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

Mert Cokelek*, Halit Ozsoy*, Nevrez Imamoglu, Cagri Ozcinar, Inci Ayhan, Erkut Erdem, Aykut Erdem

Abstract—Omnidirectional videos (ODVs) are redefining viewer experiences in virtual reality (VR) by offering an unprecedented full field-of-view (FOV). This study extends the domain of saliency prediction to 360° environments, addressing the complexities of spherical distortion and the integration of spatial audio. Contextually, ODVs have transformed user experience by adding a spatial audio dimension that aligns sound direction with the viewer's perspective in spherical scenes. Motivated by the lack of comprehensive datasets for 360° audio-visual saliency prediction, our study curates YT360-EyeTracking, a new dataset of 81 ODVs, each observed under varying audio-visual conditions. Our goal is to explore how to utilize audio-visual cues to effectively predict visual saliency in 360° videos. Towards this aim, we propose two novel saliency prediction models: SalViT360, a vision-transformer-based framework for ODVs equipped with spherical geometry-aware spatio-temporal attention layers, and SalViT360-AV, which further incorporates transformer adapters conditioned on audio input. Our results on a number of benchmark datasets, including our YT360-EyeTracking, demonstrate that SalViT360 and SalViT360-AV significantly outperform existing methods in predicting viewer attention in 360° scenes. Interpreting these results, we suggest that integrating spatial audio cues in the model architecture is crucial for accurate saliency prediction in omnidirectional videos. Code and dataset will be available at: `https://cyberiada.github.io/SalViT360/`.

Index Terms—Audio-Visual Saliency Prediction, 360° Videos, Vision Transformers, Adapter Fine-tuning

1 Introduction

The rapid proliferation of virtual reality (VR) and multimedia streaming platforms has led to an unprecedented surge in the popularity and usage of omnidirectional or 360° videos (ODVs). These videos offer viewers a fully immersive experience by providing a complete field-of-view (FOV), fundamentally reshaping the way users interact with visual content. Accurately predicting viewer attention, termed *visual saliency prediction*, is increasingly critical in optimizing various multimedia tasks such as video compression [1, 2, 3, 4], perceptually-driven super-resolution [5], quality assessment [6, 7, 8, 9, 10], and adaptive streaming techniques like foveated rendering in VR environments [11]. The success of these applications hinges on a precise understanding of where users naturally fixate their attention during video viewing.

- * *These authors contributed equally to this work.*
- *M. Cokelek is with the Department of Computer Science and Engineering, Koç University, Istanbul, TR, 34450.*

  E-mail: mcokelek21@ku.edu.tr
- *H. Ozsoy is with the Department of Psychology, Boğaziçi University, Istanbul, TR, 34342.*
- *N. Imamoglu is with the National Institute of Advanced Industrial Science and Technology (AIST), Intelligent Platforms Research Institute, Tokyo, JP. N. Imamoglu is also with AIST, CNRS-AIST Joint Robotics Laboratory, Tsukuba, JP.*
- *C. Ozcinar is with MSK.AI, London, UK.*
- *I. Ayhan is with the Department of Psychology, Boğaziçi University University, Istanbul, TR, 34342.*
- *E. Erdem is with the Department of Computer Engineering, Hacettepe University, Ankara, TR, 06800.*
- *A. Erdem is with the Department of Computer Science and Engineering, Koç University, and KUIS AI Center, Istanbul, TR, 34450.*

Predicting visual saliency in ODVs, however, presents unique challenges not encountered with traditional 2D video content. The inherent spherical geometry of ODVs introduces significant spatial distortions when conventional projection methods, such as equirectangular projection (ERP) or cubemaps, are employed. ERP, while computationally simple, severely distorts content near the poles, and cubemaps introduce discontinuities at the edges, both negatively impacting model prediction accuracy. To overcome these challenges, in our recent work, we proposed SalViT360 [12], a vision transformer (ViT) [13] based model explicitly tailored for omnidirectional videos. SalViT360 uniquely addressed spherical distortions using locally undistorted tangent images generated through gnomonic projection [14], and incorporated spherical geometry-aware spatio-temporal transformer attention. This model established state-of-the-art performance for visual-only saliency prediction in ODVs. Though, a critical limitation remained: the absence of audio modality integration, despite its significant role in guiding viewer attention in immersive scenarios.

Indeed, human perception naturally and continuously integrates auditory and visual stimuli. Numerous studies confirm that audio cues significantly modulate visual attention, guiding eye movements and enhancing viewer immersion [15, 16, 17]. In ODVs, spatial audio provides directional auditory cues aligned with visual elements, profoundly influencing where viewers focus their attention. Despite its recognized importance, existing computational models and datasets rarely consider audio-visual interactions comprehensively. To illustrate and emphasize the impact of spatial audio, we introduce an illustrative example in Fig. 1, which demonstrates how integrating spatial audio significantly

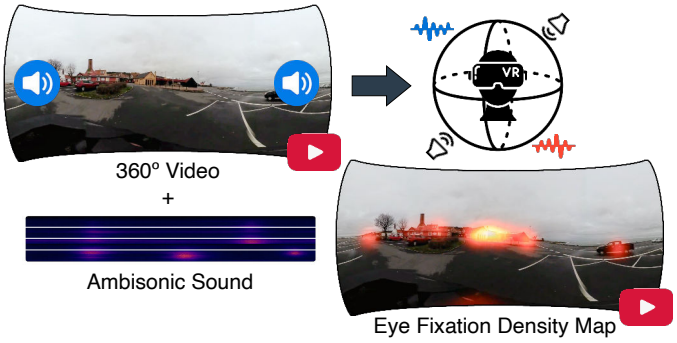


Fig. 1: **Audio-Visual Saliency in 360° Videos.** Illustration of how spatial audio cues influence visual attention in omnidirectional videos. In this example, spatial audio highlights salient regions by directing viewer attention towards audio-emitting objects such as a passing car and birds singing in the trees, emphasizing the necessity of integrating audio modalities into saliency prediction models.

alters visual attention patterns in omnidirectional contexts.

Motivated by the significant yet under-explored influence of spatial audio on visual attention, we present a novel, large-scale dataset named YT360-EyeTracking. This dataset is explicitly curated for audio-visual saliency prediction in ODVs, offering extensive annotations across various audio conditions, including mute, mono, and spatial (ambisonic) audio. YT360-EyeTracking addresses existing limitations by providing a robust foundation to systematically investigate how different audio modalities affect viewer gaze behavior in omnidirectional video scenarios.

In addition, we introduce SalViT360-AV, a novel audio-visual saliency prediction framework tailored specifically for 360° content. SalViT360-AV significantly expands upon our earlier SalViT360 model by incorporating a parameter-efficient transformer-adapter architecture inspired by AdaptFormer [18]. This enables effective integration of spatial audio cues into the transformer-based visual saliency prediction pipeline. By capturing complex interactions between audio and visual stimuli, SalViT360-AV achieves substantially improved prediction accuracy, setting a new state-of-the-art benchmark for audio-visual saliency prediction in omnidirectional contexts.

In summary, our contributions to the domain of audio-visual saliency prediction for 360° videos are as follows:

- We introduce YT360-EyeTracking, the first large-scale dataset explicitly designed for audio-visual saliency prediction in ODVs. It uniquely enables the systematic analysis of audio-visual interactions under varying audio modalities (mute, mono, ambisonics).
- We propose SalViT360-AV, a novel audio-visual saliency prediction framework that innovatively extends our previous visual-only model, SalViT360, by incorporating spatial audio through transformer adapters. This approach efficiently captures intricate audio-visual interactions, significantly improving model accuracy.
- We conduct comprehensive experiments across four benchmark datasets, including our newly introduced YT360-EyeTracking dataset, demonstrating that SalViT360-AV achieves state-of-the-art performance in predicting viewer attention, effectively highlighting the critical role of spatial audio integration in saliency prediction tasks.

2 Background and Related Work

Saliency prediction models aim 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 with eye-tracking hardware and represented as fixation and saliency maps. Fixation maps contain the actual eye gaze data of humans. Fixations are binary for each pixel in the scene, corresponding to whether an observer has fixated on that location or not. Saliency maps are heatmaps generated by convolving the fixation maps with a Gaussian kernel to create a continuous measure of saliency or likelihood that a location attracts human attention. In this section, we provide an overview of 360° video saliency prediction models and datasets.

2.1 Overview of Saliency Prediction Models

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. For 360° saliency prediction, previous works 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 drive the human visual attention system. First, we review the 360° static and dynamic saliency literature and give an overview of the audio-visual saliency models for 2D and 360° domains.

360° Static Saliency Prediction. Chao et al.[19] introduced SalGAN360, utilizing cubemap projection to reduce ERP distortion. They fine-tuned the SalGAN model[20] for cubemap faces, then reprojected the outcomes to ERP for final predictions. While this method lessens local distortion, it faces edge discontinuities due to CNNs' inability to recognize adjacency across cube faces, leading to prediction errors at edges. To address this, the authors suggest employing multiple cubemap sets with horizontal and vertical shifts, resulting in 81 projections per image, which significantly increases computational demand. In the follow-up work MV-SalGAN360 [21], the authors extended SalGAN360 with multi-view fusion. Unlike 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 $180° \times 360°$ FOV. While the multi-view representation is important for local-global scene encoding, MV-SalGAN360 requires training a separate set of parameters for each level, which is computationally expensive.

Dahou et al. [30] proposed ATSal, a two-stream model to compute global and local saliency in 360° videos. They use global prediction as a rough attention estimate, and the

TABLE 1: **A Comparison of Existing 360° Video Saliency Datasets.** Our YT360-EyeTracking dataset contains a large number of sequences with eye-tracking data collected from the highest number of subjects for both mute, mono, and ambisonics audio modalities.

Dataset	Num. Subjects	Num. Videos	Avg. Dur. (s)	Frame Rate	Min. Res.	Audio Categories	Audio Modality	Color Modality
Salient360! [22]	57	20	20s	25-30	4K	n/a	Mute	Colored
PVS-HMEM [23]	58	76	30s	24-60	2K	n/a	Mute	Colored
360-EM Dataset [24]	13	14	60s	25-60	4K	Misc.	Mono	Colored
VR-EyeTracking [25]	45	208	30s	24-60	4K	Speech, Music, Vehicle, Noise	Mono	Colored
SVGC-AVA [26]	63	57	28s	25-50	2K	Speech, Music, Sports	Ambisonics	Colored
PAV-SOD [27]	40	67	30s	25-60	4K	Speech, Music, Misc.	Ambisonics	Colored
360AV-HM [28]	45	15	25s	25-60	4K	Speech, Music, Environment	Mute, Mono, Ambisonics	Colored
AVS-ODV [29]	60	162	15s	30	8K	Speech, Music, Vehicle	Mute, Mono, Ambisonics	Colored
YT360-EyeTracking	102	81	30s	24-30	4K	Speech, Music, Vehicle	Mute, Mono, Ambisonics	Colored, Grayscale

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. Zhang et al. [31] developed a spherical U-Net for saliency prediction, applying modified kernels directly on ERP in spherical coordinates. These convolution kernels adapt to their location on the sphere, with dynamic receptive fields and a parameter-sharing approach for translation equivariance. However, computing a unique spherical crown for each convolution increases the model's runtime complexity. Djilali et al. [32] 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.

Zhu et al.[33] suggested a framework for generating separate head and eye saliency maps for static panoramas using some hand-crafted features and graphical models. Their follow-up [34] extended this pipeline by combining visual equilibrium, uncertainty, and object cues to refine head and eye saliency predictions. A deep reinforcement learning agent was later introduced in [35], encoding viewport content (CNN features, spherical coordinates, visited maps) and employing a free-energy-based reward to predict head movement trajectories in static 360° images.

360° Dynamic Saliency Prediction. Cheng et al. [36] propose cube-padding to address discontinuities at cube face boundaries, facilitating information transfer across faces through overlapping regions. Particularly for CNN models, the information exchange increases with deeper layers. The model also utilizes ConvLSTMs to capture temporal features. Qiao et al. [37] found that eye fixation distribution varies with viewport locations and treated 360° video saliency prediction as a multi-task challenge, aiming to predict global saliency on the sphere and local saliency within viewports. This insight into fixation biases based on viewport locations inspired us to incorporate *spherical position information* into our models. Zhu et al. [38] fuse RGB and motion cues over high-clarity viewports with a graph-based approach to predict saliency, demonstrating that combining appearance and dynamic information at select viewports improves temporal consistency. Yun et al. [39] developed PAVER, utilizing deformable convolutions to correct spherical distortion, assigning specific offsets per convolution. This setup allows for spatial and temporal self-attention in vision transformers using undistorted patch features.

Several recent works attempt to reduce spherical distortion through projections or modified kernels while performing dense prediction tasks. For instance, Eder et al. [14] introduced tangent image representations using gnomonic projection to convert spherical images into multiple overlapping patches, each tangent to an icosahedron face. Their method, which also inspired our approach, addresses distortion at a local level but lacks explicit modeling of global interactions across tangent viewports. Similarly, Cheng et al. [36] utilize cube projections but do not incorporate mechanisms to capture spatial dependencies across cube faces.

While such techniques do facilitate localized predictions, they fall short in capturing global saliency cues critical for comprehensive scene understanding. In particular, Zhu et al. [38] and SalViT360 do process the whole field-of-view by integrating multiple partial views, but their fusion strategies differ: Zhu et al. rely on late or local fusion of viewport-specific outputs, whereas SalViT360 explicitly models long-range dependencies across all viewports at each layer. Panoramic Vision Transformer [37] does perform spatio-temporal aggregation via deformable convolutions, but its windowed attention remains restricted to local regions. In contrast, our method enhances the tangent image approach by integrating spatial and temporal attention across viewports, allowing the model to reason globally while maintaining distortion-aware local representations. We note, however, that Yun et al. [39] perform spatio-temporal modeling through spatially deformable convolutions.

In this work, we propose using tangent images to process undistorted local viewports and develop a transformer-based model, SalViT360, to learn their global and spatio-temporal association for 360° video saliency prediction. Our transformer module is inspired by TimeSformer [40], which is developed for 2D videos. TimeSformer proposes Divided Space-Time Attention, which computes spatio-temporal attention in two stages to optimize for model complexity. We present Viewport Spatio-Temporal Attention, which extends the Divided Space-Time Attention idea to 360° videos.

2D Audio-Visual Saliency. Min et al. [41] proposed one of the earliest audio-visual saliency prediction methods, emphasizing the role of audio in guiding gaze. Their method computes separate spatial, temporal, and audio attention maps where the audio attention is generated through canonical correlation analysis between motion patterns and audio features—to produce an integrated audio-visual saliency map. Subjective eye-tracking experiments confirmed that this fusion improves prediction quality over visual-only models. Tavakoli et al. [16] proposed DAVE, an audio-visual encoder framework for 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. Tsiami et al. [15] introduced STAViS, a spatio-temporal audio-visual saliency model that also adopts a two-stream CNN-based encoder. They further proposed a bilinear audio localization module that fuses the audio and visual features before decoding them into saliency maps.

Min et al. [17] presented the MMS model for fusing audio-visual cues under the assumption of high audio-visual correspondence. Their framework emphasizes the motion-consistent audio sources and performs late fusion with existing saliency maps, though it is less effective when visual and auditory cues diverge. Chen et al. [42] proposed an alternative 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, demonstrating superior performance than fusing the features from only the final layers.

Zhu et al. [43] propose LAVS, which extracts visual features using a VGG-16 backbone with ConvLSTM and audio features via separable convolutions on MFCC, CFCC and CQT features. Cross-modal alignment is achieved through deep canonical correlation analysis, and the resulting audio map is fused with the visual map, yielding competitive accuracy with minimal computational overhead. More recently, Zhu et al. [44] proposed MTCAM, a Transformer-based weakly supervised model for audio-visual saliency prediction. MTCAM jointly models audio and visual cues using separate encoders and a unified transformer-based fusion mechanism. Xiong et al. [45] introduced DiffSal, a diffusion-based audio-visual saliency model that unifies saliency learning with generative modeling. DiffSal jointly learns audio and visual representations through a multi-stage diffusion process and enables fine-grained and high-resolution prediction with improved temporal coherence.

