Spherical Vision Transformers for Audio-Visual Saliency Prediction in 360° Videos

by

Mert Çökelek

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Science and Engineering

KOÇ ÜNİVERSİTESİ

September 12, 2023

Spherical Vision Transformers for Audio-Visual Saliency Prediction in 360° Videos

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Mert Çökelek

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Aykut ERDEM (Advisor)

_____

Assoc. Prof. Erkut ERDEM

_____

Prof. Yücel YEMEZ

Date: _____

To everyone who made this possible and never gave up on me.

ABSTRACT

Spherical Vision Transformers for Audio-Visual Saliency Prediction in 360° Videos

Mert Çökelek

Master of Science in Computer Science and Engineering

September 12, 2023

Saliency prediction aims to model human audio-visual attention mechanisms to highlight the perceptually important regions in the scenes. This problem was first addressed in the literature under three branches based on the scene characteristics: static (for images), dynamic (for videos), and audio-visual saliency prediction. Due to the growing interest in virtual reality (VR), omnidirectional videos (ODVs) that capture the full field-of-view have gained 360° saliency prediction importance in computer vision. However, predicting where humans look in 360° scenes presents novel challenges, including the representation of 360° scenes regarding spherical distortion, high resolution, and the limited amount of annotated data. This thesis proposes a novel vision-transformer-based saliency prediction model named **SalViT360** for omnidirectional videos. We introduce a spherical geometry-aware spatio-temporal self-attention mechanism among tangent image representations for effective omnidirectional video understanding. We present a consistency-based unsupervised regularization term for projection-based 360° dense-prediction models to reduce artefacts in the predictions after inverse projection. Our approach is the first to employ tangent images for undistorted omnidirectional saliency prediction. Lastly, we propose **SalViT360-AV** by extending our video saliency prediction model with audio-visual adapters to incorporate mono and spatial audio modalities for a unified 360° audio-visual saliency prediction model. Our experimental results on four ODV saliency datasets demonstrate the effectiveness of SalViT360 and SalViT360-AV compared to the state-of-the-art.

ÖZETÇE

360° Videolarda Görsel-İşitsel Belirginlik Tahmini için Küresel Görüntü Dönüştürücüleri

Mert Çökelek

Bilgisayar Mühendisliği, Yüksek Lisans

12 Eylül 2023

Belirginlik tahmini, sahnelerdeki algısal olarak önemli bölgeleri vurgulamak için insanın görsel-işitsel dikkat mekanizmalarını modellemeyi amaçlamaktadır. Literatürde bu sorun ilk olarak sahne özelliklerine göre: statik (resimler için), dinamik (videolar için) ve görsel-işitsel belirginlik tahmini olarak üç dal altında ele alınmıştır. Son zamanlarda, sanal gerçeklik (VR) teknolojilerine artan ilginin sonucunda, tam görüş alanını yakalayan çok yönlü videolar (ODV'ler), bilgisayar görüsünde 360° belirginlik tahminine önem kazanmıştır. Bununla birlikte, insanların 360° sahnelerde nereye baktığını tahmin etmek, 360° sahnelerin temsili, küresel bozulma, yüksek çözünürlük ve sınırlı miktarda etiketli veri de dahil olmak üzere yeni zorluklar sunar. Bu tezde, SalViT360 adı verilen çok yönlü videolar için yeni bir görüntü dönüştürücü tabanlı belirginlik tahmin modeli önerilmişir. 360° video anlayışı için teğet görüntü temsilleri arasında küresel geometriye duyarlı uzay-zamansal bir özdikkat mekanizması sunulmuştur. Geri projeksiyon sonrasında tahminlerdeki bozulmaları azaltmak amacıyla projeksiyon tabanlı 360° yoğun tahmin modelleri için tutarlılık bazlı denetimsiz bir kayıp fonksiyonu sunulmuştur. Bu yaklaşım, bozulmamış 360° belirginlik tahmini için teğet görüntüleri kullanan ilk yaklaşımdır. Son olarak, birleşik bir 360° görsel-işitsel belirginlik tahmin modeli için bir boyutlu ve uzamsal ses modalitelerini dahil etmek üzere video belirginliği tahmin modeli SalViT360, görsel-işitsel adaptörlerle genişletilerek SalViT360-AV sunulmuştur. Dört 360° belirginlik veri seti üzerindeki deneysel sonuçlarımız, SalViT360 ve SalViT360-AV'nin en son teknolojiyle karşılaştırıldığında etkinliğini göstermektedir.

ACKNOWLEDGMENTS

I want to thank my advisor, Assoc. Prof. Aykut Erdem for his continuous support and guidance for over four years. I also want to express my deepest gratitude, love, and respect for Assoc. Prof. Erkut Erdem, Dr. Nevrez İmamoğlu, and Dr. Çağrı Özçınar for their continuous support. It has been an amazing experience for me to work with them.

I would also like to thank Assoc. Prof. Erkut Erdem once more, and Prof. Yücel Yemez for agreeing to participate in my thesis committee and evaluating our work.

I want to thank Nevrez İmamoğlu again for everything he has done for me.

I will never forget my lovely friends in KUIS AI and AVG Lab.

I am also thankful for the financial support of KUIS AI Center and UNVEST.

I also want to thank my lovely sister Asya, mother Dilek, and father Gürsel for their continuous support.

Finally, I would like to thank NVIDIA for the 3090s.

TABLE OF CONTENTS

List of Tables **x**

List of Figures **xii**

Abbreviations **xv**

Chapter 1: Introduction **1**

1.1 Problem Definition and Motivation 1

1.2 Proposed Method and Contributions 3

1.3 Organization of the Thesis 5

Chapter 2: Background **6**

2.1 Overview of Saliency Prediction Models 6

2.2 Datasets 6

2.2.1 Omnidirectional Video Saliency Datasets 7

2.2.2 Omnidirectional Audio-Visual Saliency Datasets 7

2.3 First-Order Ambisonics Representation of Spatial Audio 8

2.4 Evaluation Metrics 9

Chapter 3: Related Work **11**

3.1 360° Static Saliency Prediction 11

3.2 360° Dynamic Saliency Prediction 12

3.3 Audio-Visual Saliency Prediction 14

3.3.1 2D Audio-Visual Saliency 14

3.3.2 360° Audio-Visual Saliency 15

Chapter 4: Spherical Vision Transformer for 360° Video Saliency Prediction **16**

4.1 Methodology . . . 16

4.1.1 Gnomonic Projection and Encoder . . . 17

4.1.2 Viewport Spatio-Temporal Attention . . . 17

4.1.3 Decoder . . . 19

4.1.4 Viewport Augmentation Consistency . . . 19

4.2 Experiments . . . 20

4.2.1 Data Pre-processing . . . 20

4.2.2 Architecture and Optimization Details . . . 20

4.2.3 Evaluation Metrics and Loss Functions . . . 21

4.2.4 Comparison with the State-of-the-art . . . 21

4.2.5 Experimental Analysis and Ablation Studies . . . 23

Chapter 5: Audio-Visual Saliency Prediction for 360° Videos **27**

5.1 Methodology . . . 27

5.1.1 Audio Backbones . . . 28

5.1.2 Audio-Visual Adapter . . . 29

5.2 Experiments . . . 29

5.2.1 Data Pre-processing . . . 30

5.2.2 Comparison with the State-of-the-art . . . 30

5.2.3 Experimental Analysis and Ablation Studies . . . 31

Chapter 6: Conclusion **33**

Bibliography **35**

Appendix A: Additional Experiments on VSTA **43**

A.1 Temporal Window Size . . . 43

A.2 Transformer Depth . . . 43

A.2.1 Comparison with Joint Spatio-Temporal Attention . . . 44

Appendix B: Additional Experiments on VAC **45**

B.1 Approaches for Augmenting Tangent Images 45

B.1.1 Shifting Viewport Centers 45

B.1.2 FOV Augmentation 45

B.1.3 Viewport Augmentation 46

B.1.4 Mask-weighted VAC Loss. 47

Appendix C: Use Case: Saliency-Guided Omnidirectional Video Quality Assessment **48**

Appendix D: Discussion on SalViT360-AV Results **50**

D.1 Marginal performance gain from ambisonics over mono modality. . . . 50

D.2 Additional Experiments on Adapter Scaling Factor 52

LIST OF TABLES

4.1 **Performance analysis of *SalViT360* against the state-of-the-art 360° saliency models** on VR-EyeTracking, PVS-HMEM, 360AV-HM datasets. While the scores in **bold** highlight the best performance, the underlined ones are the second best. 23

4.2 **Ablation study** for each component in our approach on the validation split of the VR-EyeTracking dataset. 24

4.3 **Spatio-temporal modelling performance** of *SalViT360* compared against two alternative approaches on the validation split of the VR-EyeTracking dataset. 24

5.1 **Performance analysis of *SalViT360-AV* against state-of-the-art** on 360AV-HM and YT360-EyeTracking datasets. While † represents audio models with ambisonics modality, * corresponds to the ones with mono. **Bold** highlight the best scores, and the underlined ones are the second best. 30

5.2 **Comparison of our video baseline SalViT360 on the ground truth sets** for mute, mono, and ambisonics modalities. The best scores are highlighted in **bold**. 31

5.3 **Comparison of SalViT360-AVconfigurations** using PaSST and EZ-VSL audio backbones and trained with mono and ambisonics. The best scores are highlighted in **bold**. 32

A.1 **Quantitative comparison** for space-only attention, the proposed Viewport Spatio-Temporal Attention and existing Joint Spatio-Temporal attention for 360° saliency prediction on the validation split of VR-EyeTracking dataset. Due to memory limitations, JSTA model could not be trained on our GPU. 44

B.1 **Quantitative comparison** for the proposed VAC Loss with and without mask, on the validation split of VR-EyeTracking dataset. . . 47

C.1 **Comparison with state-of-the-art** saliency models for *saliency-guided omnidirectional video quality assessment* on VQA-ODV [Li et al., 2018a] dataset, as a downstream task. 48

LIST OF FIGURES

4.1 **Overview of the proposed SalViT360 model.** The ERP video clip of F frames (1) is projected to $F \times T$ tangent images per set (2). Each tangent image is encoded and fused with spherical-geometry-aware position embeddings (3) for the 360° video transformer to aggregate global information (4). The outputs are decoded into saliency predictions in tangent space (5), which are projected back to ERP, giving the final saliency map (6). In addition to the supervised loss, the model is trainable with $\mathcal{L}_{VAC}(P,\ P')$ to minimize the tangent artefacts (7). During test time, the model works with a single tangent set. For simplicity, only one set of tangent images is shown. . . . 16

4.2 **The proposed Viewport Spatio-Temporal Attention (VSTA)** (*right*), as compared to Viewport Spatial Attention (VSA) (*left*) and Joint Spatio-Temporal Attention (*middle*). Red and Green viewports denote the self-attention neighborhood for each scheme. 18

4.3 **Qualitative comparison on VR-EyeTracking [Xu et al., 2018b], PVS-HM [Xu et al., 2018a], and 360AV-HM [Chao et al., 2020] datasets.** Our proposed approach gives better results compared to existing models. The saliency predictions of our model better resemble the ground truth fixation density maps, producing sparse estimates while covering the most dominant modes better. 22

4.4 **Qualitative comparison** for our *VSTA baseline (third row)*, with the proposed *VAC loss (fourth row)*, and the optional *late-fusion (last row)*, compared to the ground truth. 25

5.1 **System overview of the proposed SalViT360-AV pipeline.** SalViT360-AV employs a two-stream network for video and audio modalities. The video stream is pre-trained and frozen SalViT360 upgraded with learnable adapter layers in the transformer blocks for audio-visual fine-tuning. The audio stream encodes rotated ambisonics for each tangent viewport, and the Adapter layers use these features for audio-visual fine-tuning. 28

5.2 **Proposed adapter fine-tuning for audio-visual modalities.** For each of the N transformer blocks in the SalViT360 pipeline, the visual tokens after Viewport Spatio-Temporal Attention are concatenated with audio tokens. The pre-trained MLPs at the end of each VSTA block are upgraded with *Audio-Visual MLP Adapters*, which consist of a scaled combination of a frozen MLP (left) and trainable bottleneck module (right). F, T, D denote the number of frames and tangent viewports tokens, clip length, and the embedding dimension, respectively. 29

A.1 Performance of our saliency prediction model on four evaluation metrics as a function of temporal window size (F) on the validation split of VR-EyeTracking [Xu et al., 2018b] dataset. 43

A.2 Performance of our saliency prediction model on four evaluation metrics as a function of VSTA depth (N) on the validation split of VR-EyeTracking dataset. 44

B.1 A tangent viewport from the original projection, highlighted on Equirectangular Projection (ERP) *(top)*, compared with three augmentation methods *(bottom)*. 46

B.2 Performance of our VSTA baseline compared with three augmentation consistency methods on four metrics, on the validation split of VR-EyeTracking dataset. 46

B.3 **Weight mask used for VAC Loss.** Each pixel coordinate in ERP takes a value based on the number of tangent viewports it is projected onto (max. 4). Brighter colours represent increasing overlaps. 47

D.1 **t-SNE visualization of ambisonics features**, encoded by PaSST model. Each data point corresponds to a tangent viewport waveform obtained by ambisonics rotation. Each small cluster consists of four randomly rotated waveforms for a video, and each colour represents audio categories in YT360-EyeTracking dataset. 50

D.2 CC score of SalViT360-AV on YT360-EyeTracking mono dataset as a function of scale values. 52

D.3 CC scores of SalViT360-AV on VREyeTracking (mono) dataset as a function of scale values for two audio-visual scenarios (match, mismatch). 52

ABBREVIATIONS

FOV	Field-of-view
VR	Virtual Reality
ERP	Equirectangular Projection
ODV	Omnidirectional Video
1D	One-Dimensional
2D	Two-Dimensional
FFN	Feed-forward Network
CNN	Convolutional Neural Network
ViT	Vision Transformer
VSA	Viewport Spatial Attention
VTA	Viewport Temporal Attention
VSTA	Viewport Spatio-Temporal Attention
VAC	Viewport Augmentation Consistency
KLD	Kullback-Leibler Divergence
NSS	Normalized Scanpath Saliency
CC	Correlation Coefficient
SIM	Similarity Metric
PSNR	Peak Signal-to-Noise Ratio
WS-PSNR	Weighted Spherical PSNR
t-SNE	t-distributed Stochastic Neighbor Embedding

Chapter 1

INTRODUCTION

1.1 Problem Definition and Motivation

While humans perceive numerous visual stimuli daily, our visual attention system filters out the irrelevant parts and processes only a tiny fraction of our environment. These prioritized regions are referred to as *salient regions*, and understanding how they are predicted is essential in computer vision and cognitive psychology. The problem of predicting where people fixate their eyes in a scene is referred to as *visual saliency prediction*, which is widely used in computer vision tasks, not only limited to recognition and detection, but also for *attention-guided* image and video compression, super-resolution, and quality assessment. For example, [Li et al., 2018b, Wang et al., 2020, Hou and Kung, 2022] use saliency maps to guide the super-resolution models for allocating more resources to reconstruct perceptually important regions. Similarly, [Zhu and Xu, 2018] used saliency maps to represent the regions of interest from the video signal for perceptual video coding, and [Zhu et al., 2020, Mishra et al., 2021] use saliency maps for reducing the perception redundancy for video compression. Moreover, [Guan et al., 2017, Yang et al., 2019, Zhu et al., 2021] use saliency maps as human perceptual cues for visual quality assessment.

