Supplementary Material
AViSal360: Audiovisual Saliency Prediction for 360° Video

Edurne Bernal-Berdun [*] Jorge Pina Mateo Vallejo Ana Serrano Daniel Martin Belen Masia

Universidad de Zaragoza, I3A

The supplemental material of this paper is composed of the following items:

- A project web page including video results of our method (AViSal360) and the three main state-of-the-art approaches: `https://graphics.unizar.es/projects/AViSal360_2024`.
- This PDF document.

This PDF contains additional information on the following topics:

- (S.1) Ablation Studies
- (S.2) Spatial Representation of Audio: Upsampled Audio Energy Maps
- (S.3) AViSal360: Implementation Details
- (S.4) Metrics for Saliency Map Evaluation
- (S.5) Extended Comparative with Previous Works
- (S.6) Drawbacks of Optical Flow for Saliency Prediction
- (S.7) Datasets of Viewing Behavior in 360° Content
- (S.8) D-SAV360 Videos for Audiovisual Evaluation

S.1 Ablation Studies

We performed ablation studies for our model, including also an evaluation of the performance of our proposed AEMs in comparison to conventional ones [17], used by previous audiovisual saliency models [7,26]. The D-SAV360 dataset, on which we have evaluated and compared our model (Secs. 4.2 and 4.3 of the main paper), comprises a diverse collection of 87 videos where the presence and nature of audio, and its significance in visual attention varies greatly. For the ablations of our own model, we opt for videos where audio plays a significant role, more representative of the cases that our model targets. We selected a subset of videos from the dataset wherein the CC scores between the AEMs and the ground truth saliency exceeded a threshold, which we set at 0.25. This subset, consisting of 29 videos, allows us to evaluate the impact of audio without dimming its influence among videos where it is not relevant. The IDs of these videos, and the videos themselves, can be found in the supplementary material.

To evaluate the impact of audio information, we re-trained AViSal360 without incorporating any audio features: The model architecture excludes the audio branch and solely processes RGB frame sequences. Table 1 (first row) presents the results obtained for this model trained with a conventional KLD loss (i.e., without the second term of our loss function in Eq. 2). We observe a general decrease in saliency prediction accuracy when audio input is omitted from the model in the quantitative results, and the effect of audio on saliency can be clearly observed in the qualitative comparisons shown in Figure 1 and Figure 1 of the main paper, where the model without auditory information finds all individuals and motion salient, while AViSal360 can distinguish the truly salient due to audio information.

Table 1: Ablation study results. We assess the performance of various configurations of AViSal360 on the D-SAV360 dataset, following a five k-fold strategy to ensure robust evaluation. Metrics include CC, NSS, SIM, and RMSE, with the table presenting average mean scores and standard deviations (in brackets) across videos. Best performances are highlighted in bold, and the second-best are underlined. Further details can be found in Section S.1.

Model	CC ↑	NSS ↑	SIM ↑	RMSE ↓
No Audio Information	0.533 (0.104)	3.923 (1.030)	**0.400** **(0.060)**	0.104 (0.016)
Conventional AEM	0.519 (0.107)	3.786 (1.080)	0.373 (0.058)	0.071 (0.015)
Random ImageBind-ViT	0.532 (0.111)	3.945 (1.160)	0.382 (0.06)	0.064 (0.015)
Random AEM	0.488 (0.121)	3.695 (1.160)	0.345 (0.063)	0.064 (0.014)
AViSal360 (Ours)	**0.545** **(0.111)**	**4.100** **(1.180)**	0.391 (0.060)	**0.061** **(0.013)**

Furthermore, motivated by the insights of Agrawal et al. [1], we investigated the model's response to random audio information to verify that the addition of the audio branches does not merely function as a regularizer, and to assess the contribution of each of our audio features (directional and semantic). We do this by feeding our model with random AEMs and random ImageBind-ViT embeddings, respectively. The results, also presented in Table 1, highlight the model's dependency on AEMs and emphasize the necessity of precise directional and semantic audio information. A similar evaluation can be found in Section S.5 of this supplementary material for the state-of-the-art models AVS360 and SVGC-AVA.