360° Audio-Visual Saliency. Spatial audio has been increasingly recognized as a critical cue for saliency prediction in omnidirectional videos (ODVs), particularly due to its alignment with immersive VR experiences. Several recent works have explored this direction.

Chao et al. [46] proposed a two-stream pipeline where ERP frames are projected to padded cube maps and processed via a 3D ResNet, while mono audio mel-spectrograms are encoded separately. The resulting features are fused and enhanced with an audio energy map (AEM) that estimates the direction of arrival based on first-order ambisonics. Cokelek et al. [47] introduced an unsupervised, audio-only method called MCSR, which computes directional audio saliency via mel-cepstrum features across six primary directions, projecting this as a bias onto ERP-based maps. More recently, Zhu et al. [48] proposed AVPS, a unified saliency prediction model integrating both spatial audio localization and content-level audio features. Their approach combines improved group-equivariant CNNs with E3D-LSTMs to capture visual dynamics and computes spatial audio localization via AEM, further complemented by SoundNet-based attribute-aware audio embeddings. In another work, Zhu et al. [29] presented OmniAVS, a fully convolutional framework that introduces a learnable spatial audio map (SAM) based on ambisonic energy. OmniAVS combines this SAM with visual features using an attention-based fusion module that models temporal and directional saliency consistency. Yang et al. [26] introduced SVGC-AVA, a graph-based model leveraging spherical vector-based graph convolution and audio-visual attention fusion. Their model encodes spherical geometry and spatial audio jointly to predict fixations.

It is worth noting that the aforementioned recent works are concurrent with ours and some of them have not yet released code or pretrained models. As such, we are unable to provide direct quantitative comparisons. Nevertheless, we include all these works here to present a comprehensive overview of the current research landscape in 360° audio-visual saliency prediction. Our SalViT360-AV model advances this research line by integrating spatial audio via transformer adapters conditioned on ambisonic input, allowing for dynamic, fine-grained cross-modal fusion within an end-to-end trainable architecture. Compared to previous efforts, it captures deeper semantic and spatial correlations, leading to superior or competitive performance on a large number of benchmark datasets.

2.2 360° Video Saliency Datasets

Several datasets have been proposed for visual saliency prediction in omnidirectional videos (ODVs), capturing varying levels of realism, scale, and audio modality.

Salient360 dataset [22] is one of the earliest datasets to provide head movement (HM) and eye fixation (EM) annotations over ERP or cubemap projections. While it includes diverse content, it remain relatively small in size.

PVS-HMEM dataset [23] includes head movement and eye-tracking data collected from 76 $360°$ video sequences from 58 subjects. The videos vary between $10 - 80$ secs, with a minimum resolution of 2K and frame rate between $24-60$fps. They includes indoor, outdoor, movie, and sports categories. The data collection is done in a muted modality.

VR-EyeTracking dataset [25] consists of eye tracking data collected for 208 videos (134 train, 74 test) 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. At least 31 subjects view each video, with each subject watching around 35 videos. Videos are picked under various scene categories, including indoors, outdoors, sports,



Fig. 2: **Sample Input Frames from Our YT360-EyeTracking Dataset.** Rows and columns respectively represent different audio and visual categories.

games, documentation, short movies, and music shows. Subjects have viewed the videos with mono audio modality, as in traditional 2D videos. Mono modality delivers sound sources equally from all directions. Hence, the viewers cannot be guided by directional sound sources. 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).

SVGC-AVA dataset [26] is a recent benchmark proposed to evaluate audio-visual saliency in ODVs. It comprises 57 360° videos recorded with spatial audio, annotated with eye-tracking data from 63 subjects in free-viewing conditions, supporting both ERP and cubemap formats. The scenes include complex audio-visual events such as overlapping sounds and directionally distinct audio sources, providing a challenging benchmark for multimodal attention modeling.

360AV-HM dataset [28] consists of 15 ODVs with spatial audio lasting 25 seconds. Each video has a resolution of 3840×1920 in ERP format, with frame rates between $24 - 60$fps. The videos are collected from YouTube under three categories, Conversation, Music, and Environment. The videos are viewed under mute, mono, and spatial audio settings by three subject groups, where each subject has only viewed the videos with one modality. Each video was viewed by fifteen subjects, and there are a total of 16 female and 29 male subjects aged between $21 - 40$ years.

AVS-ODV [29] dataset, proposed concurrently with our work, includes 162 ODV clips spanning human activities, natural scenes, vehicles categories and provides saliency annotations under three audio conditions (mute, mono, spatial). The clips are sourced from YouTube and feature real-world audio-visual correlations. The dataset is recorded with 15 participants using a VR headset and includes a thorough psychophysical analysis of how different audio configurations influence gaze behavior.

Together, these datasets illustrate the progression of 360° saliency research from smaller-scale, visual-only datasets to larger-scale benchmarks integrating multimodal annotations, particularly spatial audio. As detailed in Table 1, our YT360-EyeTracking dataset, as compared to the recently introduced SVGC-AVA[26] and AVS-ODV[29], employs a significantly larger number of subjects. Additionally, YT360-EyeTracking explicitly explores the influence of audio modality by systematically providing annotations under three audio conditions (mute, mono, ambisonics), closely matching the concurrent AVS-ODV dataset in experimental conditions. But it also ensures a balanced gender representation among participants, featuring 53 female and 49 male subjects. These comparative strengths make our dataset particularly suitable for training deep-learning-based multimodal models such as SalViT360-AV, facilitating detailed analyses of audio's role in guiding visual attention in omnidirectional video scenarios.

3 YT360-EyeTracking Dataset

Overview. We picked 81 videos from YouTube-360 [49] by carefully considering nine audio-visual semantic categories and selecting nine videos per category, viewed under three audio conditions (mute, mono, spatial audio) and two color conditions (colored, grayscale[1]). The audio semantic categories are (1) speech, (2) music, (3) vehicle, and visual categories are (1) indoors, (2) outdoors-natural, and (3) outdoors-human-made. Sample frames from each audio and scene category are given in Fig. 2. The distribution of samples for each audio and scene category is given in detail in Fig. 3. Each video in our dataset contains dynamic scenes rather than static images and does not have cutscenes. The video clips last 30 secs, and have a resolution of 3840×1920 pixels in ERP format, with frame rates between $24 - 30$fps. Each video is watched by at least 15 observers under specific audio-visual conditions, with each observer viewing the video only once. YouTube video IDs, segment start timestamps, raw fixation data, as well as data preparation and post-processing scripts, will be made available to the public.

3.1 Stimuli

Videos. All clips within the YT-360 dataset were extracted as 10-second segments from longer videos, and there are multiple consecutive clips from the same videos. After re-downloading the original videos and trimming the reference

1. Including grayscale video allows isolating luminance-driven attention from chromatic effects, offering a clearer view of the role of color in visual saliency. Psychophysical studies have shown that color can influence gaze allocation and scene interpretation in meaningful ways [50, 51], making grayscale conditions a valuable complementary modality for exploring cross-modal and cross-condition saliency patterns. See [52] for related preliminary psychophysical analyses.

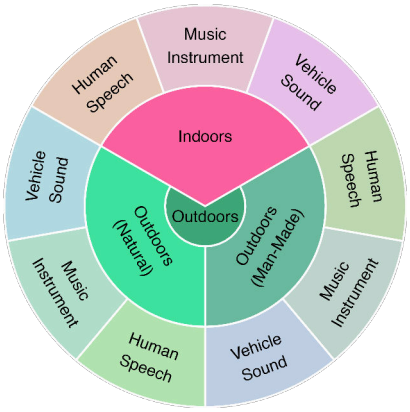


Fig. 3: **Observation Sessions in YT360-EyeTracking by Category and Format.** The dataset includes 9 clips per scene-audio combo, with 7-29 participants viewing each audio/color version, averaging 16.77 participants per clip.

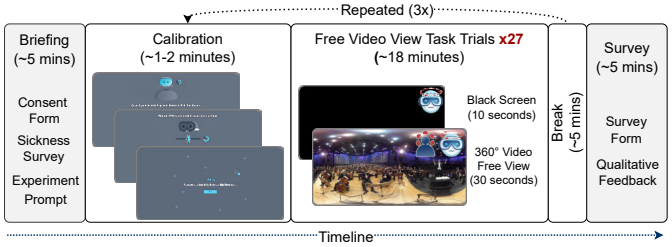


Fig. 4: **YT360-EyeTracking Data Collection Procedure.** Participants completed a free-viewing task across three sessions with 81 videos, incorporating a 5-min. break between sessions to lessen fatigue. Sessions started with eye calibration, followed by 27 audio-visual stimuli, each being a 30-sec. video clip in one of three pseudo-randomized audio (mute, mono, ambisonics) and color (grayscale, colored) conditions, separated by 10-sec. black screen intervals to reduce carry-over effects.

sequences to 30 seconds, we obtained our new stimuli for ODV saliency. Moreover, we chose to exclusively incorporate high-quality videos, with a minimum resolution of 3840×1920, and spatial audio modality. To facilitate the data filtering process while also keeping valuable semantic segmentation information in our new dataset, we deliberately chose the subset of the original sequences in YT-360.

Data Filtering and Labeling. We employed learning-based models [53, 54, 55] for initial video annotation, and selected the top 300 videos based on their annotation scores for a more detailed evaluation. For each of these videos, we manually labeled both the audio and scene categories. To streamline this process, we developed a website equipped with an intuitive interface specifically designed for video annotation. The annotation interface prompted researchers to identify the predominant type of audio in the video. The options provided were: vehicle sound, musical instrument, human speech, human vocalizations, other, and noise. For the environmental context, the interface asked whether the video was recorded indoors or outdoors, with further distinctions made between natural and human-made outdoor environments. We also incorporated several control questions to improve the overall stimuli quality. These questions aimed to identify issues such as black screens, sudden scene changes, and the origin of the audio (whether it was captured live or added in post-production). Lastly, we evaluated the ability to localize sound sources within the videos and the presence of text overlays, which helped us eliminate any erroneous content from our dataset. Following the initial annotation and filtering process, we are left with videos divided into three audio (speech, music, vehicle) and three visual (indoors, outdoors-human-made, outdoors-natural) categories.

To determine the number of stimuli for the dataset, *a priori* power analysis was conducted using G*Power version 3.1 [56]. Results indicated that to achieve 95% power for detecting a medium effect ($f = 0.25$), at a significance criterion of $\alpha = .05$, the required sample size was $N = 63$ for a mixed model ANOVA (3×3 between factors of sound and scene types, 3×2 within factors of sound and color modalities) corresponding to at least $n = 7$ videos per audio-visual category. Consequently, we chose nine representative videos for each audio-scene combination, resulting in a curated set of 81 videos, ensuring a diverse and representative dataset.

Pre-processing. Videos only available in higher resolutions were downsized to 3840×1920 pixels via FFmpeg [57] using bicubic interpolation to ensure consistency. In addition to the original first-order ambisonics, we altered all audio files to mono and mute versions by removing the audio channels. Moreover, we converted videos to grayscale by keeping only the Y (luma) component of the YUV format. As a result of these augmentations, we obtained a total of 486 distinct stimuli from 81 clips.

3.2 Eye-Tracking Data Collection

As illustrated in Fig. 4, our study employed a free-viewing task structured into three trials, with a 5-min. intermission between each to minimize participant fatigue. Each session began with eye calibration to ensure accuracy in tracking. Participants were then presented with a sequence of 27 stimuli in each trial, comprising a 30-second video clip accompanied by corresponding audio. To mitigate any carry-over effects, each stimulus was followed by a 10-second interval displaying a black screen. Overall, each video was viewed by each participant under one of six audio-visual conditions. Fig. 5 illustrates the variation in fixation density maps across these different audio modalities.

Before commencing the trials, we provided participants with guidelines to standardize the viewing experience. They were instructed to stand and view the 360° videos, allowing rotation but maintaining their position. Participants were informed about the lengths of trials, video clips, and intervening black screen periods. Although the specifics of the video content were not detailed, they were told to expect various presentations, including grayscale, colored, mute, and audiovisual formats, with an assurance of no distressing content to reduce surprises.

To examine the influence of audio and visual variables, a crossed experimental design was utilized. This design ensured that each participant experienced every video clip under specified audio and color conditions, with the sequence of these presentations randomized to control for

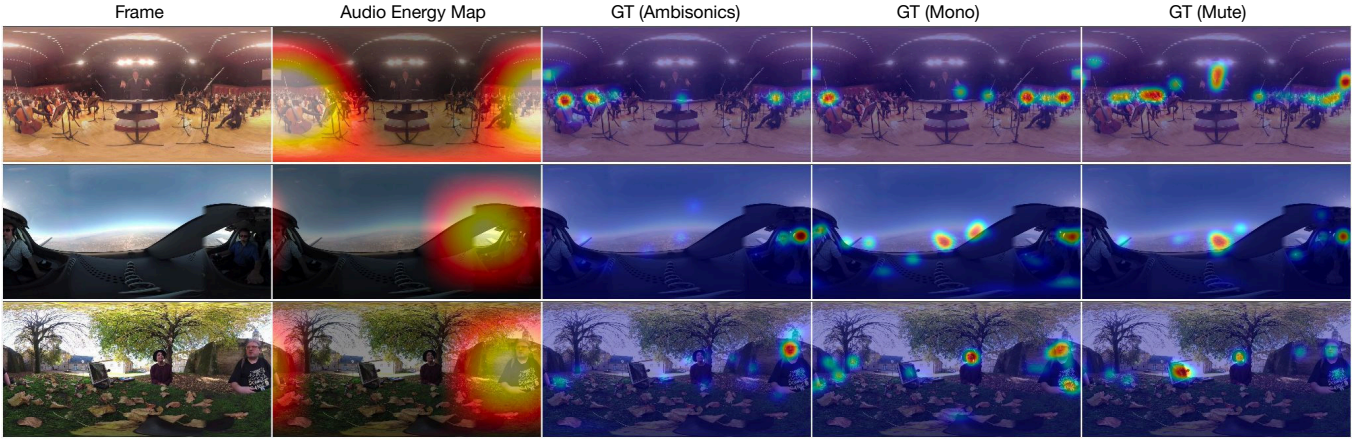


Fig. 5: **Fixation Density Maps Collected from Three Subject Groups Under Mute, Mono, and Ambisonic Audio Conditions in Our Dataset.** The analysis underscores a progressive decrease in correlation between the fixation density maps (ground-truth saliency maps) and audio energy maps, moving from ambisonics to mute settings. This trend distinctly illustrates the influence of spatial audio on guiding participants' attention.

order effects. Consistency in the viewing experience was maintained by standardizing the initial viewing angle, resetting the participant's yaw rotation to 0 degrees at the onset of each stimulus presentation.

Apparatus. We used the HTC Vive Pro-Eye[2] headset with a Tobii Eye Tracker[3] for precise eye movement tracking. The virtual viewing environment was engineered in Unity3D[4]. To facilitate free movement and rotation without cable entanglement, the headset cable was connected to a ceiling-mounted extensible and adjustable mechanism.

Participants. The study engaged 102 participants (53 females and 49 males), aged 18 to 33, predominantly affiliated with Boğaziçi University, Istanbul, TR. All had normal or corrected-to-normal vision, with corrections via contact lenses or glasses where necessary. Participant anonymity and data confidentiality were upheld through the use of randomly assigned IDs. The research protocol was approved by the Boğaziçi University Ethics Coordinating Committee[5], ensuring ethical compliance. Informed consent was obtained from all participants, who were briefed on the study's procedures and their right to withdraw at any time.