With the growing popularity of virtual reality (VR) applications and multimedia streaming, predicting saliency in 360° scenes has received more attention recently. Notably, [Patney et al., 2016a, Patney et al., 2016b] use eye fixation and gaze information to perform foveated rendering in VR environments, such that the processing and delivery of high-resolution full field-of-view ODVs are optimized based on human perception. [Ozcinar et al., 2021] leverage visual saliency models for optimal bitrate allocation process for 360° video streaming. In addition to these use cases,

[Jabbari Tofighi et al., 2023] employ saliency-selective patches to estimate quality score for viewports on 360°, and [Qiu and Shao, 2021] use saliency-guided CNNs for blind 360° image quality assessment.

The aforementioned use cases demonstrate the application areas and importance of saliency prediction, which has been studied for over thirty years. Early saliency models were based on low-level features like colour, intensity, and orientation contrast. Later, with the advancements in deep learning, more complex features that could be grouped into three semantic categories, namely, spatial, temporal, and auditory modalities, have been studied. Spatial features provide information about the appearance of objects and scenes and temporal features supply information regarding motion characteristics. Prior works use deep learning to incorporate temporal cues by including optical flow, temporal recurrence at the feature level, and spatio-temporal features. While spatio-temporal cues are crucial for visual saliency prediction, discarding audio cues contradicts real-life scenarios where humans perceive audio and visual stimuli. In this line of work, psychological and computational studies [Perrott et al., 1990, Vroomen and Gelder, 2000, Vatakis and Spence, 2007, Chen et al., 2014, Min et al., 2014] demonstrate the importance of audio in guiding human visual attention. From the computational perspective, recent works on saliency prediction [Tavakoli et al., 2019, Min et al., 2020, Tsiami et al., 2020, Chen et al., 2021, Chao et al., 2020] show that visual saliency is partly correlated with audio sources and semantics.

360° video saliency prediction requires a new dimension of complexity, which involves not only spatial characteristics but also temporal dynamics and audio information for fully capturing the immersive experience in VR environments. One primary challenge in processing 360° scenes is effectively representing omnidirectional data. Equirectangular Projection (ERP), where the whole FOV scene is projected on a 2D plane, is a typical representation due to its computational simplicity. However, ERP suffers from spherical distortion, particularly towards the poles, which can significantly affect the geometric structure of the scene and degrade model performance. While previous works have proposed kernel transformations [Su and Grauman, 2017] and spherical convolutions [Coors et al., 2018, Esteves et al., 2018] to minimize this

distortion on ERP, these methods come at the cost of computational complexity and the loss of global context in 360° scenes. Cubemap projection [Greene, 1986] is another common approach that addresses the distortion problem by approximately expressing the spherical scene with six faces of a cube. Although this approach eliminates the distortion to an extent, it breaks the continuity of neighbouring faces and introduces discrepancies in the predictions around the edges. Moreover, these approaches are limited to processing 360° images, as the temporal information is ignored. With **SalViT360**, we propose an efficient vision transformer for 360° videos, which employs gnomonic projection [Eder et al., 2020] to obtain undistorted tangent viewport tokens in both spatial and temporal domains.

The secondary challenge in 360° video saliency prediction is to mimic the perception of 360° scene, as it does not occur instantaneously but rather begins within a limited FOV. Subsequently, people continuously expand their awareness with their head and eye movements to explore the peripheral regions. In contrast, computer systems entirely represent 360° video sequences, resulting in a disparity in perceptual behaviour between machines and humans. In prior work, [Chao et al., 2020, Cokelek et al., 2021] claim that the spatial audio cues, which provide directional information in 360° space can be incorporated alongside the visual cues to supply additional insight for modelling human visual attention. Motivated by their findings, we propose **SalViT360-AV**, which incorporates mono and spatial audio information for our transformer-based video saliency model using adapter-based tuning.

1.2 Proposed Method and Contributions

In **SalViT360**, we use gnomonic projection to obtain multiple undistorted tangent images, which enables us to extract rich local spatial features using any pre-trained and fixed 2D backbone. Our extensive experiments demonstrate the effectiveness of our proposed SalViT360 model against the state-of-the-art on VR-EyeTracking [Xu et al., 2018b], PVS-HMEM [Xu et al., 2018a], and 360AV-HM [Chao et al., 2020] datasets.

With **SalViT360-AV**, we propose a parameter-efficient audio-visual saliency

prediction model employing transformer adapters between two pre-trained and frozen backbones for the audio and video modalities. While the video backbone of SalViT360-AV is SalViT360, the audio backbone can be any off-the-shelf audio encoder. We revise the recent AdaptFormer [Chen et al., 2022] tuning method for the 360° audio-visual domain. Our experiments yield superior results to the state-of-the-art and our video baseline on VR-EyeTracking, 360AV-HM, and YT360-EyeTracking datasets. Our contributions with SalViT360 are three-fold:

- *Spherical geometry-aware spatial attention.* We aggregate global information on the sphere by computing self-attention among tangent viewports. We use tangent features as viewport tokens to address the quadratic complexity associated with spatial self-attention. We introduce 360° geometry awareness to the transformer by using learnable spherical position embeddings guided by per-pixel angular coordinates (ϕ, θ), denoting *latitude* and *longitude* angles, respectively. The proposed position embedding method outperforms standard 1D embeddings and can easily be integrated into any transformer architecture designed for 360° images.

- *Viewport Spatio-Temporal Attention (VSTA).* The complexity of joint spatio-temporal self-attention increases quadratically concerning the number of frames. With VSTA, we optimize this joint computation on tangent viewports from consecutive frames, where the spatio-temporal self-attention is performed in two stages: (1) Viewport Spatial Attention (VSA) and (2) Viewport Temporal Attention (VTA). In VSA, spherical geometry-aware self-attention is computed *intra-frame* level. VTA encodes the temporal information in the videos among tangent planes that point to the same direction in the *inter-frame* level.

- *Viewport Augmentation Consistency (VAC).* We propose an unsupervised, consistency-based loss for omnidirectional images, minimizing the discrepancies in overlapping regions of tangent predictions. The loss is computed between the weight-sharing saliency predictions of two tangent image sets, which are generated with different configurations. This regularization method is suit-

able for any projection-based dense-prediction model. Importantly, VAC does not introduce any memory or time overhead during test time, as only one set of predictions is sufficient for inference.

- *Spherical geometry-guided audio-visual transformer adapter.* Since 360° ODV datasets contain mono or spatial audio, the audio-visual model is expected to work with both audio modalities. Particularly for mono audio, to provide spherical position information to adapters, as in SalViT360, we augment the waveform features with spherical coordinate-based position embeddings. Our experiments demonstrate that the proposed augmentation yields better performance qualitatively and quantitatively.

1.3 Organization of the Thesis

The structure of the thesis is organized as follows:

Chapter 2 introduces background information regarding the history of saliency prediction, datasets, and evaluation metrics.

Chapter 3 discusses the related work on 360° (audio-)visual saliency prediction.

Chapter 4 describes our SalViT360 model for saliency prediction in 360° videos.

Chapter 5 introduces our SalViT360-AV for extending the video saliency model with spatial audio modality.

Chapter 6 concludes the thesis and presents possible future research directions.

Chapter 2

BACKGROUND

This chapter provides a concise review of audio-visual saliency prediction models for 2D images/videos. We excluded 360° saliency models from this review, as the extensive analysis of these models is presented in Chapter 3. We also introduce the datasets and evaluation metrics used to assess the saliency models.

2.1 Overview of Saliency Prediction Models

As mentioned in Chapter 1, the goal of saliency models is to predict the eye fixations of humans exploring visual stimuli by producing saliency maps resembling ground-truth maps. The ground-truth annotations are collected from subjects by eye-tracking hardware and represented as fixation and saliency maps. Fixation maps contain the actual eye gaze data of humans. Fixations are binary (discrete) for each pixel in the visual scene, corresponding to whether or not an observer's gaze has fixated on that location. Saliency maps are heat maps generated by convolving the fixation maps with a Gaussian kernel to create a continuous probability distribution for all regions in a scene. Saliency maps represent the continuous measure of saliency or likelihood that a location attracts human attention.

2.2 Datasets

While extensive research has been conducted in traditional 2D saliency prediction, the number of studies in 360° videos and annotated datasets is relatively limited. As a result of this constraint, we have used the three publicly available ODV saliency datasets. Addressing this limitation, in a concurrent work, we have also collected a new dataset of 360° videos under various audio-visual conditions and semantic categories. We provide the details regarding the datasets used in the following

subsections.

2.2.1 Omnidirectional Video Saliency Datasets

VR-EyeTracking [Xu et al., 2018b]. For training our models, we use the VR-EyeTracking dataset. It consists of 134 train and 74 test videos lasting between $20 - 60$ seconds. The videos have a minimum resolution of $4K$ in ERP format with frame rates between $24 - 60$fps. There are 20 female and 25 male subjects aged between $20 - 24$ years. While at least 31 subjects view each video, each subject has viewed around 35 videos. The videos are picked under various scene categories, including indoor, outdoor, sports, games, documentation, short movies and music shows. Subjects have viewed the videos with mono audio modality.

PVS-HMEM [Xu et al., 2018a]. The dataset contains head movement and eye fixation data from 76 panoramic video sequences viewed by 58 subjects. The videos vary between $10 - 80$ seconds, with a minimum resolution of $2K$ and frame rate between $24 - 60$fps. It includes indoor, outdoor, movie, and sports categories. The data collection is done in a muted modality.

2.2.2 Omnidirectional Audio-Visual Saliency Datasets

As mentioned in Chapter 1, audio plays an essential role in the perception of videos. Traditional 2D videos may contain either mono or stereoscopic audio. Mono modality delivers sound sources equally from all directions. Hence, the viewers can not be guided by directional sound sources. Stereo audio has two channels (left and right) to provide direction information of sound sources in only one axis. In contrast, 360° videos can be delivered with spatial audio, where the sound sources are precisely located, and the perceived sound changes with head movement (roll, pitch, yaw). [Chao et al., 2020] collected the first ODV saliency dataset with spatial audio modality, and their study demonstrates the impact of spatial audio in the perception of omnidirectional videos by guiding the subjects with directional sound. Chapter 2.3 introduces the spatial audio format used in these datasets.

360AV-HM [Chao et al., 2020]. The dataset contains 15 ODVs with spatial

audio lasting 25 seconds. Each video has $4K$ resolution in ERP format, with frame rates between $24 - 60$fps. The videos are collected from YouTube under three categories, namely, *Conversation, Music*, and *Environment*. The videos are then converted to mute and mono audio settings for three subject groups, where each subject has only viewed the videos with one modality. Fifteen subjects viewed each ODV, and there are 45 (16 female, 29 male) subjects, aged between $21 - 40$ years.

YT360-EyeTracking Dataset. We address the constraint related to the availability of public datasets for 360° audio-visual saliency prediction. In a concurrent work at Bogazici University, we picked 81 videos from YT-360 [Morgado et al., 2020] by carefully designing nine audio-visual semantic categories and equally distributing nine videos per category. The audio categories are (1) speech, (2) music, (3) vehicle, and visual categories are (1) indoors, (2) outdoors-natural, and (3) outdoors-man-made. Each video lasts 30 seconds with a frame rate of 30fps. Eye tracking data from a total of 45 subjects are collected under three audio modalities, where muted, mono and ambisonic modalities are viewed by 15 subjects separately. Like the 360AV-HM dataset, the ambisonics are encoded in first-order, 4-channel B-format.

