1e-6. The training lasted for 100 epochs with a batch size of 32. During the fine-tuning of the SD v1-5 model, we only updated the parameters of matrices $A$ and $B$, keeping the original model weights $W$ unchanged. We used the Adam optimizer with a global learning rate ($\eta$) set to 1e-4, and designed specific learning rate scaling factors ($\beta_i$) for each layer. Max-norm regularization was applied, with $max_norm$ set to 1, to enhance training stability. This training lasted for 50 epochs with a batch size of 16.

## Experimental Results

**Quantitative Evaluation of TASSC** As shown in Table 1, the Silhouette Coefficient of 0.47 indicates good clustering, with data points well-grouped within clusters and separated from others. The Calinski-Harabasz index of 1345.92 shows high intra-cluster cohesion and clear inter-cluster separation. The Davies-Bouldin index of 0.23 reflects high intra-cluster similarity and distinct inter-cluster differences. Overall, these metrics demonstrate that the TASSC algorithm effectively groups similar scanpaths while maintaining distinctions between groups, providing a reliable foundation for

analyzing visual attention in 360° images.

| | Sil. Coef. | Cal-Har. Index | Dav-Bou. Index |
|---|---|---|---|
| **Value** | 0.47 | 1345.92 | 0.23 |

Table 1: Quantitative Evaluation of TASSC Algorithm

**Quantitative Results of Visual Guidance** Table 2 shows that the 200 enhanced images exhibit significant improvements across all metrics: LEV decreased by an average of 10.98%, DTW decreased by 6.78%, and REC increased by 17.84%. This indicates that the scanpaths produced by the enhanced images are closer to the target scanpath in terms of sequence structure, spatiotemporal features, and repetitive patterns. The improvement is even more pronounced when using clustered representative scanpaths for comparison, with the REC metric increasing by 42.19%. They strongly demonstrate the effectiveness of our method in guiding visual attention, successfully aligning the scanpaths of the enhanced images with the desired target scanpaths.

**Qualitative Results of Visual Guidance** Figure 3 demonstrates the effectiveness of our method across six scenes in the test set. (Scene 1) The source scanpath indicates that the user's visual attention moves from right to left, while the target scanpath aims for the user to start near the billboard on the right side before looking to the left. Our method selects two objects for editing, resulting in the scanpath of enhanced image more closely resembles the target scanpath, effectively guiding the user's attention from right to left. (Scene 2) The source scanpath shows that the user's gaze is concentrated on the left side, but our target scanpath aims for the user to focus more on the right. The model identifies that the presence of a person and a bicycle causes the left-side focus. The scanpath of the enhanced image indicates that removing the person and the bicycle results in the user's gaze being more distributed on the right side. Similarly, the enhanced image scanpaths for (Scene 3), (Scene 4), and (Scene 6) effectively guide the user's visual attention to follow the target scanpaths, which have a gaze point movement order opposite to that of the source scanpaths. (Scene 5) The source scanpath is dispersed across both sides of the scene, but the goal is to focus the user's visual attention on the central area. In the enhanced image, the scanpath shows a higher concentration of gaze points in the central area.

## Ablation Study

**Effectiveness of Fine-tuning** The Stable Diffusion v1-5 model supports inpainting using masks as conditions. To demonstrate the effectiveness of our fine-tuned SD model, we qualitatively compared its output images with those from the vanilla SD model. In terms of image consistency, the fine-tuned SD model generates more coherent results. For example, in Scene 1, the details of the shrub appear more natural, while in Scene 2, the door blends seamlessly with the background. Furthermore, the fine-tuned SD model avoids introducing unrealistic elements (Scene 3) and is more effective in removing objects (Scene 4), whereas the vanilla SD model often retains remnants of the objects.

**Impact of Inpainting Models** To evaluate the impact of different inpainting models on scanpath modification, we replaced the current model with diffusion-based approaches, Inpainting Anything (IA) (Yu et al. 2023) and LaMa (Suvorov et al. 2022). Our method improved scanpath similarity by 11.87%, compared to 10.75% for IA and 12.04% for LaMa. These results suggest that while inpainting methods affect image details, their influence on our method's ability to guide visual attention is minimal. As long as the model selects objects that significantly impact the scanpath, any modern inpainting model can be effective.

| Metric | Source Images | | Enhanced Images | | Improvement (%) | |
|---|---|---|---|---|---|---|
| | Average | Clustered | Average | Clustered | Average | Clustered |
| **LEV** ↓ | 109.24 | 105.10 | 97.25 | 85.47 | **10.98%** | **18.68%** |
| **DTW** ↓ | 4212.25 | 3904.91 | 3926.58 | 3668.45 | **6.78%** | **6.05%** |
| **REC** ↑ | 1.85 | 1.92 | 2.18 | 2.73 | **17.84%** | **42.19%** |

Table 2: Scanpath Similarity Comparison: Source Image and Enhanced Image vs. Target

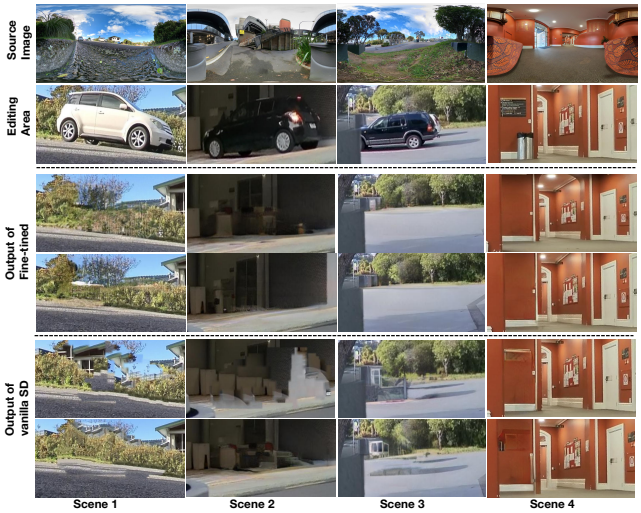


Figure 4: Results Generated by the Fine-tuning SD.