3.3 Post-processing

Initially, the experiment records were synchronized with the corresponding tracking data, providing gaze and head direction details for each stimulus per participant. Fixations were extracted using the I-DT algorithm, as described in [58]. Fixation parameters included a maximum dispersion of 1.5 visual angles and a minimum duration of 0.1 seconds. These fixation characteristics were then saved in comma-separated value text files. The chosen thresholds considered both the device's accuracy range of 0.5-1.1 degrees and established guidelines for I-DT thresholds [59, 60, 61, 62].

2. https://vive.com/sea/product/vive-pro-eye/overview/
3. https://tobii.com/products/integration/xr-headsets/
4. https://unity.com/
5. https://bogazici.edu.tr/en_US/Content/About_BU/Governance/Councils_Boards_and_Committees/Ethics_Committees

4 Methodology

In this section, we first describe SalViT360, our vision-transformer-based 360° video saliency model. Then, we present SalViT360-AV, the unified audio-visual saliency prediction pipeline for 360° videos. SalViT360-AV is a two-stream architecture for audio and visual modalities. Both streams are pre-trained and frozen, and we introduce transformer adapters conditioned on audio features for parameter-efficient fine-tuning of the video stream.

4.1 Spherical Vision Transformers for 360° Video Saliency Prediction

In Fig. 6, we present the overview of SalViT360. We start with gnomonic projection [14] to obtain tangent images for each frame in the input video clip. The tangent images are passed to an encoder-transformer-decoder architecture. The image encoder is used to extract local features for each tangent viewport and reduce 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 enriched spatial representations. The transformer utilizes our proposed Viewport Spatio-Temporal Attention (VSTA) to capture inter and intra-frame information across tangent viewports in a temporal window of $T = 8$ frames. 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 leverage large-scale pre-trained 2D models (e.g., ResNet-18) for feature extraction while the rest of the network is trained from scratch.

Gnomonic Projection and Encoder. We project the input ERP clip $x \in \mathbb{R}^{F\times 3\times H\times W}$ to a set of tangent clips $\{x_t \in \mathbb{R}^{F\times 3\times p\times p}\}_{t=1}^{T}$, with F, 3, H, and W indicating the number

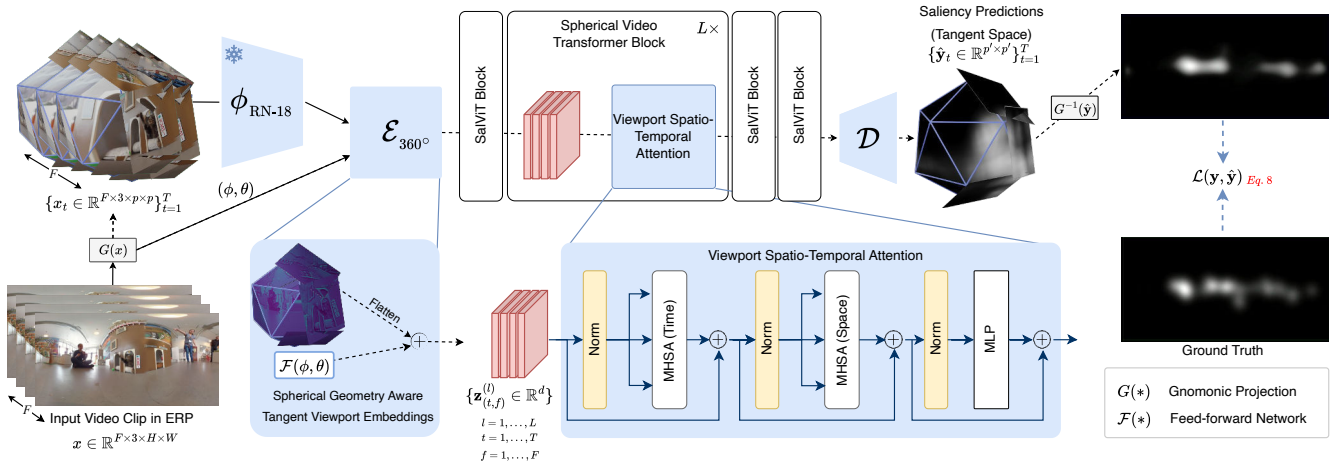


Fig. 6: **Overview of SalViT360.** The ERP video clip of F frames (1) is projected to $F \times T$ tangent images per set (2). Each tangent image is encoded and fused with spherical-geometry-aware position embeddings (3) for the 360° video transformer to aggregate global information (4). The outputs 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 $\mathcal{L}_{\text{VAC}}(P,\ P')$ to minimize the tangent artifacts (7). During test time, the model works with a single tangent set. For simplicity, only one set of tangent images is shown.

of frames, channel dimension (RGB), height, and width of the video, respectively. These tangent images have a patch size of $p \times p = 224 \times 224$ pixels. We select $T = 18$ tangent images per frame and a 80° field-of-view based on [63]. After downsampling and flattening the encoder features, we obtain tangent feature vectors $\{z_t \in \mathbb{R}^{F\times d}\}_{t=1}^{T}$. In a parallel stream, we map the angular coordinates (ϕ, θ) for each pixel of the tangent viewports to the feature dimension $d = 512$ using an FC layer and element-wise add these embeddings with encoder features to obtain the proposed spherical geometry-aware tangent embeddings $\{\mathbf{z}_t \in \mathbb{R}^{F\times d}\}_{t=1}^{T}$ that are used in the transformer.

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 understanding. 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, thus the computational complexity, inspired by the divided space-time attention mechanism in TimeSformer [40], 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 $F^2 + T^2$, effectively capturing the required global context for 360° video analysis. Our Viewport Spatio-Temporal Attention for 360° videos illustrated in Fig. 7 is defined as:

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

where $\mathbf{z}_{(t,f)}^{(l)} \in \mathbb{R}^d$ denote the tangent features 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.

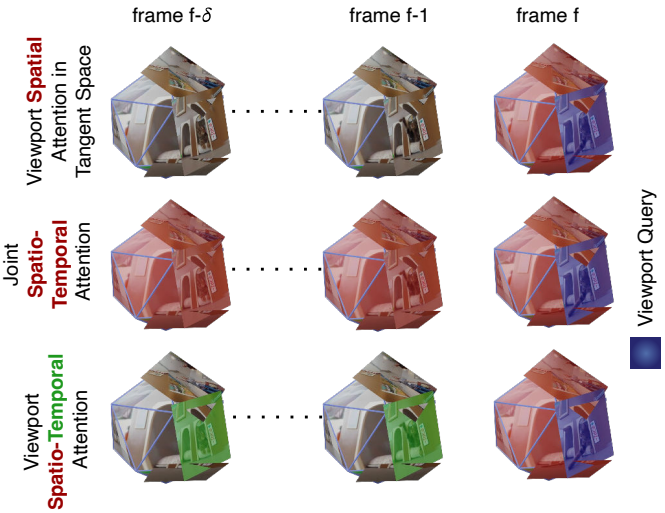


Fig. 7: **Viewport Spatio-Temporal Attention (VSTA)** (right), compared to Viewport Spatial Attention (VSA) (left) and Joint Spatio-Temporal Attention (middle) approaches. Red and Green viewports denote the self-attention neighborhood for each scheme.

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 $\{\hat{\mathbf{z}}_t \in \mathbb{R}^{512\times 7\times 7}\}_{t=1}^{T}$ of the last frame as input and outputs saliency prediction $\{\hat{\mathbf{y}}_t \in \mathbb{R}^{p'\times p'}\}_{t=1}^{T}$ on tangent planes. The final ERP saliency maps are obtained by passing the tangent predictions to inverse gnomonic projection. While we aggregate global

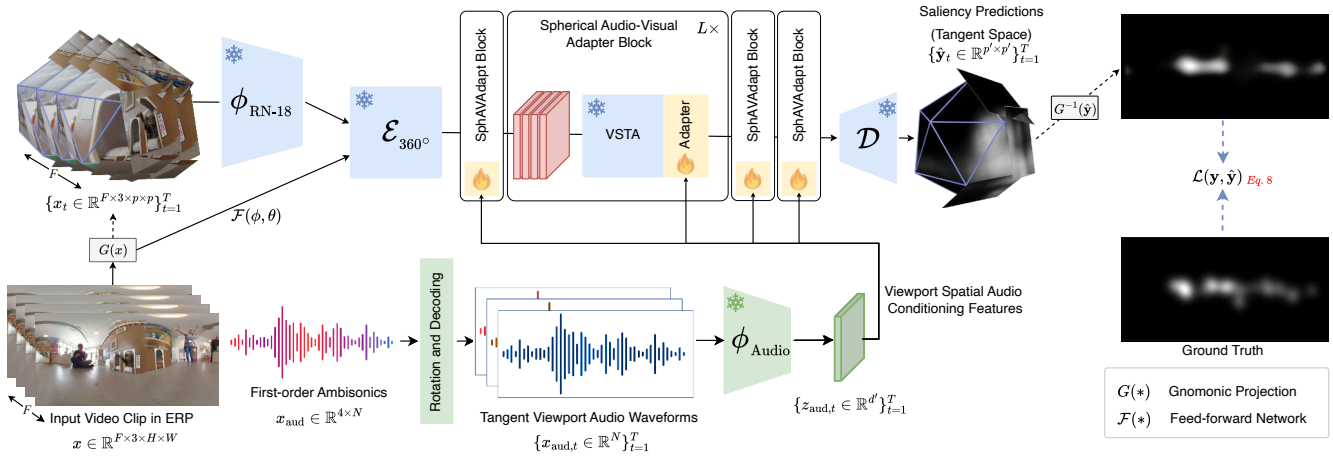


Fig. 8: **Overview of Our Proposed SalViT360-AV Pipeline.** We use the SalViT360 model as the video saliency module (top), and our implementation allows for any audio model as the audio backbone (bottom). The audio stream takes input spatial audio waveforms $x_{\text{aud}} \in \mathbb{R}^{4\times N}$ encoded as 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 tangent viewport $\{x_t\}_{t=1}^{T}$ (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 video pipeline is $37M$, the additional adapter layers require only $600k$ parameters for fine-tuning.

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 [63] proposes iterative refining to compensate for these artifacts, 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.

Viewport Augmentation Consistency. Our model learns saliency distribution with a supervised loss but faces discrepancies in ERP saliency maps due to separate predictions for each tangent plane. To address this, we propose an unsupervised *Viewport Augmentation Consistency* (VAC) loss to enhance prediction consistency across tangent projections. 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 appendix. 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:

$$\mathcal{L}_{\text{VAC}}(\hat{\mathbf{y}},\, \bar{\mathbf{y}}) = \mathcal{L}_{KLD}^{\text{weighted}}(\hat{\mathbf{y}},\, \bar{\mathbf{y}}) + \mathcal{L}_{CC}^{\text{weighted}}(\hat{\mathbf{y}},\, \bar{\mathbf{y}}), \quad (2)$$

$$\mathcal{L}_{\text{KLD}}^{\text{weighted}}(\hat{\mathbf{y}},\, \bar{\mathbf{y}}) = \textstyle\sum_{i,j} \hat{\mathbf{y}}_{i,j} \log\left(\epsilon + \frac{\hat{\mathbf{y}}_{i,j}}{\bar{\mathbf{y}}_{i,j}+\epsilon}\right) \cdot w_{i,j}, \quad (3)$$

$$\mathcal{L}_{\text{CC}}^{\text{weighted}}(\hat{\mathbf{y}},\, \bar{\mathbf{y}}) = 1 - \frac{\sum(\hat{\mathbf{y}}\cdot\bar{\mathbf{y}})\cdot w_{i,j}}{\sum(\hat{\mathbf{y}}\cdot\hat{\mathbf{y}})\cdot \sum(\bar{\mathbf{y}}\cdot\bar{\mathbf{y}})} \quad (4)$$

where $\hat{\mathbf{y}},\ \bar{\mathbf{y}}$ 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 appendix.

4.2 Spherical Vision-Transformer Adapters for 360° Audio-Visual Saliency

In our prior work, SalViT360, we successfully leveraged spatio-temporal features for understanding 360° scenes. However, SalViT360 solely relied on visual stimuli, overlooking the influential role of audio cues. Motivated by the strong impact of audio in guiding human visual attention, as discussed in Section 2, we have enhanced SalViT360 with spatial audio encoding capabilities, leading to the development of SalViT360-AV. As illustrated in Fig. 8, SalViT360-AV integrates two streams for jointly processing audio and video: the video branch (top) uses the original SalViT360 pipeline, while the audio branch (bottom) can flexibly incorporate any pre-trained audio backbone. It is important to note that when ambisonics are not available, the spatial audio stream in SalViT360-AV can be conveniently substituted with mono audio.

Audio Backbone. The audio stream receives spatial audio waveforms $x_{\text{aud}} \in \mathbb{R}^{4\times N}$ encoded as first-order ambisonics in 4-channel B-format containing audio streams (W, Y, Z, X). To simulate what a subject would hear while gazing in a specific direction, we rotate the ambisonics for each tangent viewport, defined by angular coordinates $\{(\theta^t, \phi^t)\}_{t=1}^{T}$, producing mono directional waveforms $\{x_t\}_{t=1}^{T}$. This is achieved by applying a spherical harmonics rotation transformation to the ambisonics channels [64, 65]. This approach enables the use of any pre-trained audio model such as PaSST [66], EZ-VSL [67] or AVSA [49] for

feature extraction. The resulting audio features $\mathbf{x}_{\text{aud}}^{t} \in \mathbb{R}^{\hat{d}}$ are then aligned with their corresponding visual tokens for each viewport.

In our experiments, we evaluated all three backbone models, and empirically selected PaSST as the main audio encoder. PaSST stands out with its self-attention mechanism on mel-spectrogram patches and its unique training approach, involving 'patchout', a dropout-like mechanism. Its training leverages DeiT's pre-trained weights from ImageNet, further adapted on AudioSet for enhanced audio classification capabilities. The model processes input mel-spectrograms, extracted using a window of 25ms and a hop length of 10ms, to produce feature vectors $\mathbf{x}_{\text{aud}}^{t} \in \mathbb{R}^{\hat{d}}$ for audio clips. The duration for these audio clips is set to 4 seconds, based on empirical observations.

During the development of our framework, we considered several alternatives to this ambisonics decoding strategy. One baseline fed the raw B-format channels (W, Y, Z, X), as either waveforms or mel-spectrograms, directly into the model. However, this configuration consistently underperformed compared to our viewport-based decoding. The core advantage of our adopted strategy is perceptual alignment: saliency in 360° scenes often relies on spatially localized audio cues, and decoding ambisonics into view-aligned signals exposes these cues in a geometry that directly matches the tangent visual representations. This alignment enables more effective audio–visual fusion, as features share a common spatial frame of reference.

We also experimented with learning a mapping from four global channels to eighteen spatial bins through an additional trainable module such as a cross-attention mapper. While this partially addressed the alignment issue, it increased both model size and inference time, without matching the efficiency or predictive accuracy of the viewport-based strategy. As a result, we selected the current design, which ensures a direct and efficient one-to-one correspondence between audio and visual streams and allows flexible integration of state-of-the-art audio backbones. Further implementation details and a comparative analysis of backbone choices are provided in Appendix B.