2.3 First-Order Ambisonics Representation of Spatial Audio

The most commonly used spatial audio format in ODVs is first-order ambisonics in 4-channel B-format. The datasets we introduce (360AV-HM and YT360-EyeTracking) use this format. The four channels are named (W, X, Y, Z), each representing a different directionality in 360°: center (mono), front-back, left-right, and up-down. Channel W encodes audio from all directional equally, equivalent to mono modality. The remaining channels have different spherical harmonic components, which produce the polar patterns of figure-of-eight microphones. For instance, channel Y contains a single waveform with two (positive and negative) phases, encoding audio received from the left and right of the reference coordinate, respectively.

Rotation and Decoding

During recording, the ambisonic microphones are calibrated with the center of the video, referenced with $(\theta, \phi, \psi) = (0°, 0°, 0°)$. While playing the audio within head-mounted display devices (VR headsets), the audio waveforms are decoded based on the rotation of the head. In this way, the subjects can perceive sound coming from the direction they are looking at in real time.

2.4 Evaluation Metrics

We employ the four most commonly used saliency evaluation metrics: Normalized Scanpath Saliency (NSS), Kullback-Leibler Divergence (KLD), Pearson's Correlation Coefficient, and Similarity Metric (SIM). Each metric evaluates a different aspect of saliency predictions. While the first metric is fixation-based, the remaining ones are distribution-based metrics.

NSS is a fixation-based metric, which is computed as the average normalized saliency prediction at ground-truth fixations. Given a saliency map P and the ground-truth binary map of fixations F, NSS is computed as:

$$NSS(P, F) = \frac{1}{N} \sum_i \bar{P}_i \times F_i \tag{2.1}$$

$$\text{where } N = \sum_i F_i \text{ and } \bar{P}, = \frac{P - \mu(P)}{\sigma(P)}, \tag{2.2}$$

where i is the fixation pixel index, and N is the total number of fixations. Due to normalization, NSS is sensitive to both true positives and false positives. A higher NSS value indicates more significant correspondence between prediction and ground truth, and a negative value indicates anti-correspondence. In the case of no correspondence, the NSS value is calculated as zero.

KLD evaluates the dissimilarity between two distributions, which is computed as:

$$KLD(P, S) = \sum_i S_i \log\ (\epsilon + \frac{S_i}{\epsilon + P_i}), \tag{2.3}$$

where ϵ is a regularization constant, P and S are predicted and ground-truth saliency maps, respectively. Since KLD computes the dissimilarity, better predictions will have lower KLD values.

CC is another distribution-based metric that measures the linear correspondence between predicted P and ground-truth S saliency maps. CC is computed as:

$$CC(P, S) = \frac{\sigma(P, S)}{\sigma(P) \times \sigma(S)}, \tag{2.4}$$

where $\sigma(P, S)$ corresponds to the covariance of P and S. CC equally penalizes false positives and false negatives, and unlike KLD, it is a symmetric metric. Hence, CC cannot distinguish if disparities in the maps are from false positives or false negatives. A higher value of CC indicates better results.

SIM is another distribution-based metric that measures the intersection between predicted and ground-truth saliency maps. It computes the sum of minimum values at each pixel i by:

$$SIM(P, S) = \sum_i \min\ (P_i, S_i) \tag{2.5}$$

$$\text{where } \sum_i P_i = \sum_i S_i = 1. \tag{2.6}$$

SIM takes values between $0 - 1$. While a SIM value of zero indicates no overlap, a SIM of one indicates that the distributions are the same. [Bylinskii et al., 2018] discuss that CC and NSS make fewer assumptions on the input data. Moreover, since they equally penalize false positives and negatives, CC and NSS provide more reliable and fair results.

Chapter 3

RELATED WORK

Saliency prediction models can be divided into two categories based on the scene characteristic, namely, static and dynamic saliency prediction. Static saliency considers input in the image level and ignores temporal information between frames. Conversely, dynamic saliency includes both spatial and temporal cues in the videos. In this chapter, we provide a detailed literature review of 360° static and dynamic saliency models. Previous works in 360° saliency prediction primarily focused on addressing the representation of 360° scenes, with each method trying to balance representative power and computational complexity. Nevertheless, most works ignore the audio modality, which contains essential stimuli that affect the human visual attention system. In Chapter 3.3, we introduce the audio-visual saliency models for 2D and 360° domains.

3.1 360° Static Saliency Prediction

[Chao et al., 2018] proposed SalGAN360, which employs cubemap projection to minimize the distortion that occurs in ERP. Then, they fine-tune the 2D image saliency model SalGAN [Pan et al., 2017] on each cube face and project the cube predictions back to ERP to obtain the final prediction. Although the distortion is reduced in each face locally, cubemap projection suffers from discontinuities at the edges. CNN-based models cannot distinguish the neighbourhood among cube faces, which results in discrepancies in the predictions around edges. To solve the discrepancies, authors propose using multiple sets of cube maps, each containing horizontally (9×) and vertically (9×) shifted faces. Using 81 projections per ERP image results in a dramatic computational overhead. In the follow-up work, [Chao et al., 2020] proposed MV-SalGAN360, which extends SalGAN360 with multi-view fusion. Un-

like the single-scale cubemap projection, MV-SalGAN360 uses three scales. The first scale corresponds to the cubemap of six faces with 90° FOV each. The second scale consists of five faces, each with an FOV of 120°. The last scale is the original ERP, with 180° × 360° FOV. While the multi-view representation is important for local-global scene encoding, MV-SalGAN360 requires a separate set of parameters for each level, which is computationally expensive. [Zhang et al., 2018] proposed a spherical U-Net model for saliency prediction. The authors do not project the ERP input to other formats; instead, they define modified kernels on ERP based on spherical coordinates. The convolution kernel is defined on a spherical crown, which is stretched and rotated depending on the convolution location on the sphere. While the receptive field of the kernels is dynamic, there is a parameter-sharing mechanism for the kernels, which also makes the model equivariant to translation. Although the number of parameters is optimized ($6.7M$), the runtime complexity is high due to the separate computation of a spherical crown for each convolution. [Dahou et al., 2021] proposed ATSal, a two-stream architecture to compute global and local saliency in omnidirectional videos. They use global prediction as a rough attention estimate, and the local stream on cube faces predicts local saliency. The local predictions are projected back to ERP and linearly combined with the global map for the final saliency prediction. [Djilali et al., 2021] used a self-supervised pre-training based on learning the association between several different views of the same scene and trained a supervised decoder for 360° saliency prediction as a downstream task. Their framework maximizes the mutual information between different views on the spherical image, and the decoder is trained with these embeddings. Although their approach can model the global relationship between viewports, it ignores the crucial temporal dimension for video understanding.

3.2 360° Dynamic Saliency Prediction

[Cheng et al., 2018] propose cube-padding to eliminate the discontinuities on the face boundaries. Cube-padding transfers information among cube faces by introducing overlapping regions around face boundaries. Particularly for CNN models,

the information exchange increases with deeper layers. The model also utilizes ConvLSTMs for extracting temporal features. [Qiao et al., 2020] showed that the eye fixation distribution bias depends on the viewport locations. They also considered the 360° video saliency prediction as a multi-task problem, where the goal is to predict the saliency map over the spherical scene and local saliency inside viewports. The demonstrated bias of ground-truth fixations on viewport locations motivated us to introduce *spherical position information* into our model. [Yun et al., 2022] propose PAVER - Panoramic Vision Transformer, which uses deformable convolutions to handle distortion on spherical projection, where a separate offset is computed for each convolution. They then feed the local undistorted patch features to vision transformers to compute self-attention across space and time. The authors also demonstrate a use case of their video saliency model on an omnidirectional visual quality assessment task, using saliency maps as weights for the assessment. They compare their saliency predictions alongside other models with ground-truth saliency maps for assessing the quality of videos. We also include our SalViT360 model to this comparison in Appendix C.

The methods mentioned above share a standard limitation in processing 360° data through projections or modified kernels and ignore the whole field-of-view, which is critical for global scene understanding and saliency prediction. Thus, there is a need for an effective omnidirectional data processing method that minimizes spherical distortion while preserving the global context and avoiding artefacts introduced by projection. Recently, [Eder et al., 2020] proposed tangent image representations, which use gnomonic projection to map a spherical image into multiple overlapping patches, where each patch is tangent to the faces of an icosahedron. This method tackles the problem of spherical distortion on the scene. However, to our interest, the dense-prediction models particularly suffer from discrepancies and artefacts on overlapping regions of tangent image patches after inverse projection to ERP. In this work, we propose using tangent images to process undistorted local viewports and develop a transformer-based model to learn their global association for saliency prediction in 360° videos. This is the first work employing tangent images for omnidirectional saliency prediction, which can also model the

temporal dimension. The spatio-temporal attention mechanism is adapted to 360° videos from TimeSformer [Bertasius et al., 2021], which is proposed for traditional 2D videos. TimeSformer proposes Divided Space-Time Attention, which computes spatio-temporal attention in two stages to optimize for model and data complexity. In Chapter 4.1.2, we present Viewport Spatio-Temporal Attention, which extends Divided Space-Time Attention in the TimeSformer model to 360° videos.

3.3 Audio-Visual Saliency Prediction

As mentioned in Chapter 1, previous works [Vroomen and Gelder, 2000, Chen et al., 2014, Min et al., 2014], demonstrated the impact of audio modality on human attention systems. On the other hand, most of the visual saliency estimation models ignore audio. Only a few works showed that the visually salient sources are partly correlated with the audio sources and semantics and obtained performance gain from audio information on visual saliency prediction from the computational perspective.

3.3.1 2D Audio-Visual Saliency

[Tavakoli et al., 2019] proposed DAVE, a deep-learning-based audio-visual encoder framework for traditional 2D video saliency prediction. DAVE comprises a two-stream encoder architecture for audio and visual modalities. The video clips and audio mel-spectrograms are encoded by 3D ResNets in each stream, concatenated in the feature space, and then decoded with a 2D CNN to obtain the saliency predictions. [Min et al., 2020] propose a multi-modal saliency (MMS) model for videos with high audio-visual correspondence. They first detect visual saliency maps in spatio-temporal dimension using traditional methods and fuse audio and visual modalities by localizing moving sound sources in the scene. MMS is proposed as an audio-visual saliency bias for late-fusion with saliency maps and promotes the existing models‘ accuracy by an average of 5%. The MMS model assumes that the sound sources align with the visible and moving objects in a scene. As mentioned in the MMS paper, the model does not provide performance gain when this assumption does not hold. [Tsiami et al., 2020] proposed a spatio-temporal audio-visual saliency

(STAViS) model for 2D videos. Similar to DAVE, it employs a two-stream encoder for audio and visuals. Audio waveforms and video clips are encoded by 1D and 3D CNNs, respectively. They propose a bilinear audio localization module. They then concatenate the localization volume with audio and visual features and pass it to a decoder to obtain saliency predictions. [Chen et al., 2021] proposed a deep learning architecture for two-stream feature extraction, semantic interaction, and their fusion for auditory and visual inputs. Semantic interaction brings the audio-visual features from lower and higher levels to the same dimensionality, which demonstrates superior performance than fusing the features from only the final layers.

3.3.2 360° *Audio-Visual Saliency*

Spatial audio modality has been used for saliency prediction in two prior works [Chao et al., 2020, Cokelek et al., 2021]. [Chao et al., 2020] propose a two-stream pipeline for audio and visual modalities. In the visual pipeline, they project ERP frames to padded cube maps and pass them to a 3D ResNet. In the audio stream, they extract log mel-spectrograms for all four channels of first-order-ambisonics (FOA) and pass the spectrogram features to 3D ResNet. They then concatenate the audio-visual features and fuse the resulting feature map with audio energy maps (AEM). AEMs represent the direction of arrival of hearable sound on the sphere, weighted according to the amplitude of each direction. AEMs are obtained from FOA by calculating the energy of each channel (W, X, Y, Z) over time as squared sums in short time windows. These energy values are then spatially rendered onto ERP using angular coordinates. [Cokelek et al., 2021] proposed a 360° salient spatial sound localization (SSSL) method without including visual modalities. SSSL employs a mel-cepstrum-based spectral residual saliency (MCSR) module, which takes input directional audio waveforms and outputs saliency values for each sample in the time domain. They apply MCSR to six directions and rescale the saliency values to the unit vector on the sphere, which points to the direction of the salient sounds for each timestep. Finally, they project these predictions onto ERP saliency maps as audio saliency biases for any existing 360° (audio-)visual saliency model.

