SGST-Transformer: A Spherical Geometry-Aware Spatio-Temporal Transformer for 360° Video Saliency Prediction

Supplementary Material

1. Comparison of Different Projection Strategies

To verify the effectiveness of the proposed adaptive spherical patch projection mechanism, we compared it with two commonly used projection strategies, namely ERP and cubemap projection. Since both ERP and cubemap are conventional planar image representations, the projected images were partitioned into patch sequences in the same manner as in ViT, that is, each projected image was evenly divided into a set of image patches before being fed into the network. The three projection strategies shared the same backbone architecture and training settings, and differed only in the input projection format.

As shown in Table 4, the choice of projection strategy has a clear impact on performance. ERP projection performs better than cubemap projection, with AUC-J improved by 3.5% and NSS improved by 15.3%. A likely reason is that although cubemap projection reduces distortion within each individual face, the discontinuities between adjacent faces introduce semantic fragmentation at patch boundaries, which limits overall performance.

Table 4. Comparison of different projection strategies on the AVS-ODV dataset.

Model	AUC-J ↑	NSS ↑	KLD ↓	SIM ↑	CC ↑
Baseline (ERP)	0.906	2.750	2.733	0.354	0.414
Baseline (Cube)	0.875	2.385	3.098	0.327	0.330
Baseline (Ours)	**0.919**	**2.953**	**1.859**	**0.379**	**0.483**

The proposed method outperforms both comparison strategies on all metrics. Compared with ERP, it improves AUC-J by 1.4%, NSS by 7.4%, and reduces KLD by 32.0%, while SIM and CC increase by 7.1% and 16.7%, respectively. This improvement can be attributed to two aspects. First, the adaptive partition mechanism adjusts the number of patches according to spherical geometry, which keeps the partitioning more consistent across different regions of the sphere. Second, the local ERP reprojection strategy avoids the distortion caused by projecting a large spherical area onto a single planar surface. These results indicate that the proposed projection design can effectively alleviate spherical distortion and help the model better capture the spatial structure of panoramic content.

2. Generalization Experiment

To evaluate the generalization ability of the proposed model, we conducted cross-dataset testing. Specifically, the model was trained on the larger AVS-ODV dataset and then directly tested on the SVGC-AVA and Sports360 datasets without any fine-tuning or parameter adjustment. This setting provides a direct way to examine whether the model overfits the distribution characteristics of a specific dataset.

The results are reported in Table 5. On SVGC-AVA, the model achieves an AUC-J of 0.919, an NSS of 2.990, and a KLD of 1.170. Compared with its original performance on SVGC-AVA, there is some degradation, but the gap remains within a reasonable range. On Sports360, the model reaches an AUC-J of 0.900, an NSS of 2.486, and a CC of 0.452. Since this dataset contains a large number of fast-moving targets, the result suggests that the temporal modeling module retains a certain degree of transferability across scenes. Overall, the model maintains stable predictive performance on both unseen datasets, which supports its generalization capability.

Table 5. Generalization results on the SVGC-AVA and Sports360 datasets.

Dataset	AUC-J ↑	NSS ↑	KLD ↓	SIM ↑	CC ↑
SVGC-AVA	0.919	2.990	1.170	0.418	0.604
Sports360	0.900	2.486	1.503	0.346	0.452

3. Analysis of Systematic Bias

Although the prior feature is fused with the reconstructed feature by channel concatenation rather than direct multiplication, the subsequent convolution layers may still assign an excessively large weight to the prior channel in order to reduce the training loss. This may introduce a systematic bias. Since the ground-truth saliency maps are not always strictly concentrated around the equator, valid salient regions located in polar areas may yield relatively low activation scores under the constraint of such a prior, as illustrated in Fig. 4. To examine whether the proposed panoramic prior module introduces such systematic error, we further analyzed the model behavior on a high-latitude subset.

For this purpose, a high-latitude subset was constructed from the AVS-ODV dataset. We measured the proportion of salient regions located in high-latitude areas, defined as re-

gions with latitude $< 30^\circ$ or $> 150^\circ$. To ensure that the subset contained a sufficient number of samples while reducing the influence of random variation, the threshold for the proportion of high-latitude salient regions was set to 15%.

Figure 4. Illustration of salient regions in high-latitude areas.

Table 6 presents the ablation results on this subset. The model with the equatorial prior outperforms the version without the prior on all metrics. In particular, AUC-J increases from 0.817 to 0.819, NSS rises from 2.309 to 2.35, and KLD decreases from 3.915 to 2.865, corresponding to a reduction of 26.8%. SIM and CC also show modest improvements. These results indicate that even in regions above 30° latitude or below 150° latitude, the equatorial prior does not suppress valid salient responses. Instead, it helps the model produce a saliency distribution that is closer to the ground truth.

Table 6. Ablation results on the high-latitude subset of the AVS-ODV dataset.

Model	AUC-J ↑	NSS ↑	KLD ↓	SIM ↑	CC ↑
Without prior	0.817	2.309	3.915	0.375	0.423
Equatorial prior	**0.819**	**2.350**	**2.865**	**0.383**	**0.450**

The substantial decrease in KLD further suggests that introducing the prior makes the predicted probability distribution more consistent with the annotated saliency map, without sacrificing accuracy in polar regions due to equatorial bias. Therefore, the proposed equatorial prior can effectively guide the model to focus on the attention distribution of the human visual system, avoiding systematic bias against non-equatorial regions.