Adapter Architecture. As shown in Fig. 9, SalViT360-AV uses a parallel transformer adapter structure inspired by AdaptFormer [18], with a frozen Viewport Spatio-Temporal Attention (VSTA) block from SalViT360 and a lightweight bottleneck adapter for audio-visual fusion. The model processes audio tokens $\{z_{\text{aud}} \in \mathbb{R}^{\hat{d}}\}_{t=1}^{T}$ alongside visual tokens $\{\mathbf{z}_{(\text{t, f})}^{(l)} \in \mathbb{R}^{d}\}_{(t=1,f=1)}^{(T,F)}$. This dual-stream setup involves a stack of down-projection and up-projection layers, interconnected with ReLU non-linearity, functioning as a bottleneck with significantly lower dimensions than the original embeddings. Specifically, this stream is designed as a bottleneck module with a dimension of k, as the projections are performed by parameters $\mathbf{W}_{down} \in \mathbb{R}^{(d+\hat{d})\times k}$, and $\mathbf{W}_{up} \in \mathbb{R}^{k\times\hat{d}}$, where $k \ll d, \hat{d}$. The concatenated audio-visual embeddings $\{\bar{z}_{\text{av, t}} \in \mathbb{R}^{d+\hat{d}}\}_{t=1}^{T}$ for each tangent viewport undergo a transformation to generate audio-conditioning features $\hat{z}_{\text{av, t}}$ as:

$$\hat{z}_{\text{av, t}} = \text{ReLU}(\text{LN}(\bar{z}_{\text{av, t}}) \cdot \mathbf{W}_{down}) \cdot \mathbf{W}_{up} \quad , \tag{5}$$

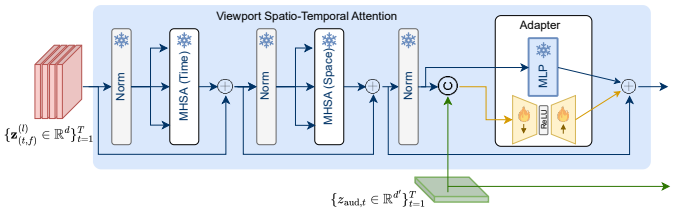


Fig. 9: **Audio-Visual Adapter Fine-Tuning of SalViT360-AV.** For each of the L transformer blocks in the frozen SalViT360 pipeline, the visual tokens after Viewport Spatio-Temporal Attention are concatenated with audio tokens. The pre-trained MLPs at the end of each VSTA block are upgraded with **Audio-Visual MLP Adapters**, which consist of a scaled combination of the frozen MLP and trainable bottleneck module. F, T, D denote the number of frames and tangent viewports tokens, clip length, and the embedding dimension, respectively.

which are then fused with the visual features $\bar{z}_{\text{av, t}}$ to enhance the saliency prediction capability of the model:

$$\mathbf{z}_{\text{av, t}} = \text{MLP}(\bar{z}_{\text{av, t}}) + s \cdot \hat{z}_{\text{av, t}} + \bar{z}_{\text{av, t}} \tag{6}$$

Our adapter-based design ensures that when no audio input is provided such as in purely visual 360° saliency datasets, the audio branch transmits no signal to the adapter layers. In this case, SalViT360-AV behaves identically to SalViT360, and the model effectively reduces to single-modal visual saliency prediction without any loss of performance or the need for re-training. This flexibility allows our framework to generalize seamlessly across both single-modal (visual-only) and multi-modal (audio-visual) saliency prediction tasks.

5 Experiments

Our experimental evaluation was structured in a two-fold approach. In the first part, focusing on vision-only settings, we trained the SalViT360 model using the VR-EyeTracking dataset [25]. This dataset comprises 134 videos for training and 74 videos for testing. For validation, we randomly selected a subset of 30 videos from the training set and assessed the performance of SalViT360 on the test set. Additionally, to evaluate the model's generalization ability, we conducted cross-dataset evaluations using the PVS-HMEM [23], 360AV-HM [28] datasets, and the test split of our YT360-EyeTracking dataset. These datasets include 76, 21, and 27 videos, respectively, and each were viewed by 58, 15, and 15 subjects. In the second part of our experimentation, we focused on audio-visual tuning of the SalViT360-AV model. We began by utilizing the pre-trained weights of SalViT360 from the VR-EyeTracking dataset, which we kept frozen throughout this phase. The training of SalViT360-AV was conducted using the training split of the YT360-EyeTracking dataset. We then evaluated SalViT360-AV's performance on various datasets, including the test set of VR-EyeTracking, the YT360-EyeTracking test split, and the 360AV-HM dataset.

Data Pre-processing. The audio waveforms are resampled to a frequency of 32.0 kHz to align with the PaSST backbone. The duration of audio clips is empirically chosen

to be a maximum of 4 seconds. This choice is based on discussions in [47], which suggest that shorter clips tend to capture short-term audio saliency, while longer clips may encompass more than local saliency. A detailed comparative analysis of clip durations can be found in the Appendix.

After rotating the 4-channel ambisonics for each viewport, we decode them to extract a single-channel directional waveform F for mono audio backbones (like PaSST and EZ-VSL), using the formula:

$$F = (\sqrt{2}W + X) \cdot 2. \tag{7}$$

Evaluation Metrics and Loss Functions. We evaluate the performance of the models using four commonly used metrics, namely, Normalized Scanpath Saliency (NSS), KL-Divergence (KLD), Pearson's Correlation Coefficient (CC), and Similarity Metric (SIM). We compute the training objective as a weighted differentiable combination of KLD, CC, and Selective-MSE (MSE on normalized saliency maps at only eye-fixation points [68]) as given below:

$$\begin{aligned} \mathcal{L}(P, Q_s, Q_f) &= \text{KLD}(P, Q_s) + \text{CC}(P, Q_s) \\ &+ \alpha\, \text{SMSE}(P, Q_s, Q_f) \end{aligned} \tag{8}$$

where $P,\ Q_s,\ Q_f$ are the predicted saliency, ground truth density and fixation maps, respectively, and $\alpha = 0.05$.

Architecture and Optimization Details. We trained SalViT360 for 100K iterations using AdamW [69] with momentum parameters $(\beta_1,\ \beta_2) = (0.9,\ 0.999)$, a weight decay of $1e{-}2$ and an initial learning rate of $1e{-}5$. A cosine scheduler was applied, decaying the lr to $2e-6$ over the course of training. The batch size was set to 16, and all experiments were conducted on a single Tesla V100 GPU with 32 GB of memory. Subsequently, we performed parameter-efficient fine-tuning of SalViT360-AV for an additional 20,000 steps using the same optimizer with a fixed learning rate of 2e-6.

5.1 Empirical Evaluation Across Visual and Audio-Visual Saliency Benchmarks

We conduct an extensive evaluation of our proposed SalViT360 and SalViT360-AV models, systematically examining their performance across a diverse set of benchmarks. Our analysis is structured into three main parts: (1) evaluation on datasets without spatial audio, which primarily assess visual saliency models; (2) assessment on our proposed YT360-EyeTracking dataset to evaluate the benefits of integrating spatial audio; and (3) cross-dataset generalization on unseen spatial audio-visual datasets.

Evaluation on Datasets without Spatial Audio. We first evaluate the effectiveness of our visual transformer model SalViT360 using the widely adopted VR-EyeTracking dataset [25], which, despite containing mono audio, lacks spatial audio annotations and thus serves primarily as a visual-only benchmark. VR-EyeTracking provides an extensive test scenario, comprising 134 training and 74 testing videos. To comprehensively assess the model's generalization beyond training data, we further evaluate its performance on the PVS-HMEM dataset [23], which features purely visual annotations without any audio content. Quantitative results provided in Table 2 clearly demonstrate that

TABLE 2: **Quantitative Evaluation on VR-EyeTracking and PVS-HMEM.** Although the VR-EyeTracking dataset contains mono audio, it lacks spatial audio cues and is thus treated as a visual saliency benchmark. The PVS-HMEM dataset does not include sound. The models are grouped by category using background color: blue for 360° visual-only, orange for 2D audio-visual, and green for 360° audio-visual saliency models. **Bold** scores indicate the best performance, while underlined scores denote the second best.

	VR-EyeTracking				PVS-HMEM			
Method	**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
ATSal	1.317	12.259	0.336	0.318	0.732	4.303	0.183	0.219
SalGAN360	1.753	10.845	0.370	0.355	1.513	4.394	0.314	0.291
MV-SalGAN360	1.818	8.713	0.382	0.357	1.546	4.112	0.316	0.295
PAVER	1.511	13.267	0.307	0.294	0.750	3.736	0.224	0.269
Djilali et al.	**3.183**	6.570	0.565	0.475	1.688	2.430	0.447	0.404
SalViT360	2.630	**5.744**	**0.586**	**0.492**	**2.191**	**1.841**	**0.626**	**0.495**
DAVE	1.821	12.199	0.331	0.304	1.644	4.252	0.309	0.273
STAViS	1.642	13.148	0.316	0.297	1.539	4.320	0.306	0.266
DiffSal	2.599	5.979	0.566	0.473	1.926	2.539	0.443	0.399
SalViT360-AV	**2.821**	**5.334**	**0.599**	**0.511**	**2.191**	**1.841**	**0.626**	**0.495**

SalViT360 achieves top-tier performance, frequently surpassing existing state-of-the-art visual-only saliency models. In particular, our transformer-based architecture shows substantial advantages in capturing complex visual contexts and accurately predicting viewer fixation patterns.

Regarding the audio-visual evaluation on these datasets, we note that 360° audio-visual saliency models, apart from our SalViT360-AV, could not be directly evaluated on the PVS-HMEM dataset, as it entirely lacks audio data. Moreover, they were incompatible with the VR-EyeTracking dataset, as they require spatial audio to process. On the other hand, our SalViT360-AV model can effectively process mono audio waveforms or even no audio (see our discussion at the end of Sec. 4.2); thus, we can provide results on both datasets. As expected, in the presence mono audio, SalViT360-AV demonstrate improved predictive accuracy compared to existing 2D audio-visual models on the VR-EyeTracking dataset. As shown in Fig. 10, SalViT360 and SalViT360-AV generate saliency maps that accurately localize human gaze across the provided panoramic scenes. Our models reliably focus on the most visually salient regions, even in cases where attention is narrowly concentrated. In contrast, previous visual-only 360° models as well as 2D audio-visual models often produce saliency maps that are either overly blurred or misplace attention, highlighting irrelevant parts of the scene. Our models are less prone to center or equator bias and more reliably highlight peripheral yet visually relevant regions, yielding predictions that are much closer to the ground truth fixations.

Effectiveness of Integrating Spatial Audio. Next, we systematically investigate the benefit of spatial audio integration. Leveraging our newly introduced YT360-EyeTracking dataset, which explicitly captures viewer attention under spatial audio conditions, we fine-tune our audio-visual SalViT360-AV model, initialized with weights from SalViT360 pretrained on VR-EyeTracking. Our transformer-based audio adapters effectively incorporate ambisonic spa-

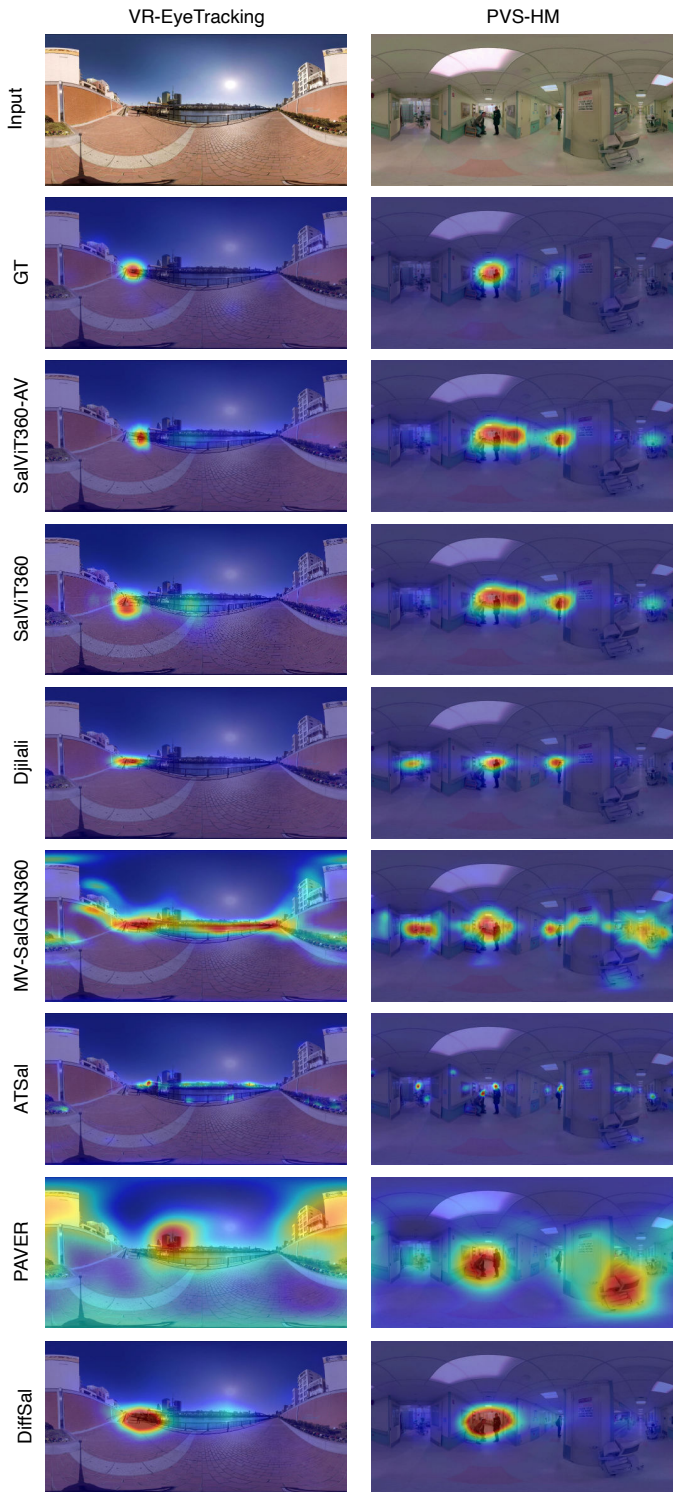


Fig. 10: **Qualitative Comparison of Predicted Saliency Maps on Sample Frames from VR-EyeTracking and PVS-HMEM.** Both SalViT360 and SalViT360-AV accurately capture human gaze patterns, effectively highlighting salient cues in complex panoramic scenes. Competing visual-only models and the 2D audio-visual DiffSal model often miss or oversmooth key regions.

tial audio cues into the visual backbone. Quantitative outcomes detailed in Table 3 highlight that SalViT360-AV consistently outperforms visual-only and state-of-the-art 2D

TABLE 3: **Evaluation on YT360-EyeTracking to Assess the Impact of Spatial Audio Integration.** We compare visual-only, 2D audio-visual, 360° audio-only, and 360° audio-visual saliency models. Categories are color-coded: blue for 360° visual, orange for 2D audio-visual, green for 360° audio-visual, and gray for the 360° audio-only models.

Method	NSS↑	KLD↓	CC↑	SIM↑
CP-360	1.041	22.356	0.120	0.118
ATSal	1.711	13.948	0.242	0.244
SalGAN360	1.415	14.412	0.236	0.239
MV-SalGAN360	1.433	14.381	0.244	0.246
PAVER	1.093	14.830	0.254	0.226
Djilali et al.	1.738	11.085	0.369	0.326
SalViT360	**2.346**	**9.861**	**0.484**	**0.373**
DAVE	2.290	8.858	0.462	0.363
STAViS	1.535	12.039	0.343	0.294
DiffSal	2.293	8.458	0.477	0.383
SSSL	1.212	13.398	0.214	0.244
AVS360	2.170	**7.801**	0.457	0.374
SVGC-AVA	2.191	8.860	0.488	0.399
SalViT360-AV (w/ raw FOA)	**2.443**	8.539	**0.508**	**0.406**
SalViT360-AV (proposed)	**2.449**	8.341	**0.512**	**0.407**

audio-visual models across all metrics. Notably, the results emphasize the significant advantage of our audio-adaptive strategy, clearly demonstrating its ability to harness spatial audio information effectively. As shown in Fig. 11, SalViT360-AV stands out in aligning attention with salient audio-driven events such as off-screen conversations, approaching vehicles, or spatialized environmental sounds. Compared to both visual-only and other 2D/360° audio-visual baselines, SalViT360-AV's maps are more precise, especially in challenging situations where salient cues are strongly dictated by audio. Competing methods tend to produce fragmented or overly diffused predictions under such conditions, failing to adapt their attention when the salient event is not visually dominant but audibly salient. In addition, we evaluate a model variant that directly uses raw FOA (4-channel B-format) signals without applying our proposed ambisonics rotation and decoding. As shown in Table 3, this baseline (SalViT360-AV w/ raw FOA) delivers competitive results; however, SalViT360-AV consistently outperforms it across all metrics. We attribute this improvement to two key factors: (1) while FOA channels offer a compact representation, they do not explicitly encode directional information, whereas decoding into 18 directional waveforms more closely replicates how spatial cues are perceived in VR environments; and (2) using only four channels in the cross-attention integration reduces the size of the attention matrix, leading to coarser audio-visual correspondences. In contrast, our decoded directional signals support finer, view-specific alignments. These findings confirm the importance of spatially aligning audio inputs prior to fusion for more accurate and robust saliency prediction.