Chapter 4

SPHERICAL VISION TRANSFORMER FOR 360° VIDEO SALIENCY PREDICTION

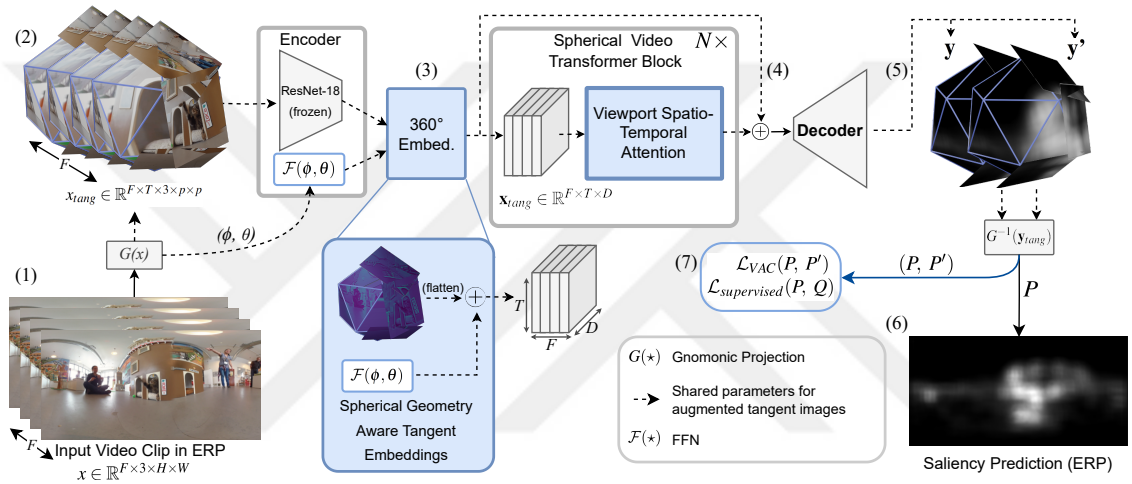

Figure 4.1: **Overview of the proposed SalViT360 model.** The ERP video clip of F frames (1) is projected to $F \times T$ tangent images per set (2). Each tangent image is encoded and fused with spherical-geometry-aware position embeddings (3) for the 360° video transformer to aggregate global information (4). The outputs are decoded into saliency predictions in tangent space (5), which are projected back to ERP, giving the final saliency map (6). In addition to the supervised loss, the model is trainable with $\mathcal{L}_{VAC}(P,\ P')$ to minimize the tangent artefacts (7). During test time, the model works with a single tangent set. For simplicity, only one set of tangent images is shown.

4.1 Methodology

In Fig. 4.1, we present an overview of our proposed model SalViT360. We start with gnomonic projection [Eder et al., 2020] to obtain tangent images for each frame in the input video clip. These are passed to an encoder-transformer-decoder architecture.

The image encoder extracts local features for each tangent viewport and reduces the input dimension for the subsequent self-attention stage. We map the pixel-wise angular coordinates to produce the proposed spherical geometry-aware position embeddings $\mathcal{F}(\phi, \theta)$ for the 360° transformer, enabling better learning of spatial representations. The transformer utilizes Viewport Spatio-Temporal Attention (VSTA) to capture inter and intra-frame global information across tangent viewports in a temporal window. The transformed embeddings are then fed into a 2D CNN-based decoder, which predicts saliency on the tangent images. We then apply inverse gnomonic projection on the tangent predictions to obtain the final saliency maps in ERP. We propose an unsupervised consistency-based Viewport Augmentation Consistency Loss to mitigate the discrepancies after inverse gnomonic projection. The learnable parameters of the network are in tangent space, allowing us to use large-scale pre-trained 2D models for feature extraction while the rest of the network is trained from scratch.

4.1.1 Gnomonic Projection and Encoder

We first project the input ERP clip $x \in \mathbb{R}^{F\times 3\times H\times W}$ to a set of tangent clips $x_{tang} \in \mathbb{R}^{F\times T\times 3\times p\times p}$, where F, 3, H, and W are the number of frames, channel dimension (RGB), height, and width of a given video, respectively. The resulting tangent images have a patch size of $p \times p = 224 \times 224$ pixels. Number of tangent images per frame T, and FOV are the projection hyperparameters which vary between 10/18 and 120°/80°. We downsample and flatten the encoder features to obtain tangent feature vectors with dimension $D = 512$. We map the angular coordinates (ϕ, θ) for each pixel of the tangent viewports to the same feature dimension using an FC layer and sum these embeddings with encoder features to obtain the proposed *spherical geometry-aware embeddings* $\mathbf{x}_{tang} \in \mathbb{R}^{F\times T\times D}$ that are used in the transformer.

4.1.2 Viewport Spatio-Temporal Attention

While the pre-trained encoder extracts rich spatial features for each tangent image locally, aggregating the global context in the full FOV is essential for 360° scene un-

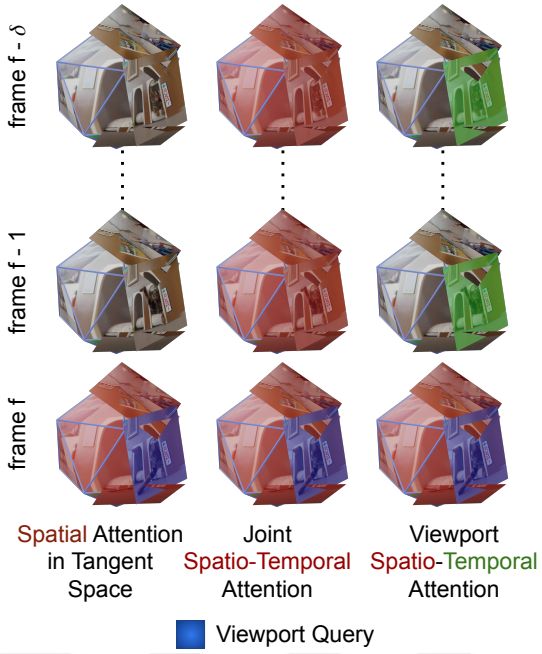

Figure 4.2: **The proposed Viewport Spatio-Temporal Attention (VSTA)** (*right*), as compared to Viewport Spatial Attention (VSA) (*left*) and Joint Spatio-Temporal Attention (*middle*). Red and Green viewports denote the self-attention neighborhood for each scheme.

derstanding. We propose a self-attention mechanism on tangent viewport features to achieve this. However, since incorporating the temporal dimension of the videos increases the number of tokens and thus the computational complexity, we approximate spatio-temporal attention with two stages: we apply temporal attention (1) among the same tangent viewports from consecutive F frames, then, spatial attention (2) among T tangent viewports in the same frame. This reduces the overall self-attention complexity from $F^2 \times T^2$ to $F^2 + T^2$. In this way, we effectively model the global context required for 360° video understanding. The proposed *Viewport Spatio-Temporal Attention* for 360° videos illustrated in Fig. 4.2 is expressed as:

$$\begin{aligned} \mathrm{VSTA}(\mathbf{z}^{(l)}_{(t,f)}) &= \mathrm{VSA}(\mathrm{VTA}(\mathbf{z}^{(l)}_{(t,f)})) \\ \mathrm{VTA}(\mathbf{z}^{(l)}_{(t,f)}) &= \mathrm{SM}\left(\mathbf{q}^{(l)}_{(t,f)} \cdot \left\{\mathbf{k}^{(l)\top}_{(t,f')}\right\}\right) \cdot \left\{\mathbf{v}^{(l)}_{(t,f')}\right\}_{f'=1\ldots F} \\ \mathrm{VSA}(\mathbf{z}^{(l)}_{(t,f)}) &= \mathrm{SM}\left(\mathbf{q}^{(l)}_{(t,f)} \cdot \left\{\mathbf{k}^{(l)\top}_{(t',f)}\right\}\right) \cdot \left\{\mathbf{v}^{(l)}_{(t',f)}\right\}_{t'=1\ldots T} \end{aligned} \tag{4.1}$$

where $\mathbf{z}^{(l)}_{(t,f)} \in \mathbb{R}^{1\times D}$ corresponds to the tangent features ($\mathbf{x}_{tang}$) of viewport t in frame f at l-th transformer block, $\mathbf{q}, \mathbf{k}, \mathbf{v}$ are the query, key, and value projections

of $\mathbf{z}$, and SM is the softmax operator. Attention heads and the scale multiplication of dot-product attention are not given for presentation clarity.

4.1.3 Decoder

The decoder comprises four upsample layers followed by 3×3 convolutions and normalization layers. For a set of tangent clips, it takes the skip connection of encoder and transformer features $\mathbf{x}^{f}_{tang} \in \mathbb{R}^{T \times D \times 7 \times 7}$ of the last frame as input and outputs saliency prediction $\mathbf{y} \in \mathbb{R}^{T \times 56 \times 56}$ on tangent planes. The final ERP saliency maps are obtained by passing the tangent predictions to inverse gnomonic projection. While we aggregate global information among tangent images through the transformer, each tangent plane is predicted separately in the decoder. This causes discrepancies in the overlapping regions on ERP. Prior work on omnidirectional depth estimation [Li et al., 2022] proposes iterative refining to compensate for these artefacts, requiring the whole model to run twice per sample during training and inference. To address this, we propose *Viewport Augmentation Consistency*, an unsupervised loss strategy as a regularizer that requires zero parameters and has no time overhead. It is trained on multiple tangent scales in parallel and requires only one tangent set for inference.

4.1.4 Viewport Augmentation Consistency

Our model effectively learns the overall saliency distribution with the supervised loss. However, since each tangent plane is predicted separately, the final ERP saliency map contains discrepancies in overlapping regions of tangent patches. To tackle this issue, we propose an unsupervised loss strategy, called *Viewport Augmentation Consistency* (VAC), for improving the consistency between the saliency predictions P and P' from two tangent projection sets. Specifically, we generate the second tangent set by applying different configurations, such as horizontally shifting the tangent planes on the sphere, using a larger FOV for the same viewports, and varying the number of tangent images at different viewports. We provide a detailed comparison of these approaches in the supplementary. VAC does not require any

additional memory or time overhead since it uses the shared parameters of the whole model, and the forward pass is done in parallel. Furthermore, since the ERP predictions from the original P and augmented P' tangent sets are expected to be consistent, only one tangent set is sufficient for testing. The VAC loss is defined as:

$$\begin{aligned} \mathcal{L}_{VAC}(P,\ P') &= \mathcal{L}_{KLD}^{weighted}(P,\ P')\ +\ \mathcal{L}_{CC}^{weighted}(P,\ P'), \\ \mathcal{L}_{KLD}^{weighted}(P,\ P') &= \sum_{i,j} P_{i,j} \log\left(\epsilon + \frac{P_{i,j}}{P'_{i,j} + \epsilon}\right) \cdot w_{i,j}, \\ \mathcal{L}_{CC}^{weighted}(P,\ P') &= 1 - \frac{\sum(P \cdot P') \cdot w_{i,j}}{\sum(P \cdot P) \cdot\ \sum(P' \cdot P')} \end{aligned} \tag{4.2}$$

where $P,\ P'$ are the saliency predictions from original and augmented viewports, and w is an optional weight matrix obtained from gnomonic projection to weigh the overlapping pixels of gnomonic projection on ERP predictions. Details for viewport augmentation approaches and the weighting operation are provided in the supplementary.

4.2 Experiments

4.2.1 Data Pre-processing

We use the publicly available *VR-EyeTracking* [Xu et al., 2018b] dataset for training, which consists of 134 train and 74 test videos viewed by at least 31 subjects, lasting between $20 - 60$ seconds. We sampled the videos at 16 fps with a resolution of 960×1920. For cross-dataset evaluation, we use the *PVS-HMEM* [Xu et al., 2018a] and *360AV-HM* [Chao et al., 2020] datasets, which respectively contain 76 and 21 videos viewed by 58 and 15 subjects. The videos in the *PVS-HMEM* dataset have varying durations between $10 - 80$ secs, while those in the *360AV-HM* dataset have a duration of 25 secs. The videos in both datasets have a frame rate between 24-60 fps.

4.2.2 Architecture and Optimization Details

We use a ResNet-18 encoder pre-trained on ImageNet and keep it frozen while extracting features. Our baseline 360° video transformer consists of 6 blocks with

an embedding dimension of 512 and 8 attention heads. For the spatial position embeddings, we use linear projections of flattened pixel-wise angular coordinates, while we use rotary embeddings are used for the time dimension with a window size of $F = 8$. As in [Bertasius et al., 2021], we apply temporal self-attention followed by spatial self-attention in alternation. We train our SalViT360 model using the AdamW optimizer [Loshchilov and Hutter, 2017] with an initial learning rate of $1e-5$, default weight decay, and momentum parameters of $1e-2$ and $(\beta_1,\ \beta_2) = (0.9,\ 0.999)$, with a batch size of 16 for five epochs with early stopping. The experiments are conducted on a Tesla V100 GPU.

4.2.3 Evaluation Metrics and Loss Functions

As described in Chapter 2.4, we evaluate the performance of the models using Normalized Scanpath Saliency (NSS), KL-Divergence (KLD), Pearson's Correlation Coefficient (CC), and Similarity Metric (SIM). For the supervised loss, we use a weighted differentiable combination of KLD, CC, and Selective-MSE (MSE on normalized saliency maps at only eye-fixation points [Ding et al., 2022]) as given below:

$$\mathcal{L}_{supervised}(P,\ Q_s,\ Q_f) = \mathcal{L}_{KLD}(P,\ Q_s)\ +\ \mathcal{L}_{CC}(P,\ Q_s)\ +\ \alpha\mathcal{L}_{SMSE}(P,\ Q_s,\ Q_f) \tag{4.3}$$

where $P,\ Q_s,\ Q_f$ are the predicted saliency, ground truth density and fixation maps, respectively, and $\alpha = 0.005$.