## Conclusion

We propose a target scanpath-guided 360° image enhancement task and introduce the Progressive Scanpath-Guided Enhancement Model (PSEM) to address it. Our method extracts a representative scanpath from diverse source scanpaths, models the differences between source and target visual attention, and identifies objects for editing. Finally, a fine-tuned Stable Diffusion model generates an enhanced 360° image aligned with the target scanpath. Experimental results validate the effectiveness of our approach.

Currently, our model is limited to modifying scanpaths by removing specific objects and cannot perform other pixel-level edits (e.g., enhancing contrast or adjusting colors) or object-level edits based on scanpath differences. This limitation stems from the small size of our manually annotated training dataset, which could be mitigated by creating a larger and more diverse dataset. Future work will focus on optimizing the editing stage to account for the spherical properties of 360° images and conducting real experiments with participants to expand the dataset with real human data.

## Acknowledgements

This work was supported by the Marsden Fund Council managed by the Royal Society of New Zealand (No. MFP-20-VUW-180).

## References

Aberman, K.; He, J.; Gandelsman, Y.; Mosseri, I.; Jacobs, D. E.; Kohlhoff, K.; and Rubinstein, M. 2022. Deep saliency prior for reducing visual distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19851–19860. IEEE.

Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J. M.; and Perona, I. 2013. An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recognition*, 46(1): 243–256.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402. IEEE.

Deng, B.; Song, S.; French, A. P.; Schluppeck, D.; and Pound, M. P. 2024. Advancing Saliency Ranking with Human Fixations: Dataset Models and Benchmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28348–28357.

Froese, V.; Jain, B.; Rymar, M.; and Weller, M. 2023. Fast Exact Dynamic Time Warping on Run-Length Encoded Time Series. *Algorithmica*, 85(2): 492–508.

Gava, C.; Mukunda, V.; Habtegebrial, T.; Raue, F.; Palacio, S.; and Dengel, A. 2023. SphereGlue: Learning Keypoint Matching on High Resolution Spherical Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6134–6144. IEEE/CVF.

Hessels, R. S.; Niehorster, D. C.; Kemner, C.; and Hooge, I. T. 2017. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods*, 49: 1802–1823.

Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; and Van de Weijer, J. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. OUP Oxford.

Hu, B.; Tunison, P.; RichardWebster, B.; and Hoogs, A. 2023. Xaitk-saliency: An open source explainable ai toolkit for saliency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15760–15766.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Jiang, L.; Xu, M.; Wang, X.; and Sigal, L. 2021. Saliency-Guided Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16509–16518.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.-Y.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026. IEEE/CVF.

Li, Y.; Barnes, C.; Huang, K.; and Zhang, F.-L. 2022. Deep 360° Optical Flow Estimation Based on Multi-Projection Fusion. In *European Conference on Computer Vision*.

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.

Mejjati, Y. A.; Gomez, C. F.; Kim, K. I.; Shechtman, E.; and Bylinskii, Z. 2020. Look here! a parametric learning based approach to redirect visual attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, 343–361. Springer International Publishing.

Miangoleh, S. M. H.; Bylinskii, Z.; Kee, E.; Shechtman, E.; and Aksoy, Y. 2023. Realistic Saliency Guided Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 186–194.

Murtagh, F.; and Contreras, P. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1): 86–97.

Petitjean, F.; Ketterlin, A.; and Gancarski, P. 2011. A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering. *Pattern Recognition*, 44(3): 678–693.

Qiao, M.; Xu, M.; Wang, Z.; and Borji, A. 2020. Viewport-dependent saliency prediction in 360 video. *IEEE Transactions on Multimedia*, 23: 748–760.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. IEEE/CVF.

Salvucci, D. D.; and Goldberg, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, 71–78.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2023. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *arXiv preprint arXiv:2310.06313*. arXiv.

Sui, X.; Fang, Y.; Zhu, H.; Wang, S.; and Wang, Z. 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*, 6989–6999. IEEE/CVF.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with Fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2149–2159.

Wang, Y.; Zhang, F.-L.; and Dodgson, N. A. 2024. ScanTD: 360° Scanpath Prediction based on Time-Series Diffusion. In *ACM Multimedia 2024*.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint Anything: Segment Anything Meets Image Inpainting. *arXiv preprint arXiv:2304.06790*.

Zhang, F.-L.; Zhao, J.; Zhang, Y.; and Zollmann, S. 2023. A Survey on 360° Images and Videos in Mixed Reality: Algorithms and Applications.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847. IEEE.

Zhang, Z.; Xu, Y.; Yu, J.; and Gao, S. 2018. Saliency detection in 360 videos. In *Proceedings of the European conference on computer vision (ECCV)*, 488–503.

Zhou, H.; Qiao, B.; Yang, L.; Lai, J.; and Xie, X. 2023. Texture-guided saliency distilling for unsupervised salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7257–7267.