Cross-Dataset Generalization under Spatial Audio. Finally, we examine the robustness and generalizability of our approach by evaluating the fine-tuned SalViT360-AV model on three additional datasets containing spatial audio cues: 360AV-HM[28], AVS-ODV [29], and SVGC-AVA [26]. These datasets vary substantially in content complexity, audio-visual dynamics, and participant viewing conditions, thus providing rigorous tests of generalization. As summarized

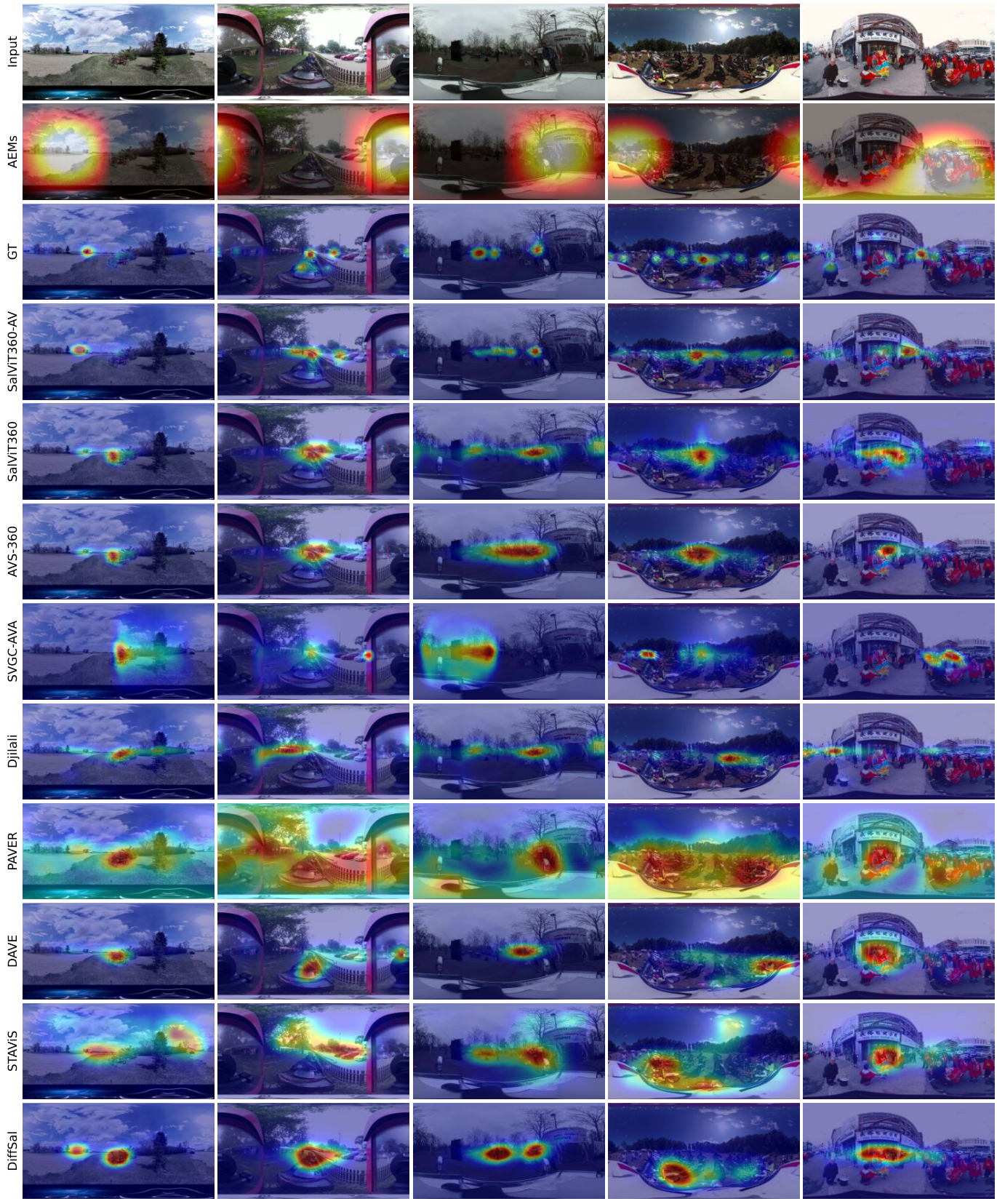


Fig. 11: **Predicted Saliency Maps Illustrating the Integration of Spatial Audio Cues in the YT360-EyeTracking Dataset.** SalViT360-AV clearly leverages ambisonic sound to pinpoint visually salient regions strongly influenced by audio events (e.g., approaching vehicles, off-screen conversations, and environmental sounds). Compared to visual-only or 2D audio-visual baselines, our model produces richer and contextually accurate attention maps aligned closely with human fixations.

TABLE 4: **Cross-Dataset Generalization Results on 360AV-HM, AVS-ODV, and SVGC-AVA.** All datasets include ambisonic spatial audio. The models are grouped by category using background color: blue for 360° visual saliency, orange for 2D audio-visual saliency, gray for the 360° audio-only saliency, green for 360° audio-visual saliency prediction.

	360AV-HM				AVS-ODV				SVGC-AVA			
Method	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
ATSal	1.322	**14.141**	0.156	0.121	1.430	13.159	0.233	0.229	1.510	11.886	0.314	0.311
PAVER	1.019	19.136	0.176	0.115	1.619	13.252	0.251	0.241	1.641	11.302	0.422	0.396
Djilali et al.	2.050	16.618	0.308	0.209	2.014	10.167	0.344	0.300	2.020	10.239	0.511	0.419
SalViT360	**2.285**	15.879	**0.349**	**0.246**	2.192	**8.626**	**0.390**	**0.321**	2.179	**9.320**	**0.557**	**0.434**
DAVE	2.178	14.693	0.344	0.256	2.050	12.065	0.341	0.279	1.997	10.451	0.446	0.401
STAViS	1.479	17.934	0.237	0.162	2.012	12.195	0.340	0.277	1.991	10.543	0.443	0.400
DiffSal	2.460	13.991	0.376	0.271	2.368	11.866	**0.398**	0.317	2.276	**8.274**	0.541	0.408
SSSL	1.181	14.192	0.203	0.184	1.019	13.225	0.198	0.184	1.190	11.469	0.211	0.187
AVS360	**2.501**	**13.581**	**0.380**	**0.290**	2.386	**8.130**	0.394	0.332	2.320	9.245	0.559	0.433
SVGC-AVA	2.448	14.012	0.367	0.266	2.387	8.379	0.394	**0.334**	**2.371**	9.241	0.560	0.434
SalViT360-AV	2.473	13.830	0.379	0.278	**2.389**	8.481	**0.397**	0.332	2.342	9.251	**0.564**	**0.440**

TABLE 5: **Ablation Study of SalViT360 Components on VR-EyeTracking.** We evaluate the impact of core architectural components. Replacing 1D positional embeddings with spherical ones improves performance. Introducing our VSTA module yields further gains, and the VAC module—with masking—achieves the best results, demonstrating the benefit of spatial priors and temporal context.

Method	**# params**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
VSA (w/ 1D Pos. Emb.)	18.76M	2.518	6.445	0.560	0.472
+ Spherical Pos. Emb.	19.26M	2.575	6.221	0.563	0.475
VSTA (w/ Sph. Pos. Emb.)	25.56M	**2.664**	6.174	0.570	0.479
+ VAC (w/o mask)	25.56M	2.624	6.011	0.576	0.490
+ VAC (w/ mask)	25.56M	2.630	5.744	0.586	0.492

in Table 4, our model consistently achieves top-level performance, outperforming or closely matching recent state-of-the-art 360° audio-visual methods.[6] These results confirm the broad adaptability of SalViT360-AV to unseen datasets characterized by complex spatial audio environments. The qualitative results in Fig. 12 further highlight the robustness and adaptability of SalViT360-AV. Even when tested on datasets it was not trained on, our model consistently identifies salient audio-visual regions. For example, in the AVS-360 scene where a vehicle passes behind the camera, SalViT360-AV accurately shifts attention to the rear-facing tangent viewport, unlike competing models that either miss the cue or overemphasize the front-facing view. Similarly, in the SVGC-AVA clip with multiple sound sources, our model selectively highlights the dominant auditory event while suppressing distractions. These examples demonstrate that SalViT360-AV effectively understands subtle interactions between auditory and visual modalities. Unlike earlier methods, which tend to blur attention or focus on visually prominent but semantically irrelevant areas, our model produces sharper, context-aware predictions that align with human-like attention shifts across diverse content types.

6. We note that the AVS360 model is trained on 360AV-HM dataset, and, thus, obtains the highest scores on the test set of 360AV-HM.

TABLE 6: **Comparison of spatio-temporal modeling strategies.** We compare our VSTA mechanism against a 2+1D CNN and offline EMA aggregation. VSTA outperforms both alternatives across all metrics, confirming the advantage of end-to-end learnable spatio-temporal attention in 360° video saliency prediction.

Method	**# params**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
VSA + 2+1D-CNN Enc.	17.31M	2.568	**5.915**	0.568	0.477
VSA + Offline EMA	19.27M	2.591	6.018	0.566	0.477
VSTA	25.56M	**2.664**	6.174	**0.570**	**0.479**

5.2 Ablation Studies

To provide a more comprehensive analysis of each component in our audio-visual saliency pipeline, we perform experiments on the validation splits of VR-EyeTracking and YT360-EyeTracking datasets, for our SalViT360 and SalViT360-AVmodels, respectively.

Spherical Position Embeddings. We compare the performance of our proposed *spherical geometry-aware spatial position embeddings* with regular 1D learnable position embeddings in Table 5. 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.

Spatio-Temporal modelling. In Table 6, we compare Viewport Spatial Attention (VSA) and Viewport Spatio-Temporal Attention (VSTA) blocks to assess the contribution of temporal information processing in omnidirectional videos. We conducted several experiments to assess the effectiveness of our proposed VSTA mechanism. We include two distinct approaches, 2+1D-CNN [49] backbone and Offline EMA in this analysis. 2+1D-CNN backbone is an R2+1D model [70] which performs convolution over consecutive frames, pretrained 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. We also exper-

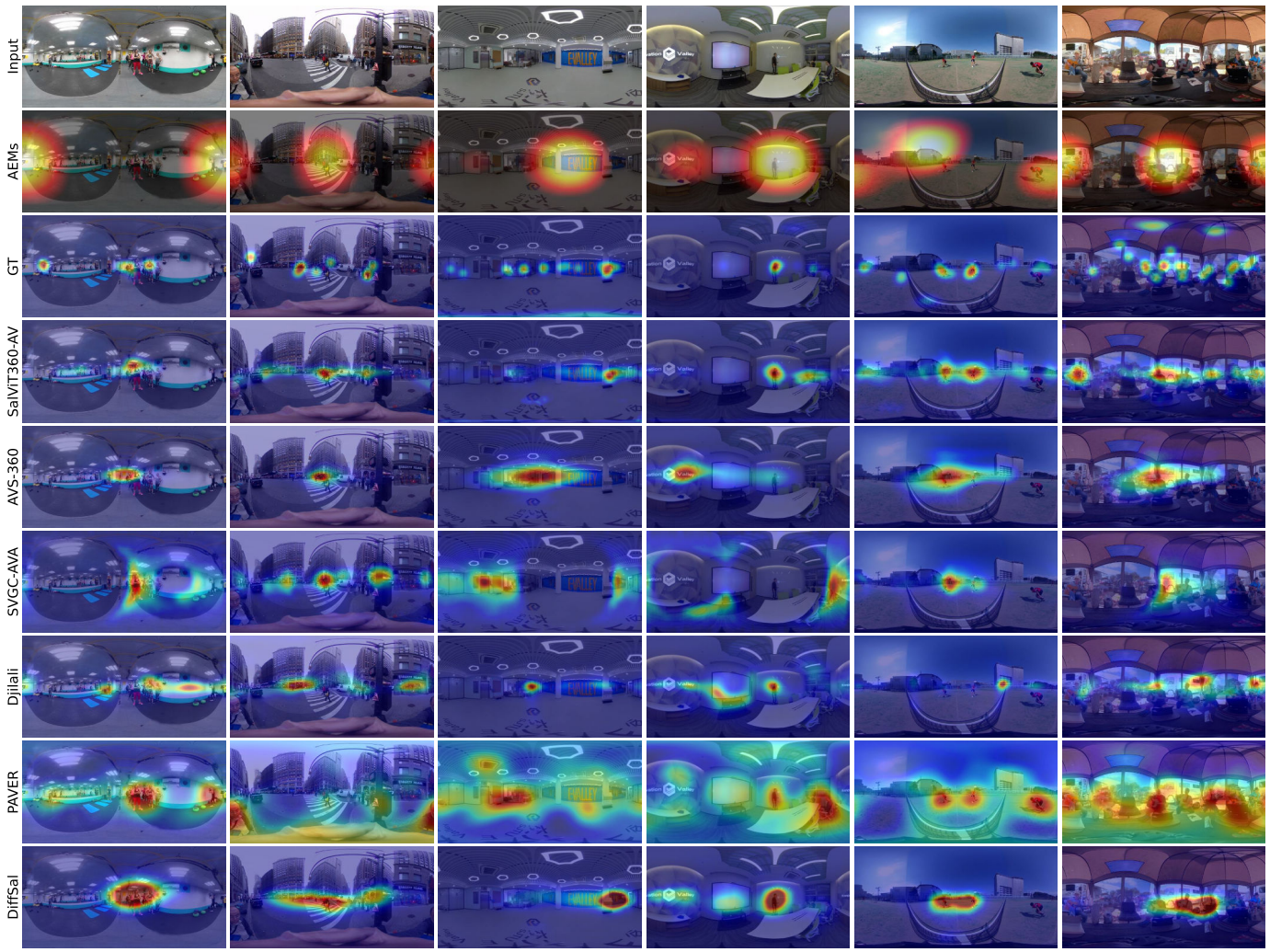


Fig. 12: **Cross-Dataset Generalization Results on 360AV-HM, AVS-ODV, and SVGC-AVA Datasets.** Qualitative saliency predictions on sample frames from 360AV-HM (columns 1–2), AVS-ODV (columns 3–4), and SVGC-AVA (columns 5–6) datasets. These datasets span diverse indoor and outdoor scenes with varying spatial audio complexity. Our model consistently localizes salient regions influenced by spatial audio cues, outperforming recent state-of-the-art models.

iment on transformer depth, which shows that our VSTA blocks gradually learn better spatio-temporal representations in deeper layers. Our results show that the proposed VSTA mechanism outperforms the spatial-only setting and the other two spatio-temporal approaches. We refer the reader to the appendix for comprehensive experiments on the transformer depth, and temporal window size F, along with a comparison of joint spatio-temporal attention.