4.2.4 Comparison with the State-of-the-art

In Table 4.1, we present the results of our SalViT360 model and the existing models. We evaluate the performance of SalViT360 with six state-of-the-art models for 360° image and video saliency prediction, namely CP-360 [Cheng et al., 2018], SalGAN360 [Chao et al., 2018], MV-SalGAN360 [Chao et al., 2020], Djilali et al. [Djilali et al., 2021], ATSal [Dahou et al., 2021], PAVER [Yun et al., 2022]. ATSal, PAVER, and CP-360 are video saliency models; the rest are image-based models developed

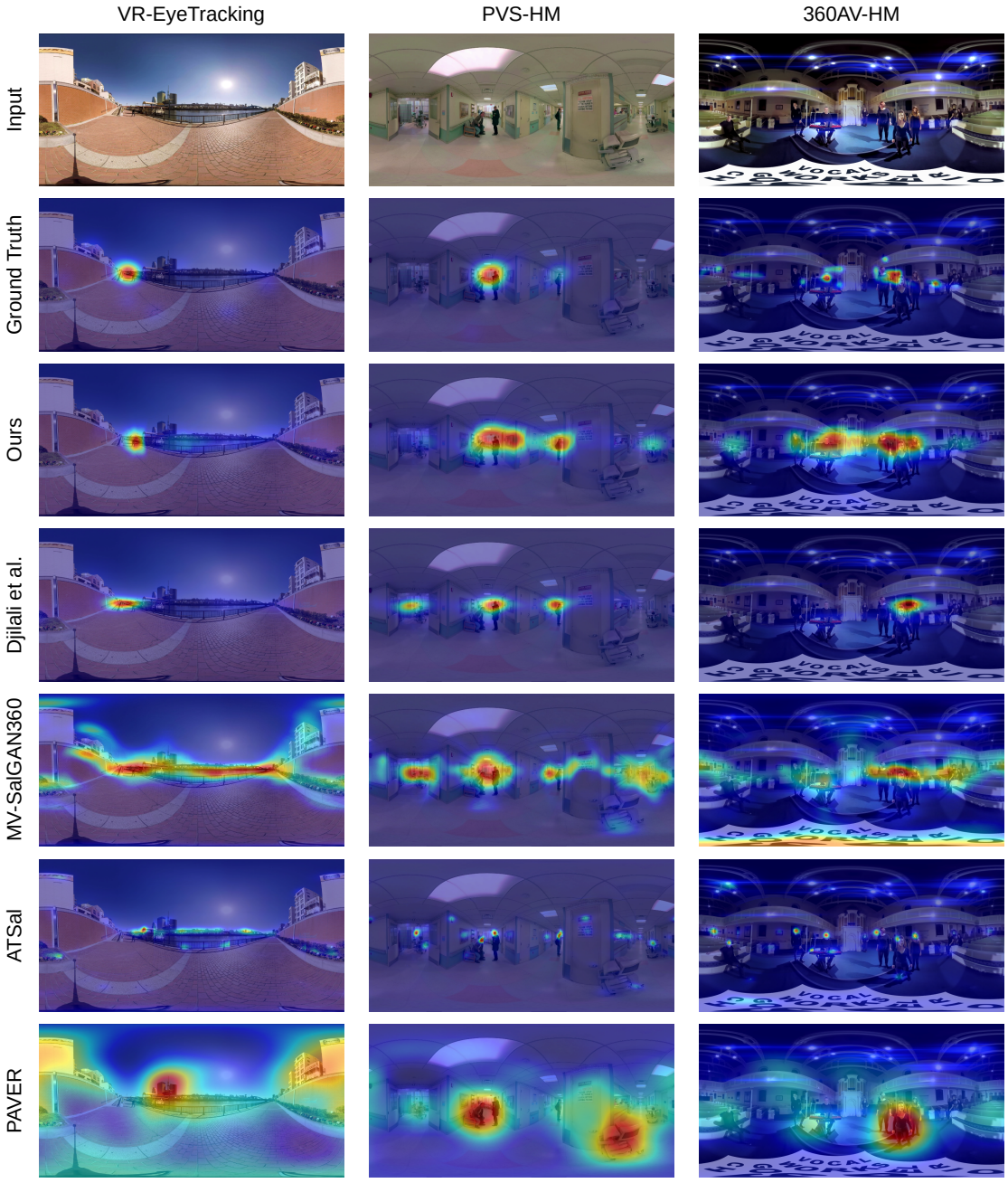

Figure 4.3: **Qualitative comparison on VR-EyeTracking [Xu et al., 2018b], PVS-HM [Xu et al., 2018a], and 360AV-HM [Chao et al., 2020] datasets.** Our proposed approach gives better results compared to existing models. The saliency predictions of our model better resemble the ground truth fixation density maps, producing sparse estimates while covering the most dominant modes better.

for the omnidirectional domain. On VR-EyeTracking test set, SalViT360 outperforms the state-of-the-art on three metrics and gives competitive results on NSS. On PVS-HMEM and 360-AVHM datasets, it outperforms the state-of-the-art by a mar-

Table 4.1: **Performance analysis of *SalViT360* against the state-of-the-art 360° saliency models** on VR-EyeTracking, PVS-HMEM, 360AV-HM datasets. While the scores in **bold** highlight the best performance, the underlined ones are the second best.

	VR-EyeTracking				PVS-HMEM				360AV-HM			
Method	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
CP-360	0.624	15.338	0.165	0.240	0.576	4.738	0.162	0.198	0.689	24.426	0.061	0.041
SalGAN360	1.753	10.845	0.370	0.355	1.513	4.394	0.314	0.291	0.719	25.301	0.065	0.036
MV-SalGAN360	1.818	8.713	0.382	0.357	1.546	4.112	0.316	0.295	0.716	25.322	0.066	0.036
ATSal	1.317	12.259	0.336	0.318	0.732	4.303	0.183	0.219	0.727	24.141	0.058	0.041
PAVER	1.511	13.267	0.307	0.294	0.750	3.736	0.224	0.269	0.732	23.944	0.065	0.035
Djilali et al.	**3.183**	6.570	0.565	0.475	1.688	2.430	0.447	0.404	1.727	22.889	0.148	0.085
SalViT360 (Ours)	2.630	**5.744**	**0.586**	**0.492**	**2.191**	**1.841**	**0.626**	**0.495**	**1.946**	**22.711**	**0.168**	**0.093**

gin on all metrics, demonstrating our proposed model's cross-dataset performance. These results demonstrate that SalViT360 has a better generalization capability than the existing methods. The qualitative comparison in Fig. 4.3 also shows the effectiveness of our approach in highlighting the salient regions more accurately.

4.2.5 Experimental Analysis and Ablation Studies

To assess the contribution of each component of our approach and to provide an in-depth analysis of spatio-temporal modelling, we perform additional experiments on the validation split of VR-EyeTracking dataset and report our findings in Table 4.2 and Table 4.3. In these experiments, we consider a baseline model comprised of a 2D ResNet-18 backbone, a vision transformer with Viewport Spatial Attention, and a CNN-based decoder.

Spherical Position Embeddings. We compare the performance of our proposed *spherical geometry-aware spatial position embeddings* with regular 1D learnable position embeddings. The results on all four metrics show that our proposed embedding method outperforms it, demonstrating that it is more suitable for processing spherical data with Vision Transformers.

Viewport Augmentation Consistency and Late-Fusion. We train our VSTA baseline using only the supervised loss. We then compare this single-scale baseline

Table 4.2: **Ablation study** for each component in our approach on the validation split of the VR-EyeTracking dataset.

Method	**# params**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
VSA (*w/ 1D Pos. Emb.*)	30.28M	2.518	6.445	0.560	0.472
+ *Spherical Pos. Emb.*	30.78M	2.575	6.221	0.563	0.475
VSTA (*w/ Sph. Pos. Emb.*)	37.07M	**2.664**	6.174	0.570	0.479
+ *VAC (w/o mask)*	37.07M	2.624	6.011	0.576	0.490
+ *VAC (w/ mask)*	37.07M	<u>2.630</u>	<u>5.744</u>	<u>0.586</u>	<u>0.492</u>
+ *Late-Fusion*	37.07M	2.578	**4.654**	**0.592**	**0.495**

to the one trained with a weighted combination of the supervised loss ($\mathcal{L}_{supervised}$) and the consistency loss ($\mathcal{L}_{VAC}$). Table 4.2 shows that VAC outperforms the VSTA baseline on three distribution-based saliency evaluation metrics with a performance gain of 3.2% on KLD, 6.4% on CC, and 5.8% on SIM, respectively. Lastly, we investigated the effect of using the predictions of two tangent sets in the final prediction. We perform this with an optional late-fusion as element-wise multiplication of two ERP predictions to highlight the consistently predicted salient regions better. This simple optional fusion improves the performance on three metrics significantly, with zero memory- and 0.5× time-overhead. We provide sample results for qualitatively comparing these components in Fig. 4.4, and detail them in the supplementary.

Table 4.3: **Spatio-temporal modelling performance** of *SalViT360* compared against two alternative approaches on the validation split of the VR-EyeTracking dataset.

Method	**# params**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
VSA + *2+1D-CNN Enc.*	28.82M	2.568	**5.915**	0.568	0.477
VSA + *Offline EMA*	30.78M	2.591	6.018	0.566	0.477
VSTA	37.07M	**2.664**	6.174	**0.570**	**0.479**

Spatio-Temporal modelling. We conducted several experiments to evaluate the effectiveness of our proposed Viewport Spatio-Temporal Attention (VSTA) mechanism. In Table 4.2, we compare Viewport Spatial Attention (VSA) and VSTA

Figure 4.4: **Qualitative comparison** for our *VSTA baseline (third row)*, with the proposed *VAC loss (fourth row)*, and the optional *late-fusion (last row)*, compared to the ground truth.

blocks to assess the contribution of temporal information processing in omnidirectional videos. In Table 4.3, we compare our VSTA with two distinct approaches, namely *2+1D-CNN* [Morgado et al., 2020] backbone and *Offline EMA*. *2+1D-CNN* backbone is an R2+1D model [Tran et al., 2018] which performs convolution over consecutive frames, pre-trained on undistorted *normal-FOV* crops in 360° videos. We replace ResNet-18 + VSTA with *2+1D-CNN* + VSA to introduce temporal features for spatial self-attention. In the other setting, we keep ResNet-18 and VSA and apply a weighted exponential moving average on F consecutive predictions for temporal aggregation. Additionally, we ablate on transformer depth, which shows that our VSTA blocks gradually learn better spatio-temporal representations in deeper

layers. Our experiments demonstrate that the proposed VSTA mechanism outperforms the spatial-only setting and the other two spatio-temporal approaches. We refer the reader to the supplementary for comprehensive experiments on the effect of *the transformer depth,* and *temporal window size* F, along with the performance of *joint spatio-temporal attention.*

Chapter 5

AUDIO-VISUAL SALIENCY PREDICTION FOR 360° VIDEOS

While the proposed SalViT360 model in Chapter 4 effectively uses spatio-temporal features for 360° scene understanding, it only considers visual stimuli input and ignores audio information. As mentioned in Chapter **??**, audio is essential in guiding human visual attention. In this chapter, we propose SalViT360-AV model for audio-visual saliency prediction in 360° videos, with two pre-trained and frozen backbones for audio and visual modalities. We employ an adapter fine-tuning method for data and parameter-efficient audio-visual tuning.

5.1 Methodology

The overview of SalViT360-AV pipeline is illustrated in Fig 5.1. We use the SalViT360 model as the vision backbone, and our implementation allows for any audio model as the audio backbone. The audio stream takes input ambisonic waveforms $a \in \mathbb{R}^{4\times N}$ in first-order ambisonics in 4-channel B-format. To simulate what the subjects are hearing while looking at a particular location, we rotate the ambisonics depending on the angular coordinates (θ, ϕ) for each of T tangent viewports of x_{tang} (1). The rotated waveforms are mono, which enables us to use any pre-trained audio backbone for feature extraction (2). The extracted features are passed to the Adapter layers in each upgraded transformer block (3) for audio-visual tuning. While the total number of parameters in the main (visual) pipeline is $37M$, the additional Adapter layers require only $700K$ parameters for tuning.

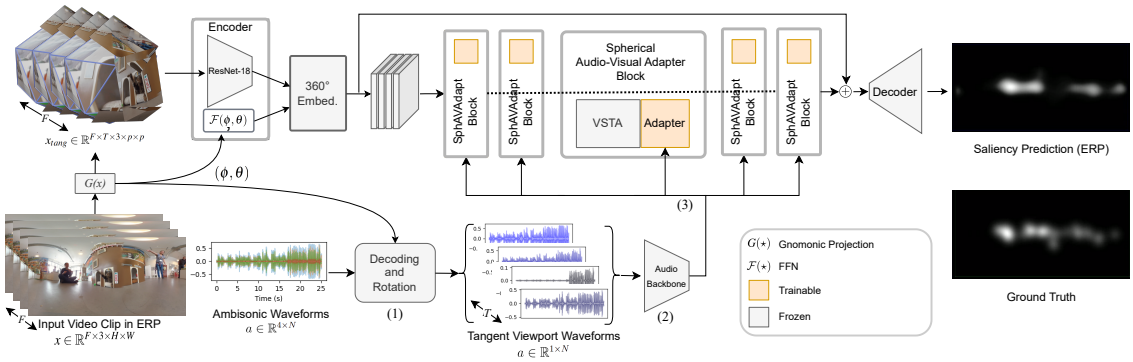

Figure 5.1: **System overview of the proposed SalViT360-AV pipeline.** SalViT360-AV employs a two-stream network for video and audio modalities. The video stream is pre-trained and frozen SalViT360 upgraded with learnable adapter layers in the transformer blocks for audio-visual fine-tuning. The audio stream encodes rotated ambisonics for each tangent viewport, and the Adapter layers use these features for audio-visual fine-tuning.

5.1.1 Audio Backbones

SalViT360-AV model can work with any pre-trained backbone for audio feature extraction, and we empirically picked PaSST [Koutini et al., 2022] and EZ-VSL [Mo and Morgado, 2022] models for testing. **PaSST** is an Audio Spectrogram Transformer [Gong et al., 2021] variant, which performs self-attention among mel-spectrogram patches. It is trained for audio classification on AudioSet [Gemmeke et al., 2017] with 527 classes. On the other hand, **EZ-VSL** is a CNN-based visual sound localization model, which is pre-trained on VGG-Sound dataset [Chen et al., 2020] with a multiple-instance contrastive loss objective between audio-visual pairs.

Using PaSST and EZ-VSL models as audio backbones gives us insight into the importance of audio pre-training objectives in audio-visual saliency prediction. We perform comprehensive experiments to determine whether a classification or a localization model performs better for the saliency downstream task and present the discussion on the results.

5.1.2 Audio-Visual Adapter

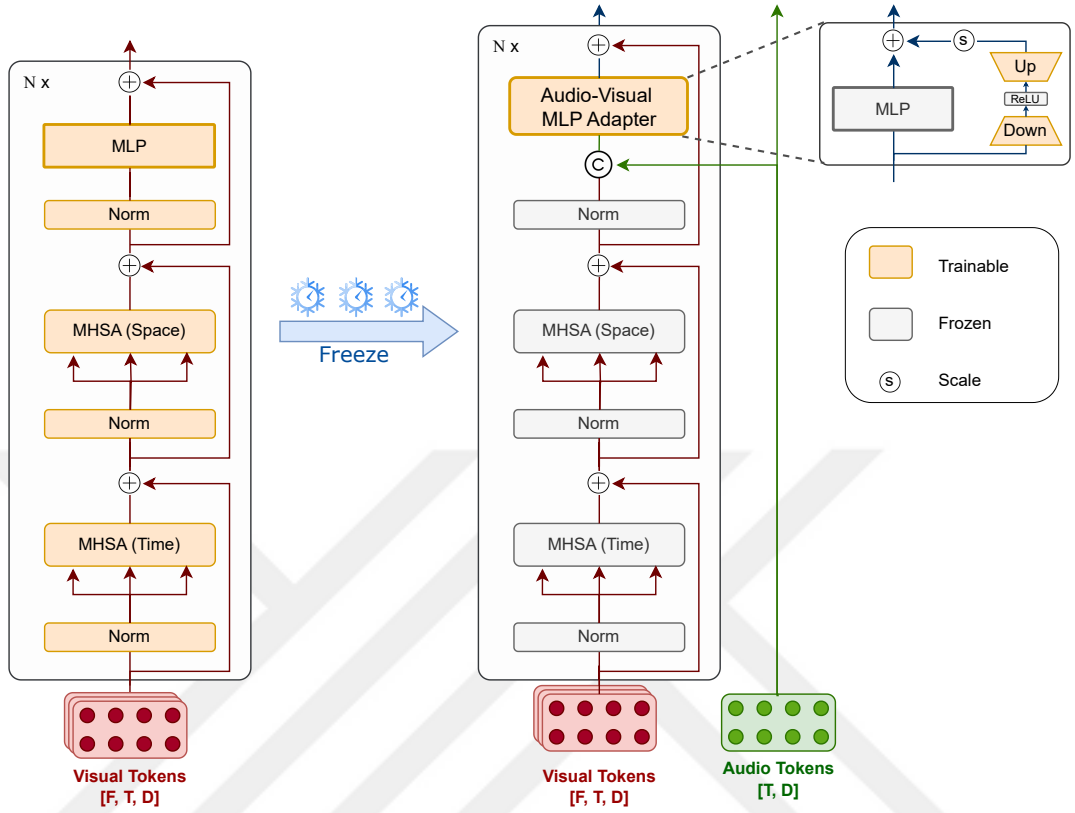

Figure 5.2: **Proposed adapter fine-tuning for audio-visual modalities.** For each of the N transformer blocks in the SalViT360 pipeline, the visual tokens after Viewport Spatio-Temporal Attention are concatenated with audio tokens. The pre-trained MLPs at the end of each VSTA block are upgraded with *Audio-Visual MLP Adapters*, which consist of a scaled combination of a frozen MLP (left) and trainable bottleneck module (right). F, T, D denote the number of frames and tangent viewports tokens, clip length, and the embedding dimension, respectively.

5.2 Experiments

We use the newly collected YT360-EyeTracking dataset for training. Also, we report the test results and state-of-the-art comparison on YT360-EyeTrackingand 360AV-HM datasets. As mentioned in Chapter 2.2, VR-EyeTracking includes only mono audio modality, while the others have ambisonics. At this point, we conduct our experiments on two main settings: (1) using the mono modality of all datasets and (2) leaving the VR-EyeTracking dataset out and using the ambisonic modality of the remaining two datasets.

5.2.1 Data Pre-processing

Sampling and Clipping While the EZ-VSL backbone does not require a specified sample rate, the audio waveforms are resampled 32.0kHz for the PaSST backbone. Audio clip duration is empirically selected as up to 4-seconds[1], i.e., for a video frame at 10'th second, the audio input to the model is considered between $6-10$'th seconds. A comparative analysis of clip duration is left for future work.

Obtaining Viewport-Specific Waveforms

As mentioned in Chapter 2.3, the ambisonics are rotated and decoded in real-time for the subjects to perceive the sounds from the direction they are looking at. To achieve the same effect in our tangent-viewport-based saliency model, we propose rotating the ambisonics for each of $T = 18$ tangent viewports by taking the product of ambisonic channels and the spherical harmonics rotation matrices for given angular coordinates (θ, ϕ) [Kronlachner, 2014, Morgado et al., 2018].

5.2.2 Comparison with the State-of-the-art

Table 5.1: **Performance analysis of *SalViT360-AV* against state-of-the-art** on 360AV-HM and YT360-EyeTracking datasets. While † represents audio models with ambisonics modality, * corresponds to the ones with mono. **Bold** highlight the best scores, and the underlined ones are the second best.

	360AV-HM *[Chao et al., 2020]*				**YT360-EyeTracking**			
Method	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
STAViS* *[Tsiami et al., 2020]*	1.826	18.216	0.267	0.221	2.009	13.311	0.339	0.281
+ **SSSL**† *[Cokelek et al., 2021]*	1.937	17.997	0.281	0.238	2.141	10.961	0.366	0.299
DAVE* *[Tavakoli et al., 2019]*	1.712	18.519	0.262	0.219	1.911	13.295	0.336	0.276
+ **SSSL**†	1.829	18.247	0.277	0.237	2.042	10.899	0.362	0.297
AVS360† *[Chao et al., 2020]*	**2.633**	13.981	**0.382**	**0.296**	2.342	9.611	0.465	0.373
SalViT360-AV† (Ours)	2.473	**13.830**	0.379	0.278	**2.553**	**7.847**	**0.521**	**0.411**

In Table 5.1, we report the results of our SalViT360-AV model alongside existing mono and ambisonics audio-visual saliency methods, namely, STAViS [Tsi-

[1]Discussion in [Cokelek et al., 2021] show that shorter clips are sensitive to short-term audio saliency, and longer clips overshoot local saliency.

ami et al., 2020], DAVE [Tavakoli et al., 2019], AVS360 [Chao et al., 2020], and SSSL [Cokelek et al., 2021] models. We train our model on the training set of YT360-EyeTracking, and report performance on all videos in 360AV-HM and the test set on YT360-EyeTracking dataset. While STAViS and DAVE use mono modality, SSSL, AVS360, and SalViT360-AV leverages directional sound information in ambisonics. As mentioned in Chapter 3.3, SSSL is proposed as an audio saliency bias on top of an existing model, and we report SSSL predictions via the late-fusion with STAViS and DAVE. The results demonstrate the significant performance gain of using spatial audio over mono, which complements our motivation and discussion on using spatial audio. For instance, SSSL bias outperforms its mono baselines in two datasets, and the overall best results on both datasets correspond to AVS360 and SalViT360-AV models that encode ambisonics modality.

5.2.3 Experimental Analysis and Ablation Studies

To provide a more comprehensive analysis of each component in our audio-visual saliency pipeline, we perform experiments on the validation split of YT360-EyeTracking.

Impact of audio modality. The datasets including spatial audio are collected under three audio conditions: (1) muted, (2) mono, and (3) ambisonics and each audio modality is viewed by different subject groups to prevent biases. We evaluate our video-only baseline on each ground truth separately. The results reported in Table 5.1.1 demonstrate that the increasing representative power of audio modality yields higher inter-subject correlation on our video-only model.

Table 5.2: **Comparison of our video baseline SalViT360 on the ground truth sets** for mute, mono, and ambisonics modalities. The best scores are highlighted in **bold**.

	360AV-HM				**YT360-EyeTracking**			
Audio Modality of GT	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
Mute	1.961	16.819	0.289	0.225	2.115	**9.798**	0.461	0.370
Mono	2.230	16.356	0.329	0.238	2.315	9.993	0.478	0.371
Ambisonics	**2.285**	**15.879**	**0.349**	**0.246**	**2.346**	9.861	**0.484**	**0.373**

Performance of audio Backbones and modalities. As mentioned in Chapter 5.1.1, we compare the performance of two audio backbones in our pipeline, namely, PaSST [Koutini et al., 2022] and EZ-VSL [Mo and Morgado, 2022]. We also conduct our experiments on mono and ambisonics modalities to compare the performance of each modality on audio-visual tuning. Table 5.3 compares the two backbones on two modalities. We get an average performance gain of over 25% on all four metrics, compared to our video baseline SalViT360. Also, using ambisonics performed better than mono in both audio backbones. However, we observe that EZ-VSL and PaSST backbones perform head-to-head in all metrics for both modalities. The detailed comparison and tuning for audio backbones are left for future work. Lastly, the table shows that the performance gain of ambisonics over mono is less significant than the gain of mono over muted setting. In Appendix D, we discuss the possible reason for this situation by comparing the features obtained from ambisonics and mono modalities. We also introduce an analysis of the performance of adapter scaling on VR-EyeTracking and YT360-EyeTracking datasets.

Table 5.3: **Comparison of SalViT360-AVconfigurations** using PaSST and EZ-VSL audio backbones and trained with mono and ambisonics. The best scores are highlighted in **bold**.

	360AV-HM				**YT360-EyeTracking**			
Method	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
SalViT360	1.961	16.819	0.289	0.225	2.346	9.861	0.484	0.373
SalViT360-AV w/ PaSST								
On Mono	2.470	13.869	0.371	0.273	2.540	7.689	0.513	0.405
On Ambisonics	**2.473**	**13.830**	**0.379**	**0.278**	2.553	**7.847**	**0.521**	**0.411**
SalViT360-AV w/ EZ-VSL								
On Mono	2.471	13.947	0.370	0.274	2.538	7.701	0.513	0.406
On Ambisonics	2.473	13.933	**0.379**	**0.278**	**2.554**	7.853	**0.521**	**0.411**

Chapter 6

CONCLUSION

In this thesis, we proposed two models to address 360° video and audio-visual saliency prediction. SalViT360, our first model has an encoder-transformer-decoder architecture executing on undistorted tangent image representations. We introduced a spatio-temporal attention mechanism on tangent viewports to model the global spatio-temporal information on omnidirectional scenes effectively. Our framework employs new model-agnostic spherical geometry-aware position embeddings based on angular coordinates (ϕ, θ). Lastly, we proposed a consistency-based regularization term as an unsupervised loss to reduce projection artefacts in dense-prediction models that occur after inverse projection. Our ablation studies clearly show the contribution of each component to saliency prediction accuracy. The experimental results obtained from three omnidirectional video saliency datasets demonstrate that our proposed SalViT360 model outperforms the state-of-the-art qualitatively and quantitatively. We also evaluate our model on a downstream task for visual attention-guided omnidirectional video quality assessment. Our analysis in Appendix C shows that our model outperforms state-of-the-art saliency models for quality assessment in two PSNR variants when replaced with ground truth head movement weights. These results demonstrate the performance of SalViT360 model in highlighting the perceptually important regions in 360° videos in a practical example.

While SalViT360 effectively models the human visual attention system using spatial and temporal information in 360° scenes, it ignores auditory cues that have a significant impact on guiding our perception. With SalViT360-AV, upgrade our SalViT360 model by incorporating audio modality using a parameter-efficient adapter fine-tuning. SalViT360-AV employs pre-trained and frozen backbones for auditory and visual inputs. The audio pipeline comprises an ambisonics decoding and rota-

tion step for the spatial audio modality to generate unique waveforms representing the audio that are actually perceived in each tangent viewport. Our extensive experiments demonstrate the performance gain obtained from the audio information with a simple fine-tuning in two omnidirectional video datasets with spatial audio. To further boost the model performance with spatial audio over mono modality, developing spatial-audio-specific encoders is left for future work, alongside investigating the impact of audio clip duration on saliency accuracy.

Tackling the limitation on publicly available ODV saliency datasets with spatial audio modality, we introduce YT360-EyeTracking, the second large-scale dataset with ambisonics. It consists of 81 videos carefully distributed by audio-visual semantics, which are viewed by 45 participants in total. The code, pre-trained models, and the dataset will be publicly available.

BIBLIOGRAPHY

[Bertasius et al., 2021] Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

[Bylinskii et al., 2018] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2018). What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757.

[Chao et al., 2020] Chao, F., Zhang, L., Hamidouche, W., and Deforges, O. (2020). A multi-fov viewport-based visual saliency model using adaptive weighting losses for 360-degree images. *IEEE Transactions on Multimedia*, pages 1–1.

[Chao et al., 2020] Chao, F.-Y., Ozcinar, C., Wang, C., Zerman, E., Zhang, L., Hamidouche, W., Deforges, O., and Smolic, A. (2020). Audio-visual perception of omnidirectional video for virtual reality applications. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

[Chao et al., 2020] Chao, F. Y., Ozcinar, C., Zhang, L., Hamidouche, W., Deforges, O., and Smolic, A. (2020). Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 355–358.