Impact of Ground-Truth Audio Modality The saliency datasets having spatial audio are collected under (1) muted, (2) mono, and (3) ambisonics audio conditions, each audio modality viewed by a different subject group to prevent biases. We evaluate our video-only baseline on each ground truth separately. The results in Table 7 demonstrate that the increasing representative power of audio modality yields higher inter-subject correlation on our video-only model.

TABLE 7: **Effect of audio modality in ground truth saliency.** We evaluate SalViT360 using saliency maps collected under mute, mono, and ambisonics audio. Performance improves with richer audio conditions, highlighting how spatial audio influences human attention—even when the model is trained without audio.

	360AV-HM				YT360-EyeTracking			
Audio	**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**

6 Conclusion

In this study, we introduced SalViT360and SalViT360-AV. two models for predicting saliency in 360° video and audio-visual contexts. SalViT360, leverages an encoder-transformer-decoder architecture that processes undistorted tangent image representations using a spatio-temporal attention mechanism and spherical geometry-aware position embeddings. Evaluations on four benchmark datasets, supported by ablation studies, confirm consistent improvements over state-of-the-art methods.

To model the auditory dimension of attention, SalViT360-AV extends SalViT360 with a parameter-efficient adapter fine-tuning strategy, frozen pre-trained backbones, and a spatial audio pipeline that decodes and rotates ambisonics per viewport. This integration significantly boosts performance on datasets with audio–visual content.

Beyond saliency prediction, both models have broader applications, including improving audio-visual quality assessment [71, 72, 73], automatically generating viewing paths for immersive navigation [74], and guiding compression to prioritize perceptually important regions [75].

ACKNOWLEDGMENT

This work was supported in part by the KUIS AI Center Research Award, Unvest R&D Center, TÜBİTAK-1001 Program (No. 120E501), the TÜBA-GEBİP 2018 Award to E. Erdem, and the BAGEP 2021 Award to A. Erdem.

REFERENCES

[1] Shengxi Li, Mai Xu, Yun Ren, and Zulin Wang. Closed-form optimization on saliency-guided image compression for hevc-msp. *IEEE Trans. Multimed.*, 20(1):155–170, 2018.

[2] Dipti Mishra, Satish Kumar Singh, Rajat Kumar Singh, and Divanshu Kedia. Multi-scale network (mssg-cnn) for joint image and saliency map learning-based compression. *Neurocomputing*, 460:95–105, 2021.

[3] Shiping Zhu, Chang Liu, and Ziyao Xu. High-definition video compression system based on perception guidance of salient information of a convolutional neural network and hevc compression domain. *IEEE Trans. Circuits Syst. Video Technol.*, 30(7):1946–1959, 2020.

[4] Shiping Zhu and Ziyao Xu. Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275:511–522, 2018.

[5] Zejiang Hou and Sun-Yuan Kung. Multi-dimensional dynamic model compression for efficient image super-resolution. In *Proc. IEEE/CVF WACV*, pages 633–643, 2022.

[6] Jingwei Guan, Shuai Yi, Xingyu Zeng, Wai-Kuen Cham, and Xiaogang Wang. Visual importance and distortion guided deep image quality assessment framework. *IEEE Trans. Multimed.*, 19(11):2505–2520, 2017.

[7] Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang. SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *Proc. ACM Int. Conf. Multimedia*, pages 1383–1391, 2019.

[8] Mengmeng Zhu, Guanqun Hou, Xinjia Chen, Jiaxing Xie, Haixian Lu, and Jun Che. Saliency-guided transformer network combined with local embedding for no-reference image quality assessment. In *Proc. IEEE/CVF ICCV*, pages 1953–1962, 2021.

[9] Miaomiao Qiu and Feng Shao. Blind 360-degree image quality assessment via saliency-guided convolution neural network. *Optik*, 240:166858, 2021.

[10] Nafiseh Jabbari Tofighi, Mohamed Hedi Elfkir, Nevrez Imamoglu, Cagri Ozcinar, Erkut Erdem, and Aykut Erdem. ST360IQ: No-Reference Omnidirectional Image Quality Assessment with Spherical Vision Transformers. In *Proc. IEEE ICASSP*, 2023.

[11] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph.*, 35(6):1–12, 2016.

[12] Mert Cokelek, Nevrez Imamoglu, Cagri Ozcinar, Erkut Erdem, and Aykut Erdem. Spherical vision transformer for 360-degree video saliency prediction. In *BMVC*, 2023.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proc. IEEE CVPR*, pages 12426–12434, 2020.

[15] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *Proc. IEEE CVPR*, pages 4766–4776, 2020.

[16] Hamed R. Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *CoRR*, abs/1905.10693, 2019.

[17] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Trans. Image Process.*, 29:3805–3819, 2020.

[18] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Proc. NeurIPS*, 35:16664–16678, 2022.

[19] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *Proc. IEEE ICME Workshops*, pages 01–04, 2018.

[20] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier and Giro-i Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. In *arXiv*, January 2017.

[21] F. Chao, L. Zhang, W. Hamidouche, and O. Deforges. A multi-fov viewport-based visual saliency model using adaptive weighting losses for 360-degree images. *IEEE Trans. Multimed.*, pages 1–1, 2020.

[22] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. A dataset of head and eye movements for 360 videos. In *Proceedings of the 9th ACM multimedia systems conference*, pages 432–437, 2018.

[23] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[24] Ioannis Agtzidis, Mikhail Startsev, and Michael Dorr. 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proc. ACM MM*. ACM, 2019.

[25] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *Proc. IEEE/CVF CVPR*, pages 5333–5342, 2018.

[26] Qin Yang, Yuqi Li, Chenglin Li, Hao Wang, Sa Yan, Li Wei, Wenrui Dai, Junni Zou, Hongkai Xiong, and Pascal Frossard. SVGC-AVA: 360-degree video saliency prediction with spherical vector-based graph convolution and audio-visual attention. *IEEE Trans. Multimed.*, 2023.

[27] Yi Zhang, Fang-Yi Chao, Wassim Hamidouche, and Olivier Deforges. Pav-sod: A new task towards panoramic audio-visual saliency detection. *ACM Trans. Multimed. Comput. Commun. Appl.*, 19(3):1–26, 2023.

[28] Fang-Yi Chao, Cagri Ozcinar, Chen Wang, Emin Zerman, Lu Zhang, Wassim Hamidouche, Olivier Deforges, and Aljosa Smolic. Audio-visual perception of omnidirectional video for virtual reality applications. In *IEEE ICME Workshops*, pages 1–6. IEEE, 2020.

[29] Yuxin Zhu, Huiyu Duan, Kaiwei Zhang, Yucheng Zhu, Xilei Zhu, Long Teng, Xiongkuo Min, and Guangtao Zhai. How does audio influence visual attention in omnidirectional videos? database and model. *IEEE Trans. Image*

Process., pages 1–1, 2025.

[30] Yasser Dahou, Marouane Tliba, Kevin McGuinness, and Noel O'Connor. ATSal: An attention based architecture for saliency prediction in 360 videos. In *Proc. ICPR*, pages 305–320, 2021.

[31] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360-degree videos. In *Proc. ECCV*, September 2018.

[32] Yasser Abdelaziz Dahou Djilali, Tarun Krishna, Kevin McGuinness, and Noel E. O'Connor. Rethinking 360deg image visual attention modelling with unsupervised learning. In *Proc. ICCV*, pages 15414–15424, October 2021.

[33] Yucheng Zhu, Guangtao Zhai, and Xiongkuo Min. The prediction of head and eye movement for 360 degree images. *Signal Process. Image Commun.*, 69:15–25, 2018.

[34] Yucheng Zhu, Guangtao Zhai, Xiongkuo Min, and Jiantao Zhou. The prediction of saliency map for head and eye movements in 360 degree images. *IEEE Trans. Multimed.*, 22(9):2331–2344, 2020.

[35] Yucheng Zhu, Guangtao Zhai, Xiongkuo Min, and Jiantao Zhou. Learning a deep agent to predict head movement in 360-degree images. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(4), December 2020.

[36] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proc. IEEE/CVF CVPR*, pages 1420–1429, 2018.

[37] Minglang Qiao, Mai Xu, Zulin Wang, and Ali Borji. Viewport-dependent saliency prediction in 360 video. *IEEE Trans. Multimed.*, 23:748–760, 2020.

[38] Yucheng Zhu, Guangtao Zhai, Yiwei Yang, Huiyu Duan, Xiongkuo Min, and Xiaokang Yang. Viewing behavior supported visual saliency predictor for 360 degree videos. *IEEE Trans. Circuits Syst. Video Technol.*, 32(7):4188–4201, 2022.

[39] Heeseung Yun, Sehun Lee, and Gunhee Kim. Panoramic vision transformer for saliency detection in 360-degree videos. In *Proc. ECCV*, pages 422–439. Springer, 2022.

[40] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. ICML*, July 2021.

[41] Xiongkuo Min, Guangtao Zhai, Chunjia Hu, and Ke Gu. Fixation prediction through multimodal analysis. In *Proc. VCIP*, pages 1–4, 2015.

[42] Jiazhong Chen, Qingqing Li, Hefei Ling, Dakai Ren, and Ping Duan. Audiovisual saliency prediction via deep learning. *Neurocomputing*, 428:248–258, 2021.

[43] Dandan Zhu, Xuan Shao, Qiangqiang Zhou, Xiongkuo Min, Guangtao Zhai, and Xiaokang Yang. A novel lightweight audio-visual saliency model for videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(4), July 2023.

[44] Dandan Zhu, Kun Zhu, Weiping Ding, Nana Zhang, Xiongkuo Min, Guangtao Zhai, and Xiaokang Yang. MTCAM: a novel weakly-supervised audio-visual saliency prediction model with multi-modal transformer. *IEEE Trans. Emerg. Top. Comput. Intell.*, 8(2):1756–1771, 2024.

[45] Junwen Xiong, Peng Zhang, Tao You, Chuanyue Li, Wei Huang, and Yufei Zha. DiffSal: Joint audio and video learning for diffusion saliency prediction. In *Proc. IEEE/CVF CVPR*, pages 27273–27283, June 2024.

[46] Fang-Yi Chao, Cagri Ozcinar, Lu Zhang, Wassim Hamidouche, Olivier Deforges, and Aljosa Smolic. Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *Proc. VCIP.*, pages 355–358, 2020.

[47] Mert Cokelek, Nevrez Imamoglu, Cagri Ozcinar, Erkut Erdem, and Aykut Erdem. Leveraging frequency based salient spatial sound localization to improve 360 video saliency prediction. In *Proc. MVA*, pages 1–5, 2021.

[48] Dandan Zhu, Kaiwei Zhang, Nana Zhang, Qiangqiang Zhou, Xiongkuo Min, Guangtao Zhai, and Xiaokang Yang. Unified audio-visual saliency model for omnidirectional videos with spatial audio. *IEEE Trans. Multimed.*, 26:764–775, 2024.

[49] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Proc. NeurIPS*, 33, 2020.

[50] Antje Nuthmann and George L. Malcolm. Eye guidance during real-world scene search: The role color plays in central and peripheral vision. *Journal of Vision*, 16(2):3–3, 01 2016.

[51] Hans-Peter Frey, Christian Honey, and Peter König. What's color got to do with it? the influence of color on visual attention in different categories. *Journal of Vision*, 8(14):6–6, 10 2008.

[52] Halit Ozsoy. Audio-visual saliency in omnidirectional panoramic scenes. Master's thesis, Bogazici University, Istanbul, Turkey, May 2024. Thesis No: 882420, Available at https://tez.yok.gov.tr/UlusalTezMerkezi.

[53] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE ICASSP*, pages 721–725, 2020.

[54] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP*, pages 776–780, 2017.

[55] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018.

[56] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, Nov 2009.

[57] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006.

[58] Ioannis Agtzidis and Michael Dorr. Getting (more) real: Bringing eye movement classification to hmd experiments with equirectangular stimuli. In *Proc. ACM Symposium on Eye Tracking Research & Applications*, 2019.

[59] Erwan David, Jesús Gutiérrez, Melissa Le-Hoa Vo, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. The salient360! toolbox: Processing, visualising and comparing gaze data in 3d. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2023.

[60] Immo Schuetz and Katja Fiehler. Eye tracking in virtual reality: Vive pro eye spatial accuracy, precision, and calibration reliability. *J. Eye Mov. Res.*, 15(3), September 2022.

[61] Pieter Blignaut. Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, 71(4):881–895, May 2009.

[62] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, page 71–78, 2000.

[63] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 2801–2810, 2022.

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

[65] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Proc. NeurIPS*, 31, 2018.

[66] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Proc. Interspeech*, pages 2753–2757, 2022.

[67] Shentong Mo and Pedro Morgado. Localizing visual

sounds the easy way. In *Proc. ECCV*, pages 218–234, 2022.
[68] Guanqun Ding, Nevrez İmamoğlu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image Vis. Comput.*, 120:104395, 2022.
[69] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
[70] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proc. IEEE/CVF CVPR*, pages 6450–6459, 2018.
[71] Yuqin Cao, Xiongkuo Min, Wei Sun, and Guangtao Zhai. Attention-guided neural networks for full-reference and no-reference audio-visual quality assessment. *IEEE Trans. Image Process.*, 32:1882–1896, 2023.
[72] Xilei Zhu, Huiyu Duan, Yuqin Cao, Yuxin Zhu, Yucheng Zhu, Jing Liu, Li Chen, Xiongkuo Min, and Guangtao Zhai. Perceptual quality assessment of omnidirectional audio-visual signals. In Lu Fang, Jian Pei, Guangtao Zhai, and Ruiping Wang, editors, *Artif. Intell.*, pages 512–525, 2024.
[73] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylène C. Q. Farias, and Alan Conrad Bovik. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Process.*, 29(11):6054–6066, 2020.
[74] Seunghoon Cha, Jungjin Lee, Seunghwa Jeong, Younghui Kim, and Junyong Noh. Enhanced interactive 360° viewing via automatic guidance. *ACM Trans. Graph.*, 39(5), May 2020.
[75] J. C. Chiang, C. Y. Yang, B. Dedhia, and Y. F. Char. Saliency-driven rate-distortion optimization for 360-degree image coding. *Multimed. Tools Appl.*, 80:8309–8329, March 2021.
[76] F. Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic. Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *Proc. IEEE VCIP*, pages 355–358, 2020.
[77] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *Proc. ACM Int. Conf. Multimedia*, pages 932–940, 2018.
[78] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Process. Lett.*, 24:1408–1412, 2017.



Mert Cokelek received his Bachelor's degree in Computer Science from Hacettepe University, Ankara, Turkey, in 2021, and his Master's degree in Computer Science from Koç University, Istanbul, Turkey, in 2023. His research interests include computer vision focused on omnidirectional videos, audio-visual saliency prediction and generative AI.



Halit Ozsoy is currently pursuing his M.A. degree in the Cognitive Science Program at Boğaziçi University, specializing in psychology and computer science. He completed his undergraduate studies in Computer Engineering at the same university. His research primarily focuses on analyzing the impact of specific semantic contextual information on audio-visual saliency in omnidirectional environments.



Nevrez Imamoglu is a Senior Researcher at National Institute of Advanced Industrial Science and Technology (AIST), Japan since 2016. Before AIST, he was with RIKEN BSI as a JSPS Foreign Postdoctoral Fellow from 2015 to 2016. He received Ph.D. from Chiba University, Japan in 2015. He was also with Nanyang Technbological University, Singapore as a Research Associate from 2010 to 2011. His research interests are applications of signal image processing, computer vision and machine learning.



Cagri Ozcinar is a machine learning scientist with expertise in image processing, video streaming, and computer vision solutions for embedded devices. His expertise also extended to immersive media technologies and speech processing areas. He held academic research positions in Ireland, UK, and EU Universities. He earned his M.Sc. and Ph.D. degrees from the University of Surrey in the UK. He authored/co-authored over 100 scientific works, including journal and conference papers, as well as books.