[Chao et al., 2020] Chao, F.-Y., Ozcinar, C., Zhang, L., Hamidouche, W., Deforges, O., and Smolic, A. (2020). Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 355–358. IEEE.

[Chao et al., 2018] Chao, F.-Y., Zhang, L., Hamidouche, W., and Deforges, O.

(2018). Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *Proc. IEEE ICMEW*.

[Chen et al., 2020] Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

[Chen et al., 2021] Chen, J., Li, Q., Ling, H., Ren, D., and Duan, P. (2021). Audiovisual saliency prediction via deep learning. *Neurocomputing*, 428:248–258.

[Chen et al., 2022] Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., and Luo, P. (2022). Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*.

[Chen et al., 2014] Chen, Y., Nguyen, T. V., Kankanhalli, M., Yuan, J., Yan, S., and Wang, M. (2014). Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(11):1992–2003.

[Cheng et al., 2018] Cheng, H.-T., Chao, C.-H., Dong, J.-D., Wen, H.-K., Liu, T.-L., and Sun, M. (2018). Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proc. IEEE/CVF CVPR*, pages 1420–1429.

[Cokelek et al., 2021] Cokelek, M., Imamoglu, N., Ozcinar, C., Erdem, E., and Erdem, A. (2021). Leveraging frequency based salient spatial sound localization to improve 360 video saliency prediction. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE.

[Coors et al., 2018] Coors, B., Condurache, A. P., and Geiger, A. (2018). Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533.

[Dahou et al., 2021] Dahou, Y., Tliba, M., McGuinness, K., and O'Connor, N. (2021). ATSal: An attention based architecture for saliency prediction in 360 videos. In *International Conference on Pattern Recognition*, pages 305–320.

[Ding et al., 2022] Ding, G., İmamoğlu, N., Caglayan, A., Murakawa, M., and Nakamura, R. (2022). SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image and Vision Computing*, 120:104395.

[Djilali et al., 2021] Djilali, Y. A. D., Krishna, T., McGuinness, K., and O'Connor, N. E. (2021). Rethinking 360deg image visual attention modelling with unsupervised learning. In *Proc. IEEE/CVF ICCV*, pages 15414–15424.

[Eder et al., 2020] Eder, M., Shvets, M., Lim, J., and Frahm, J.-M. (2020). Tangent images for mitigating spherical distortion. In *Proc. IEEE/CVF CVPR*, pages 12426–12434.

[Esteves et al., 2018] Esteves, C., Allen-Blanchette, C., Makadia, A., and Daniilidis, K. (2018). Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68.

[Gemmeke et al., 2017] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

[Gong et al., 2021] Gong, Y., Chung, Y.-A., and Glass, J. (2021). AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575.

[Greene, 1986] Greene, N. (1986). Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications*, 6(11):21–29.

[Guan et al., 2017] Guan, J., Yi, S., Zeng, X., Cham, W.-K., and Wang, X. (2017). Visual importance and distortion guided deep image quality assessment framework. *IEEE Transactions on Multimedia*, 19(11):2505–2520.

[Hou and Kung, 2022] Hou, Z. and Kung, S.-Y. (2022). Multi-dimensional dynamic model compression for efficient image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 633–643.

[Jabbari Tofighi et al., 2023] Jabbari Tofighi, N., Hedi Elfkir, M., Imamoglu, N., Ozcinar, C., Erdem, E., and Erdem, A. (2023). St360iq: No-reference omnidirectional image quality assessment with spherical vision transformers. *arXiv e-prints*, pages arXiv–2303.

[Koutini et al., 2022] Koutini, K., Schlüter, J., Eghbal-zadeh, H., and Widmer, G. (2022). Efficient training of audio transformers with patchout. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2753–2757. ISCA.

[Kronlachner, 2014] Kronlachner, M. (2014). Spatial transformations for the alteration of ambisonic recordings. *M. Thesis, University of Music and Performing Arts, Graz, Institute of Electronic Music and Acoustics*, 7.

[Li et al., 2018a] Li, C., Xu, M., Du, X., and Wang, Z. (2018a). Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 932–940, New York, NY, USA. ACM.

[Li et al., 2018b] Li, S., Xu, M., Ren, Y., and Wang, Z. (2018b). Closed-form optimization on saliency-guided image compression for hevc-msp. *IEEE Transactions on Multimedia*, 20(1):155–170.

[Li et al., 2022] Li, Y., Guo, Y., Yan, Z., Huang, X., Duan, Y., and Ren, L. (2022).

Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proc. IEEE/CVF CVPR*, pages 2801–2810.

[Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[Min et al., 2014] Min, X., Zhai, G., Gao, Z., Hu, C., and Yang, X. (2014). Sound influences visual attention discriminately in videos. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 153–158. IEEE.

[Min et al., 2020] Min, X., Zhai, G., Zhou, J., Zhang, X.-P., Yang, X., and Guan, X. (2020). A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819.

[Mishra et al., 2021] Mishra, D., Singh, S. K., Singh, R. K., and Kedia, D. (2021). Multi-scale network (mssg-cnn) for joint image and saliency map learning-based compression. *Neurocomputing*, 460:95–105.

[Mo and Morgado, 2022] Mo, S. and Morgado, P. (2022). Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*.

[Morgado et al., 2020] Morgado, P., Li, Y., and Nvasconcelos, N. (2020). Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33.

[Morgado et al., 2018] Morgado, P., Nvasconcelos, N., Langlois, T., and Wang, O. (2018). Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31.

[Ozcinar et al., 2021] Ozcinar, C., İmamoğlu, N., Wang, W., and Smolic, A. (2021). Delivery of omnidirectional video using saliency prediction and optimal bitrate allocation. *Signal, Image and Video Processing*, 15:493–500.

[Pan et al., 2017] Pan, J., Canton, C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., and Giro-i Nieto, X. a. (2017). SalGAN: Visual saliency prediction with generative adversarial networks. In *arXiv*.

[Patney et al., 2016a] Patney, A., Kim, J., Salvi, M., Kaplanyan, A., Wyman, C., Benty, N., Lefohn, A., and Luebke, D. (2016a). Perceptually-based foveated virtual reality. In *ACM SIGGRAPH 2016 Emerging Technologies*, pages 1–2.

[Patney et al., 2016b] Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., and Lefohn, A. (2016b). Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph. (TOG)*, 35(6):1–12.

[Perrott et al., 1990] Perrott, D. R., Saberi, K., Brown, K., and Strybel, T. Z. (1990). Auditory psychomotor coordination and visual search performance. *Perception & psychophysics*, 48(3):214–226.

[Qiao et al., 2020] Qiao, M., Xu, M., Wang, Z., and Borji, A. (2020). Viewport-dependent saliency prediction in 360 video. *IEEE Trans. Multimed.*, 23:748–760.

[Qiu and Shao, 2021] Qiu, M. and Shao, F. (2021). Blind 360-degree image quality assessment via saliency-guided convolution neural network. *Optik*, 240:166858.

[Su and Grauman, 2017] Su, Y.-C. and Grauman, K. (2017). Learning spherical convolution for fast features from 360 imagery. *Advances in Neural Information Processing Systems*, 30.

[Sun et al., 2017] Sun, Y., Lu, A., and Yu, L. (2017). Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, 24:1408–1412.

[Tavakoli et al., 2019] Tavakoli, H. R., Borji, A., Rahtu, E., and Kannala, J. (2019). Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*.

[Tran et al., 2018] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

[Tsiami et al., 2020] Tsiami, A., Koutras, P., and Maragos, P. (2020). Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4766–4776.

[Vatakis and Spence, 2007] Vatakis, A. and Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & psychophysics*, 69:744–756.

[Vroomen and Gelder, 2000] Vroomen, J. and Gelder, B. d. (2000). Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology: Human perception and performance*, 26(5):1583.

[Wang et al., 2020] Wang, H., Li, Z., Li, Y., Gupta, B. B., and Choi, C. (2020). Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, 130:64–72.

[Xu et al., 2018a] Xu, M., Song, Y., Wang, J., Qiao, M., Huo, L., and Wang, Z. (2018a). Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence*.

[Xu et al., 2018b] Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., and Gao, S. (2018b). Gaze prediction in dynamic 360 immersive videos. In *Proc. IEEE/CVF CVPR*, pages 5333–5342.

[Yang et al., 2019] Yang, S., Jiang, Q., Lin, W., and Wang, Y. (2019). Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1383–1391.

[Yun et al., 2022] Yun, H., Lee, S., and Kim, G. (2022). Panoramic vision transformer for saliency detection in 360-degree videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 422–439. Springer.

[Zhang et al., 2018] Zhang, Z., Xu, Y., Yu, J., and Gao, S. (2018). Saliency detection in 360-degree videos. In *Proc. ECCV*.

[Zhu et al., 2021] Zhu, M., Hou, G., Chen, X., Xie, J., Lu, H., and Che, J. (2021). Saliency-guided transformer network combined with local embedding for no-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1953–1962.

[Zhu et al., 2020] Zhu, S., Liu, C., and Xu, Z. (2020). High-definition video compression system based on perception guidance of salient information of a convolutional neural network and hevc compression domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):1946–1959.

[Zhu and Xu, 2018] Zhu, S. and Xu, Z. (2018). Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275:511–522.

Appendix A

ADDITIONAL EXPERIMENTS ON VSTA

A.1 Temporal Window Size

To investigate the impact of temporal window size (F) on the representational power of our omnidirectional video saliency prediction model, we vary the number of frames in a video clip, and analyze its effect. These experiments are performed on our VSTA baseline, which consists of 6 transformer blocks with an embedding dimension of $D = 512$ and 8 attention heads. We present the results of our experiments using four saliency evaluation metrics in Fig. A.1, providing insights into the performance of our model across different temporal window sizes.

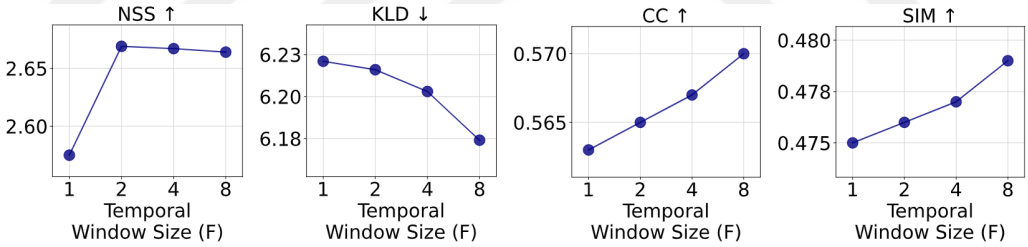

Figure A.1: Performance of our saliency prediction model on four evaluation metrics as a function of temporal window size (F) on the validation split of VR-EyeTracking [Xu et al., 2018b] dataset.

A.2 Transformer Depth

We analyze the influence of the number of transformer blocks on the performance and the computational complexity of our saliency prediction model for omnidirectional videos. In Fig. A.2, we present our experimental results, showing how the performance varies with different numbers of transformer blocks.

In Fig. A.2, the performance gain from $N = 0$ to $N = 1$, highlights the effectiveness of our VSTA mechanism. Notably, even with a single VSTA transformer block, our

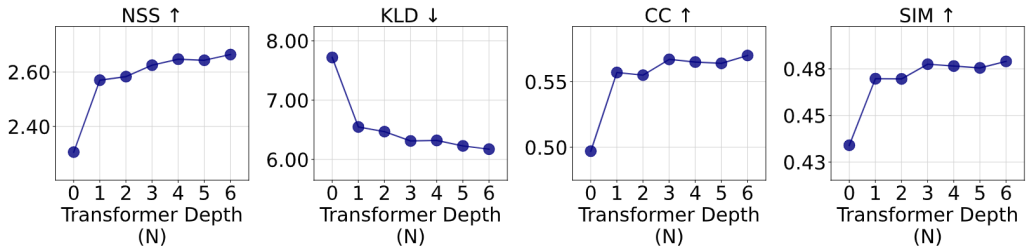

Figure A.2: Performance of our saliency prediction model on four evaluation metrics as a function of VSTA depth (N) on the validation split of VR-EyeTracking dataset.

model shows the ability to capture rich 360° spatio-temporal features. As the depth of the transformer blocks increases, the model performance continues to improve. However, we conclude our experiments after $N = 6$ transformer blocks, taking into consideration the model size as reported in Table A.1.

A.2.1 Comparison with Joint Spatio-Temporal Attention

Table A.1 compares our proposed VSA and VSTA mechanisms with joint spatio-temporal attention, which computes self-attention among all frames and tokens in a video clip. Since VSTA computes spatio-temporal attention in two stages (time and space), the model size becomes larger than VSA / JSTA. On the other hand, VSTA is computationally more efficient as its complexity grows linearly with respect to temporal window size, where it grows quadratically in JSTA.

Table A.1: **Quantitative comparison** for space-only attention, the proposed Viewport Spatio-Temporal Attention and existing Joint Spatio-Temporal attention for 360° saliency prediction on the validation split of VR-EyeTracking dataset. Due to memory limitations, JSTA model could not be trained on our GPU.

Attention	# params	GFLOPs	NSS↑	KLD↓	CC↑	SIM↑
None	11.81M	0.00	2.306	7.718	0.497	0.434
VSA	30.78M	57.08	2.575	6.221	0.563	0.475
VSTA	37.07M	63.30	2.664	6.174	0.570	0.479
JSTA	30.78M	77.57	n/a	n/a	n/a	n/a

Appendix B

ADDITIONAL EXPERIMENTS ON VAC

In this section, we describe the proposed *Viewport Augmentation Consistency* in more detail. To address the discrepancies in overlapping regions of tangent predictions, we propose to use a second -*augmented*- tangent image set and minimize the difference between the predictions of these pairs with an additional loss term. Following [Li et al., 2022], we sampled $T = 18$ tangent images at four latitudes: $-67.5°, -22.5°, 22.5°, 67.5°$ for the original set. The tangent images are sampled for each latitude level with 90° apart in longitude. We extracted each tangent image with a resolution of 224×224 and field-of-view (FOV) of 80°. We generated the augmented tangent image set under three configurations: (1) horizontally shifting viewports, (2) using a larger FOV for each viewport, and (3) varying the number, position, and FOV of the tangent viewports. It is important to emphasize that the augmented tangent images share weights with the original set, which neither requires extra parameters nor increases model complexity during training. Our experimental results in Fig. B.2 demonstrate that each augmentation method improves model consistency significantly.

B.1 Approaches for Augmenting Tangent Images

B.1.1 Shifting Viewport Centers

In this setting, we keep the number and FOV of tangent images the same. We obtain the shifted tangent image set by applying a 45° horizontal shift on each viewport.

B.1.2 FOV Augmentation

In the second set, we keep the position of each tangent image the same and generate the augmented set by increasing their FOV to 120°. Augmented FOV also provides

the model with a multi-scale representation for the same input.

B.1.3 Viewport Augmentation

In the last setting, we generate the second set with $T' = 10$ tangent images with a FOV of 120°, located in three latitudes: $-60°, 0°, 60°$. We sample $3, 4, 3$ viewports for each latitude.

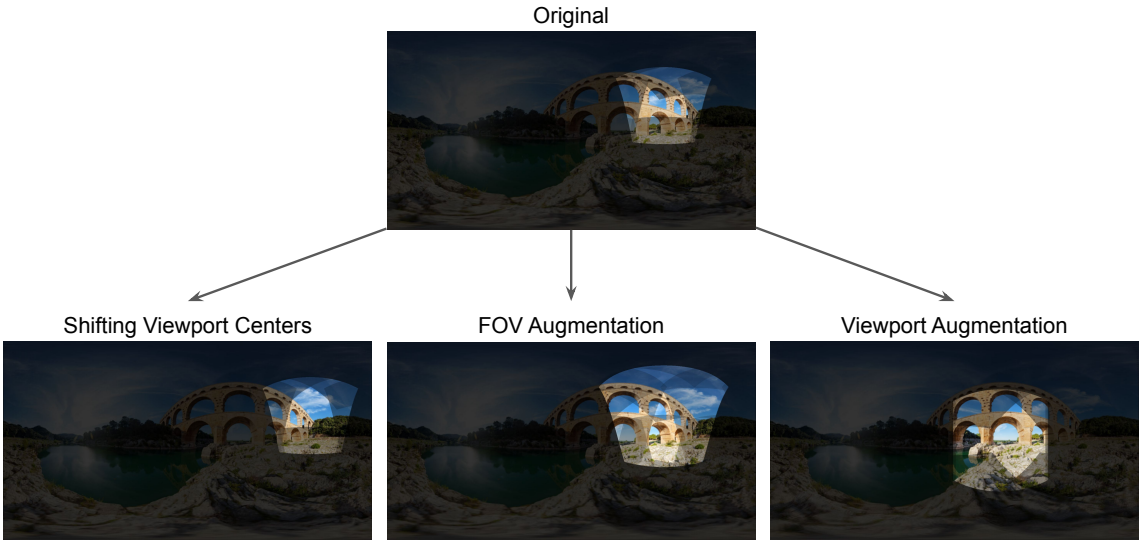

Figure B.1: A tangent viewport from the original projection, highlighted on Equirectangular Projection (ERP) *(top)*, compared with three augmentation methods *(bottom)*.

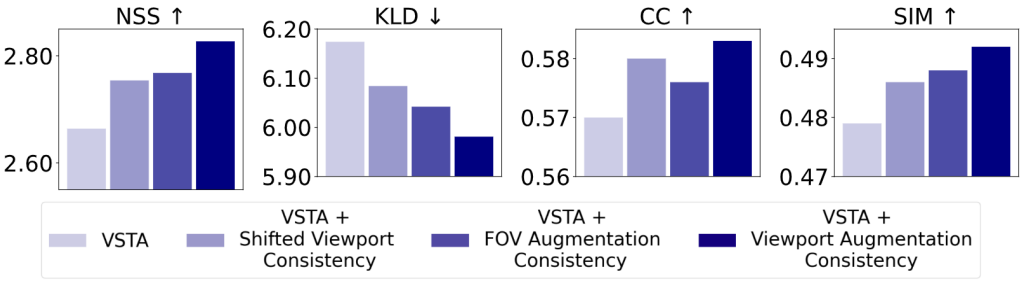

Figure B.2: Performance of our VSTA baseline compared with three augmentation consistency methods on four metrics, on the validation split of VR-EyeTracking dataset.

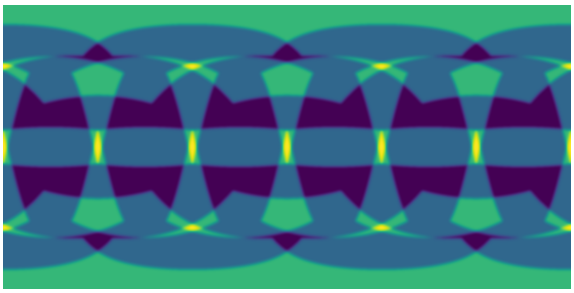

Figure B.3: **Weight mask used for VAC Loss.** Each pixel coordinate in ERP takes a value based on the number of tangent viewports it is projected onto (max. 4). Brighter colours represent increasing overlaps.

Table B.1: **Quantitative comparison** for the proposed VAC Loss with and without mask, on the validation split of VR-EyeTracking dataset.

Model	NSS↑	KLD↓	CC↑	SIM↑
VSTA + VAC *(w/o mask)*	2.624	6.011	0.576	0.490
VSTA + VAC *(w/ mask)*	2.630	5.744	0.586	0.492

B.1.4 Mask-weighted VAC Loss.

We use an optional weight for the proposed $\mathcal{L}_{VAC}(P,\ P')$ loss to increase consistency, especially on the overlapping regions on ERP. In Fig. B.3, we provide the weight mask computed from the gnomonic projection for the original tangent image set. The performance comparison in Table B.1 demonstrates the effectiveness of the proposed masking operation.

Appendix C

USE CASE: SALIENCY-GUIDED OMNIDIRECTIONAL VIDEO QUALITY ASSESSMENT

Table C.1: **Comparison with state-of-the-art** saliency models for *saliency-guided omnidirectional video quality assessment* on VQA-ODV [Li et al., 2018a] dataset, as a downstream task.

	PSNR			WS-PSNR [Sun et al., 2017]		
Weight	PCC↑	SRCC↑	RMSE↓	PCC↑	SRCC↑	RMSE↓
None	0.650	6.664	7.502	0.671	0.686	7.233
PAVER [Yun et al., 2022]	0.661	0.667	7.481	0.679	0.691	6.914
Djilali et. al. [Djilali et al., 2021]	0.648	0.721	7.336	0.684	0.721	6.829
SalViT360 (ours)	0.688	0.733	7.295	0.689	0.737	6.673
HM (Supervised)	0.764	0.759	6.601	0.759	0.756	6.612

In Table C.1, we report the performance of two PSNR variants compared to ground-truth DMOS values in the VQA-ODV [Li et al., 2018a] dataset. Each row corresponds to saliency weights that supply human-perceptual information to PSNR and WS-PSNR metrics. The ground truth head movement (HM) maps refer to the viewports that human subjects have viewed while rating the visual quality of ODVs. Predicting saliency maps that better capture human head movements will result in better performance in the PSNR metrics. The table demonstrates that our proposed saliency prediction model better highlights the perceptually important regions in 360° videos compared to the state-of-the-art for omnidirectional video quality assessment downstream task.

Following the prior work [Yun et al., 2022], the saliency-weighted PSNR and WS-PSNR values are calculated as:

$$\mathrm{PSNR}_{sal} = 10 \log_{10} \frac{Y_{max}^2 \cdot \sum_{p \in \mathbb{P}} w_{sal}(p)}{\sum_{p \in \mathbb{P}} (Y(p) - Y'(p))^2 \cdot w_{sal}(p)} \tag{C.1}$$

$$\mathrm{WS-PSNR}_{sal} = 10 \log_{10} \frac{Y_{max}^2 \cdot \sum_{p \in \mathbb{P}} w_{sal}(p) \cos\theta_p}{\sum_{p \in \mathbb{P}} (Y(p) - Y'(p))^2 \cdot w_{sal}(p) \cos\theta_p} \tag{C.2}$$

where Y_{max} is the maximum intensity of the frames, $Y(p)$ and $Y'(p)$ denote the intensities for of pixel p in the reference and impaired videos, and θ_p is the latitude at pixel p.

Appendix D

DISCUSSION ON SalViT360-AV RESULTS

D.1 Marginal performance gain from ambisonics over mono modality.

Table 5.3 demonstrates that although using ambisonics improves performance over mono modality, the performance gain is limited compared to the one between mono and muted settings. We investigate the possible reasons by visualizing the extracted ambisonics features. In Figure D.1, we used the t-SNE method to visualize PaSST features of rotated ambisonics for 15 videos.

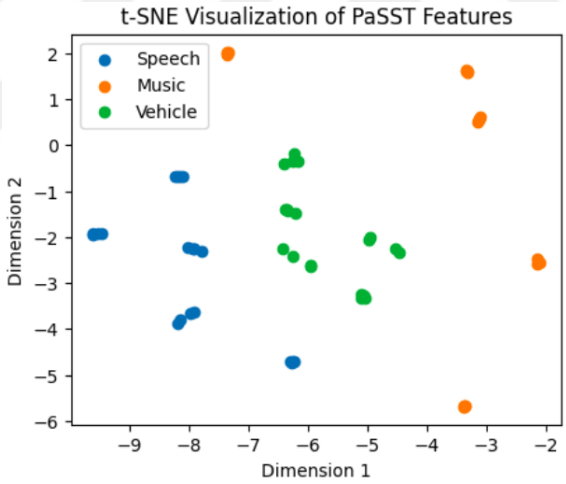

Figure D.1: **t-SNE visualization of ambisonics features**, encoded by PaSST model. Each data point corresponds to a tangent viewport waveform obtained by ambisonics rotation. Each small cluster consists of four randomly rotated waveforms for a video, and each colour represents audio categories in YT360-EyeTracking dataset.

From the figure, we make several observations. (1) Different audio samples are distinguished from each other. (2) Audio categories are easily identified. (3) However, the features of directional waveforms are very close to each other. This means that even though the perceived audio is rotated for different directions, the pre-trained audio encoders are robust to these changes. This motivates us to fine-tune

the audio encoders for learning diverse representations for rotated audio waveforms or use the direction-of-arrival of spatial audio explicitly, such as the audio energy maps used in AVS360 [Chao et al., 2020].

D.2 Additional Experiments on Adapter Scaling Factor

The learnable scaling parameter in Figure 5.2 controls the information flow from audio features to the vision transformer. A scale value of 0 means the model does not utilize the audio modality. During training, the parameter is initialized as 1.0, and learned. Our experiments on VR-EyeTracking and YT360-EyeTracking datasets demonstrate a contradiction regarding the performance gain of increasing scale values on the two datasets.

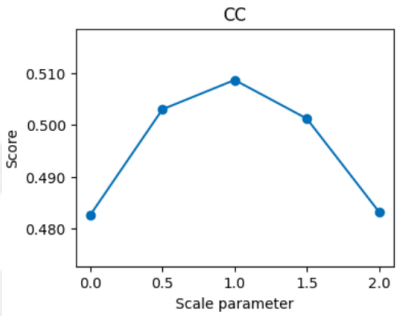

Figure D.2: CC score of SalViT360-AV on YT360-EyeTracking mono dataset as a function of scale values.

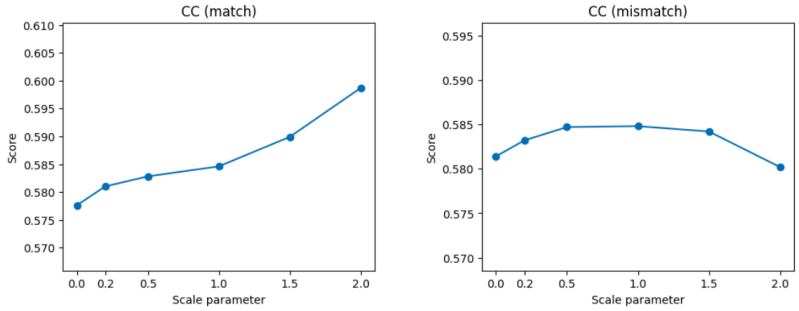

Figure D.3: CC scores of SalViT360-AV on VREyeTracking (mono) dataset as a function of scale values for two audio-visual scenarios (match, mismatch).

In YT360-EyeTracking dataset, all of the audio-video pairs align, i.e., every video has its corresponding original audio. In contrast, nearly half of the videos in VR-EyeTracking have audio that is added later as background music. This analysis demonstrates that the learned scale parameter for YT360-EyeTracking dataset is optimal. On the other hand, for the VR-EyeTracking dataset, we observe that for

the audio-visually aligning videos, the model requires a higher scale value for the adapter. For the mismatching videos, we see that the model performs worse when more audio information is passed to the visual model. These results motivate us to train scaling parameters for each sample separately, which we leave as future work. For presentation clarity, only the CC metric is reported.