Inci Ayhan is an Associate Professor in the Department of Psychology at Boğaziçi University, Turkey. She received her Ph.D. from University College London, UK, in 2010. Following her doctorate, she held postdoctoral positions at University College London and University of London. Her research primarily explores visual perception, temporal processing in the visual system, time perception, and the interactions between the visual and motor systems.



Erkut Erdem is a Professor with the Department of Computer Engineering, Hacettepe University, Turkey. He received his Ph.D. degree from Middle East Technical University in 2008. After completing his Ph.D., he continued his post-doctoral studies with Télécom ParisTech, École Nationale Supérieure des Télécommunications, France, from 2009 to 2010. His research interests include semantic image editing, visual saliency prediction, and multimodal machine learning.



Aykut Erdem is an Associate Professor of Computer Science at Koç University. He received his Ph.D. from Middle East Technical University in 2008 and was a post-doctoral researcher at the Ca'Foscari University of Venice from 2008 to 2010. Previously, he was with Hacettepe University. His research interests lie in computer vision and machine learning, with a particular focus on semantic image editing and vision–language integration.

SUPPLEMENTARY MATERIAL

Overview. The supplementary material has the following structure:

- Appendix A provides more details about our proposed audio-blind SalViT360 model, offering extensive analyses to understand its performance and underlying mechanics.
- Appendix B gives details about the implementation aspects of our SalViT360-AV model and conduct thorough analyses on the influence and effectiveness of the audio backbone in enhancing our model's capabilities.
- Appendix C presents additional analysis on our YT360-EyeTracking dataset, including a breakdown of fixation patterns across audio categories and an investigation into the role of audio modalities in guiding attention and inter-subject consistency.
- Appendix D evaluates the performance of our SalViT360 model through omnidirectional image quality assessment task.

APPENDIX A
SALVIT360 ADDITIONAL EXPERIMENTS

This section presents additional experiments on the proposed Viewport Spatio-Temporal Attention (VSTA) mechanism and the implementation details of our proposed Viewport Augmentation Consistency (VAC) loss.

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.

Fig. A.1 shows that increasing temporal window size gradually leads to a performance boost in three metrics and an insignificant performance drop in NSS for $F > 2$. We conclude our experiments at $F = 8$, considering the memory limit of a single Tesla V100 GPU.

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 proposed VSTA mechanism for saliency prediction in omnidirectional videos. Notably, even with a single VSTA transformer block, our 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.3 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. Alternatively, VSTA is computationally more efficient as its complexity grows linearly with respect to temporal window size, which grows quadratically in JSTA.

A.4 Viewport Augmentation Consistency (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 [63], 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 (Fig. A.3).

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.

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.

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.

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. A.4 show that each augmentation method improves model consistency significantly.

Mask-weighted VAC Loss. We use an optional weight for the proposed $\mathcal{L}_{\text{VAC}}(P,\ P')$ loss to increase consistency, especially on the overlapping regions on ERP. In Fig. A.5, we provide the weight mask computed from the gnomonic projection for the original tangent image set. The performance comparison in Table A.2 demonstrates the effectiveness of the proposed masking operation.

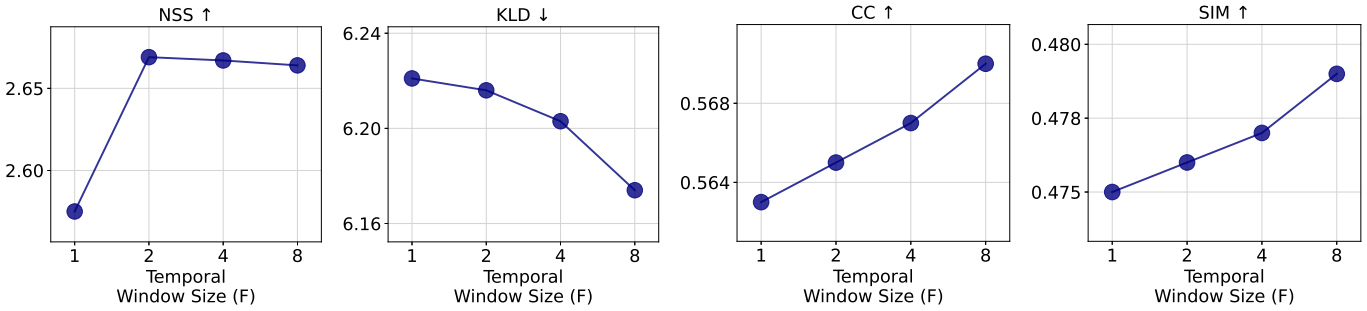


Fig. A.1: Performance of our ODV saliency prediction model in terms of four evaluation metrics (NSS, KLD, CC, SIM) as a function of temporal window size (F) on the validation split of VR-EyeTracking [25] dataset.

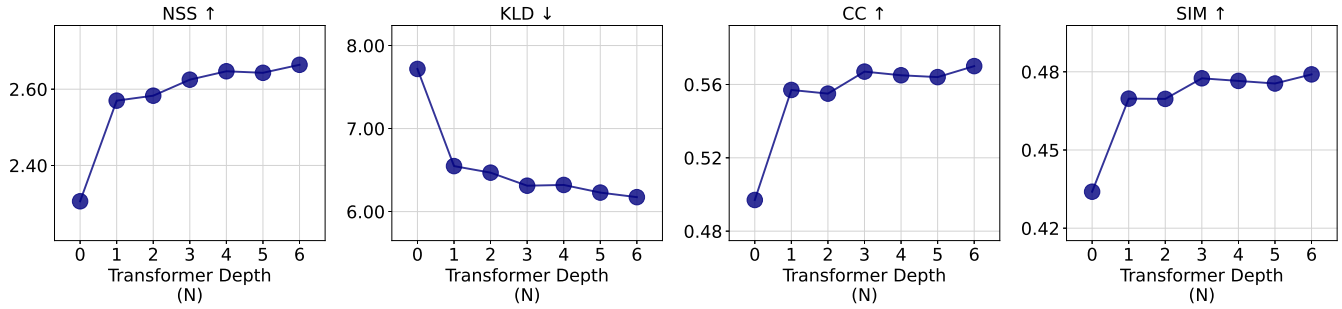


Fig. A.2: Performance of our ODV saliency prediction model in terms of four evaluation metrics (KLD, NSS, CC, SIM) as a function of VSTA depth (N) on the validation split of VR-EyeTracking dataset.

TABLE A.1: **Quantitative comparison** for space-only attention, the proposed Viewport Spatio-Temporal Attention and existing Joint Spatio-Temporal attention for omnidirectional video 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

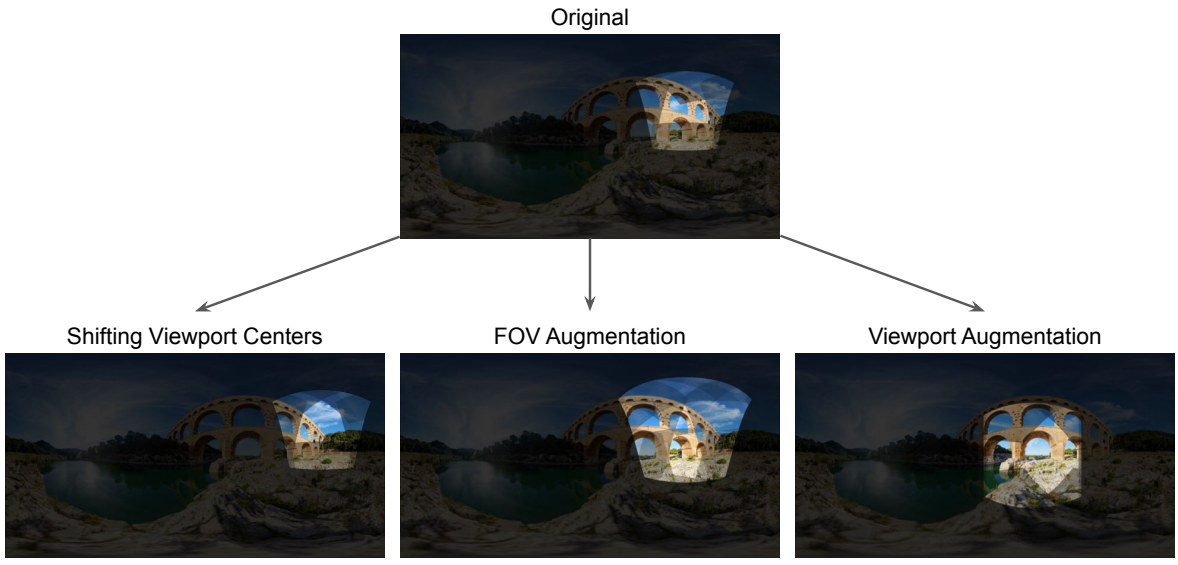


Fig. A.3: A tangent viewport from the original projection, highlighted on Equirectangular Projection (ERP) (top), compared with three augmentation methods (bottom).

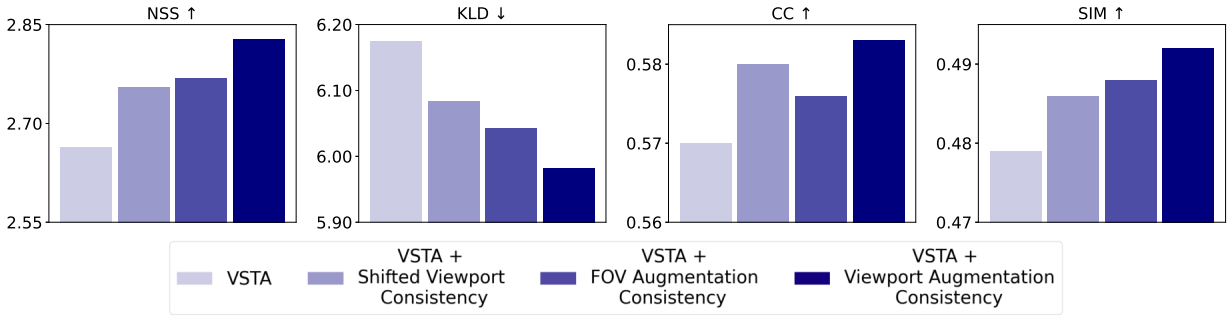


Fig. A.4: Performance of our VSTA baseline compared with three augmentation consistency methods on four metrics, on the validation split of VR-EyeTracking dataset.

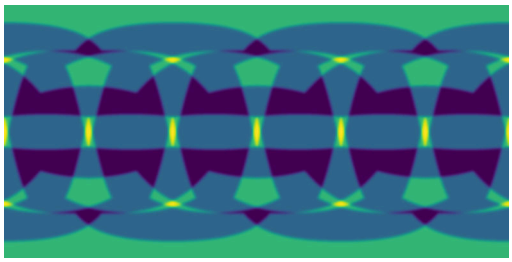


Fig. A.5: **Weight mask used for VAC Loss.** Each pixel coordinate in ERP takes a value based on the number of tangent viewports it is projected onto (max. 4). Brighter colors represent increasing overlaps.

TABLE A.2: **Quantitative comparison** for the proposed VAC Loss with and without mask, on the validation split of the 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

APPENDIX B
SALVIT360-AV DETAILS

In this section, we first formally describe the rotation and decoding of first-order ambisonics, the spatial audio encoding method used in our audio-visual pipeline. Then, we compare three audio backbones described in our main paper and investigate the effect of audio clip duration on saliency prediction performance.

B.1 Spatial Audio

First-order Ambisonics (FOA) is a sound encoding technique that captures and reproduces audio, allowing immersive listening experiences. This method diverges from the traditional stereo or surround sound systems, which rely on discrete speakers arranged around the listener to simulate spatial audio. Instead, Ambisonics adopts a spherical approach to sound representation, enabling the playback of audio from any direction surrounding the listener. FOA integrates sound information from multiple directions into a single audio signal. This signal can then be decoded and played back from any direction, offering a versatile listening experience. The encoding process thus captures the direction of sound waves by recording and mixing the sound sources according to the positions of each source relative to a reference point on the sphere. Ambisonics approximates the sound pressure field at a single point in space using a spherical harmonic decomposition.

More specifically, an audio signal $f((\theta,\phi),t)$ is represented by a spherical harmonic expansion of order N

$$f((\theta,\phi),t) = \sum_{n=0}^{N}\sum_{m=-n}^{n} Y_n^m(\theta,\phi)\alpha_n^m(t)\ , \qquad \text{(B.1)}$$

where θ and ϕ denote the zenith and the azimuth angles at time t, respectively, $Y_n^m(\theta,\phi)$ is the spherical harmonic of order n and degree m, and $\alpha_n^m(t)$ are the expansion coefficients. For simplicity, (Eq. B.1) can be written as $f((\theta,\phi),t) = \boldsymbol{y}_N\boldsymbol{\alpha}_N(t)$.

Encoding. Given a set of k audio signals $s_1(t),\ldots,s_k(t)$ from directions $\theta_1,\ldots,\theta_k$,

$$\boldsymbol{\alpha}_N(t) = \sum_{i=1}^{k} \boldsymbol{y}_N((\theta_i,\phi_i))s_i(t). \qquad \text{(B.2)}$$

The spherical harmonic coefficients, denoted as $\boldsymbol{\alpha}_N$, or ambisonic channels $(\boldsymbol{\alpha}_w,\boldsymbol{\alpha}_x,\boldsymbol{\alpha}_y,\boldsymbol{\alpha}_z)$, are integral to decoding the captured sound into a speaker array, facilitating the reconstruction of the sound field. FOA employs four distinct channels to represent the first-order coefficients of the spherical harmonic expansion: α_0^0, α_0^{-1}, α_1^0, and α_1^1. These channels encapsulate the sound information in a way that allows for a spatially immersive playback, reproducing sound from any direction on the sphere.

Rotation and Decoding. Adjusting the orientation of the sound field to match the listener's perspective involves rotating the spherical harmonic coefficients, $\boldsymbol{\alpha}_N$, with a rotation matrix, $\mathbf{R}$. For FOA, this rotation is executed using a 3x3 matrix tailored to the specific rotation defined by Euler angles (yaw, pitch, and roll). This process, applied prior to decoding, modifies the ambisonic channels directly. By manipulating these coefficients, it is possible to precisely alter the directionality of the sound field, ensuring that it aligns with the target orientation

$$\boldsymbol{\alpha}'_N(t) = \mathbf{R}\cdot\boldsymbol{\alpha}_N(t)\ . \qquad \text{(B.3)}$$

In SalViT360 and SalViT360-AV, we use $T = 18$ tangent viewports located around the sphere centered at $(\theta_t,\phi_t))_{t=1}^{T}$. For each viewport, we apply the ambisonics channel rota-

TABLE B.1: **Performance comparison of three pre-trained audio models** as the audio backbone of SalViT360-AV, on VR-EyeTracking, 360AV-HM, YT360-EyeTracking datasets.

	VR-EyeTracking [25]				360AV-HM [76]				YT360-EyeTracking			
Method	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
AVSA [49]	n/a	n/a	n/a	n/a	2.499	12.916	0.381	0.284	2.501	7.768	0.516	0.417
EZ-VSL [49]	2.817	5.342	0.598	0.509	2.472	13.833	0.380	0.277	2.449	8.288	0.511	0.407
PaSST [66]	2.821	5.334	0.599	0.511	2.473	13.830	0.379	0.278	2.449	8.341	0.512	0.407

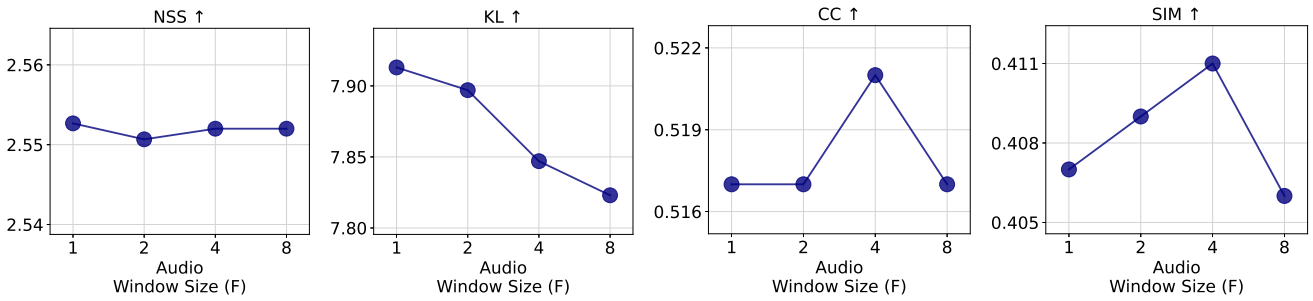


Fig. B.1: Performance of our ODV saliency prediction model in terms of four evaluation metrics (NSS, KLD, CC, SIM) as a function of audio window size (F) on the validation split of YT360-EyeTracking dataset.

tion given in (Eq. B.1) to generate new ambisonics. Then, we decode the ambisonics waveforms for the forward direction using (Eq. 5 of the main manuscript). This way, we are able to obtain viewport-specific waveforms in mono format, capturing the sound sources located in the direction of each tangent viewport.

B.2 Additional Experiments

B.2.1 On Audio Backbones

We compare the performance of three audio backbones in our SalViT360-AV pipeline, namely, PaSST [66], EZ-VSL [67], and AVSA [49] in Table B.1. PaSST [66] is an Audio Spectrogram Transformer trained for audio classification, distinguished by its self-attention mechanism on mel-spectrogram patches and its patchout-based regularization strategy. PaSST further benefits from initialization with DeiT's pre-trained ImageNet weights, which are subsequently adapted to AudioSet for improved audio classification. EZ-VSL is a ResNet-18 model fine-tuned on mel-spectrograms for visual sound localization in 2D, while PaSST, an Audio Spectrogram Transformer variant, is trained for audio classification. AVSA employs a 9-layer 2D CNN that operates on mel-spectrograms in the time-frequency domain, specifically optimized for audio-visual spatial alignment within 360° environments. While the AVSA encoder, which is specifically trained for first-order ambisonics, yields the best quantitative results, we used PaSST in our main experiments due to its availability on mono audio.

B.2.2 On Audio Window Size

We report the impact of audio window size (in seconds) on saliency prediction performance in Fig. B.1. The findings suggest that there is an optimal audio window size range that marginally improves the model's ability to predict visual attention across different metrics. Notably, a window size of approximately 4 secs tends to yield slightly better or comparable results relative to smaller or larger window sizes. This analysis highlights the importance of audio temporal context in enhancing the accuracy of SalViT360-AV.

APPENDIX C
ADDITIONAL ANALYSIS ON OUR YT360-EYETRACKING DATASET

C.1 Fixation Distributions Across Audio Categories

To further examine the presence and variability of center bias in our dataset, we visualize example frames and fixation patterns across three distinct audio categories: *vehicle*, *speech*, and *music*. As shown in Fig. C.1-C.3, each category captures diverse visual contexts, ranging from structured environments (e.g., car interiors) to more unconstrained scenes (e.g., live concerts or social gatherings). In the *vehicle* category, the camera is typically placed at the front of the vehicle, facing the road ahead. This layout introduces a strong content-driven center bias, as participants tend to fixate in the forward driving direction, near the horizon. Average fixation density maps given in Fig. C.4 confirm this behavior, especially under spatial audio. In contrast, the *speech* and *music* categories feature more dynamic and visually diverse content, often with multiple people or performers distributed across the scene. As reflected in the average fixation maps given in Fig. C.5 and C.6), subjects explore wider regions of the visual field, and the fixation distribution becomes more dispersed.



Fig. C.1: Sample frames from the *Vehicle Sound* category. Scenes typically feature forward-facing driving perspectives, leading to a stronger center bias in fixations as demonstrated in Fig. C.4.



Fig. C.2: Sample frames from the *Human Speech* category. Scenes involve multi-speaker conversations and interactions in diverse settings. Spatial audio promotes more focused fixations around the speaking individuals, resulting in more concentrated attention patterns, as further illustrated in Fig. C.5.



Fig. C.3: Sample frames from the *Music* category. These include live performances and concerts with highly dynamic visuals and distributed sound sources. Spatial audio encourages more focused fixations around relevant sound sources, as illustrated in Fig. C.6.

C.2 Effect of Audio Modalities on Fixation

To quantify the role of auditory context in directing attention, we compare fixation patterns under three audio modalities: *mute*, *mono*, and *ambisonics*. Across all content categories, spatial audio (ambisonics) consistently results in more concentrated and semantically aligned fixations, typically near the sound sources. In contrast, mute conditions often lead to less focused exploration, particularly around

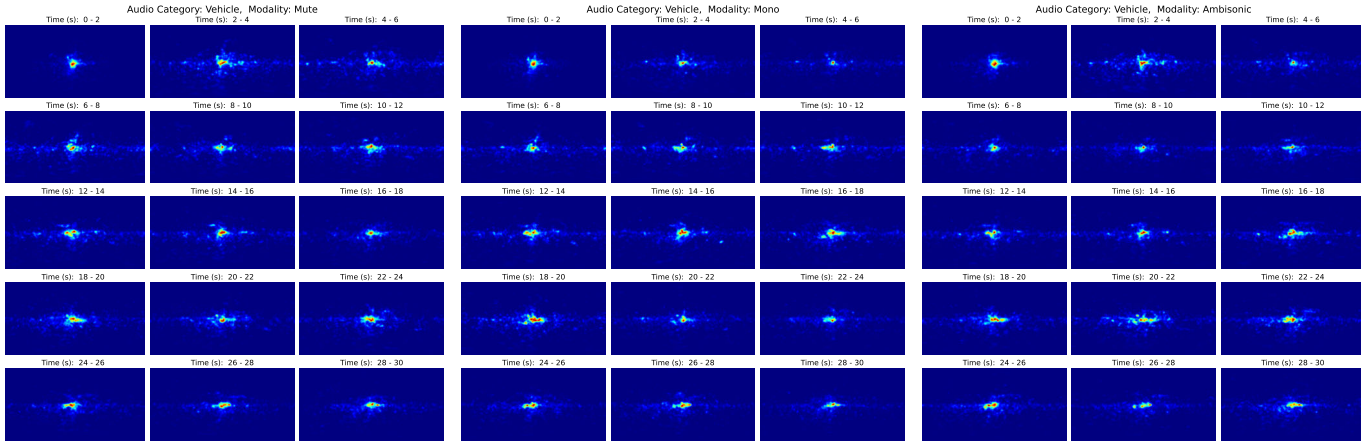


Fig. C.4: Average fixation density maps for the *vehicle* category across different time intervals (2-secs windows) under *mute*, *mono*, and *ambisonics* audio conditions. Fixations mainly suggest a stronger center bias due to the forward-facing driving behavior.

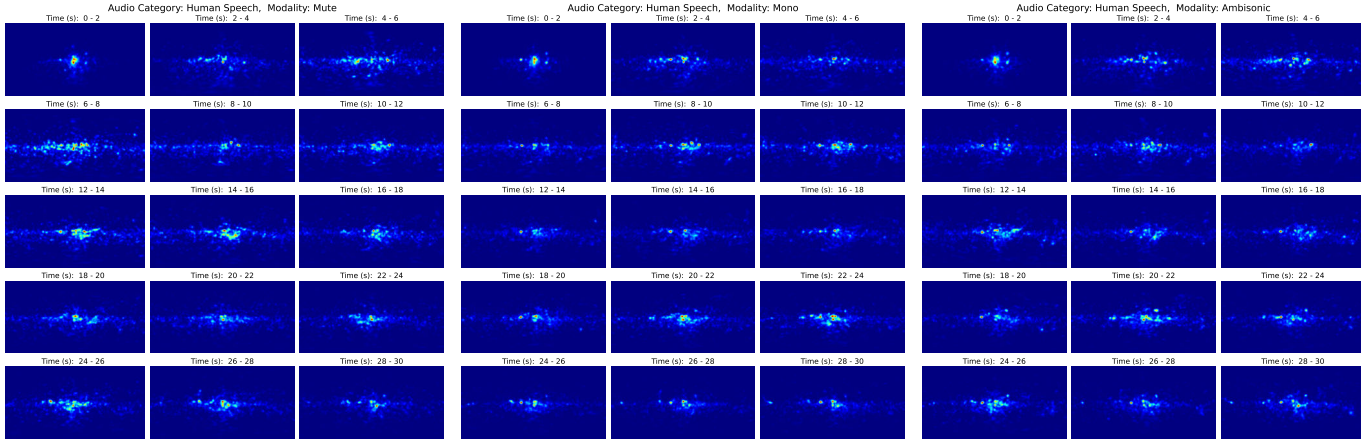


Fig. C.5: Average fixation density maps for the *speech* category across different time intervals (2-secs windows) under *mute*, *mono*, and *ambisonics* audio. Compared to vehicle scenes, fixations here are more evenly distributed and driven by the existence of mono and spatial audio.

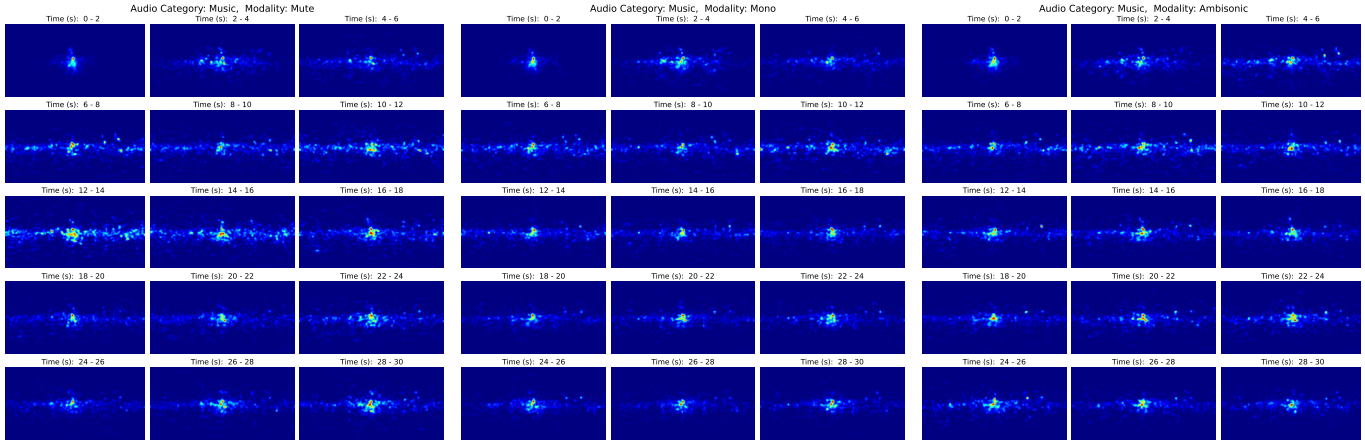


Fig. C.6: Average fixation density maps for the *music* category across different time intervals (2-secs windows) under *mute*, *mono*, and *ambisonics* audio. The presence of spatial audio helps guide attention toward musicians or instruments, whereas fixations under mute audio are more diffuse.

the equator, indicating a lack of auditory cues to guide viewer attention.

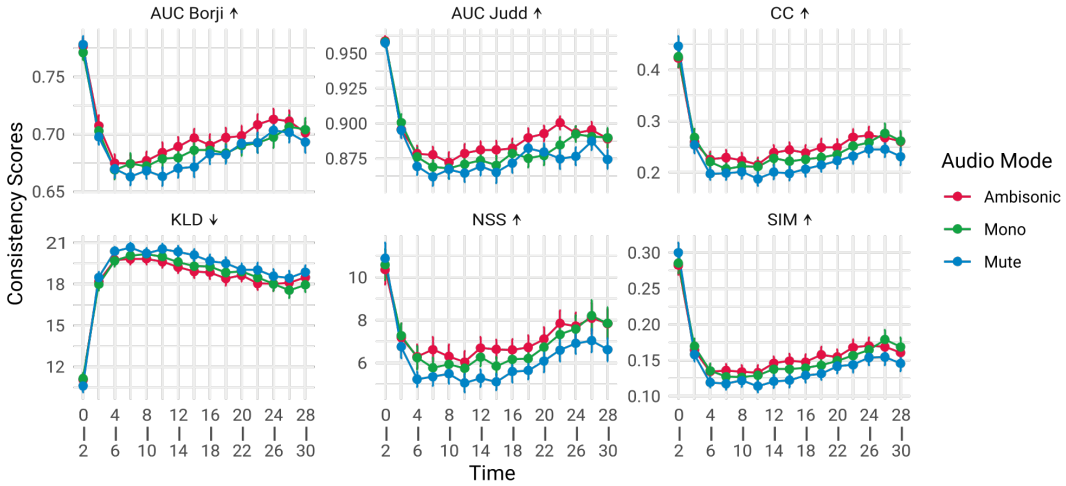


Fig. C.7: **Inter-subject consistency scores of viewer fixations under three different audio conditions, mute, mono, and ambisonics, measured over 2-secs time windows.** Higher consistency in spatial audio (ambisonics) highlights its stronger influence on guiding collective visual attention.

To further quantify the influence of audio modality on collective viewing behavior, we analyze inter-subject consistency across viewers under different audio conditions. As shown in Fig. C.7, consistency scores are computed over 2-second time windows across all participants. The results reveal a consistent trend: spatial audio (ambisonics) yields significantly higher inter-subject consistency compared to mono and mute conditions. This suggests that directional auditory cues not only enhance individual saliency but also help synchronize visual attention across viewers. Notably, consistency scores are highest at the onset of each video, likely due to the standardized ERP-centered initial viewport. A noticeable dip in consistency occurs between 4 and 10 seconds, reflecting a phase of exploratory behavior as viewers diverge in their gaze patterns. This is followed by a gradual realignment of attention, particularly under the influence of spatial audio, indicating that spatial cues guide viewers back toward semantically relevant regions. These temporal dynamics highlight the crucial role of audio modality in modulating both the distribution and convergence of attention over time, further emphasizing the importance of incorporating spatial audio in 360° saliency modeling.

APPENDIX D
USE CASE: SALIENCY-GUIDED OMNIDIRECTIONAL VIDEO QUALITY ASSESSMENT

This section shows the evaluation of our method and the competing approaches on a downstream task, which involves assessing the visual quality of omnidirectional videos.

In Table D.1, we report the performance of two PSNR variants compared to ground-truth DMOS values in the VQA-ODV [77] 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.

TABLE D.1: **Comparison with state-of-the-art** saliency models for *saliency-guided omnidirectional video quality assessment* on VQA-ODV [77] dataset, as a downstream task.

	PSNR			WS-PSNR [78]		
Weight	PCC↑	SRCC↑	RMSE↓	PCC↑	SRCC↑	RMSE↓
None	0.650	6.664	7.502	0.671	0.686	7.233
PAVER [39]	0.661	0.667	7.481	0.679	0.691	6.914
Djilali et. al. [32]	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

Following the prior work [39], the saliency-weighted PSNR and WS-PSNR values are calculated as:

$$\text{PSNR} = 10\log_{10}\left(\frac{Y_{max}^2 \cdot \sum_{p\in\mathbb{P}} w_{sal}(p)}{\sum_{p\in\mathbb{P}}(Y(p)-Y'(p))^2 \cdot w_{sal}(p)}\right) \quad \text{(D.1)}$$

$$\text{WS-PSNR} = 10\log_{10}\left(\frac{Y_{max}^2 \cdot \sum_{p\in\mathbb{P}} w_{sal}(p)\cos\theta_p}{\sum_{p\in\mathbb{P}}(Y(p)-Y'(p))^2 \cdot w_{sal}(p)\cos\theta_p}\right) \quad \text{(D.2)}$$

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.