Finally, we train our model using the conventional AEMs employed by previous works [17], which rely on conventional beamformers to decode the AEMs from first-order ambisonic audio. Table 1 shows how the performance is notably worse than AViSal360 trained with our proposed AEMs. The proposed upsampling enhances the sound field representation, and the use of an adaptive beamformer such as MVDR allows for a clearer reconstruction of the incoming audio intensity from each direction, achieving more accurate AEMs that favor learning from the directional audio features.

[*] e-mail: edurnebernal@unizar.es

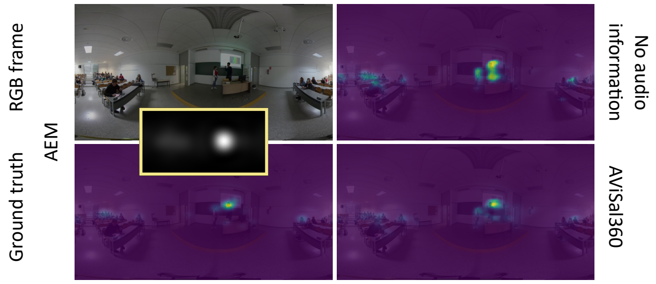

Figure 1: Qualitative ablation of AViSal360 for an example video (*video_2008* in the dataset) without audio processing features, labeled as the 'No Audio Information' configuration in Table 1 (excluding the audio branches and trained solely with a conventional KLD loss). *Left*: Sample RGB frame (*top*) and the corresponding ground-truth saliency (*bottom*). While this classroom scene features several potential points of visual interest among the audience, the ground truth saliency focuses on the presenter who is talking (as illustrated by the AEM). Under the 'No Audio Information' condition (*top right*), the model fails to discern the presenter as the primary point of interest, instead distributing saliency across several regions, including the audience to the sides of the frame. In contrast, AViSal360 with its full capabilities (*bottom right*) leverages the auditory information from the AEM and accurately focuses on the presenter as the main source of attention, matching more accuratelly the ground truth saliency.

S.2 Spatial Representation of Audio: Upsampled Audio Energy Maps

Audio in 360° content is typically played back from stereo headphones, where users receive different audio signals in each ear to simulate spatialized sounds. However, since the direction of a sound is relative to the orientation of the user's head, the audio signal for each ear should change as users orient their heads. Therefore, storing the incoming sounds within a stereo (two-channel) audio track, as in traditional videos, is not enough. There exist more sophisticated spatial representations of sound environments, yet the most common format for virtual reality (VR) applications is the *ambisonic* format. It represents the captured soundscape as a combination of different audio tracks, each encoding a directional signal according to *spherical harmonics* (see Figure 2). As such, it supports different orders, where ambisonics of higher order have an increased spatial localization at the expense of requiring specialized hardware for capture.

For real-time applications such as 360° video, it is widely adapted to employ *first order ambisonics* (FOA), which are composed of four channels that correspond to the first four spherical harmonics. Indeed, previous works in audiovisual saliency prediction [8, 26] have resorted to FOA representations. While FOA can represent the general direction of the sound, they usually lack precision. As discussed in the main document, having an accurate representation of where the audio salient regions are is a cornerstone for the development of audiovisual saliency models. We hypothesized that FOA does not necessarily suffice for this task, especially in cases where a high degree of spatial localization is required to resolve ambiguities in scenes with a large amount of nearby potential sound sources.

To alleviate this limitation, and feed AViSal360 with accurate enough audio spatial representations, we perform an upsampling (i.e., also coined *upmixing*) of the source audio files to fourth-order ambisonic (4OA) format, which assigns an audio channel for each of the first 25 spherical harmonics. We perform this upsampling from FOA to 4OA using COMPASS ambisonic upmixer [16]. We then extract their corresponding audio energy maps (AEMs, see main document for further details on this representation) with the PowerMap plug-in from the Spatial Audio Framework [15], using the Minimum Variance Distortionless Response (MVDR) method [6]. This method is a common algorithm in signal processing with phased arrays (such as ambisonic microphones) that employs an adaptative beamformer that minimizes the total power at the receiver while maintaining the power in the direction of the signal. This results in the reduction of the interfering signals while conserving the signal in the sampled direction.

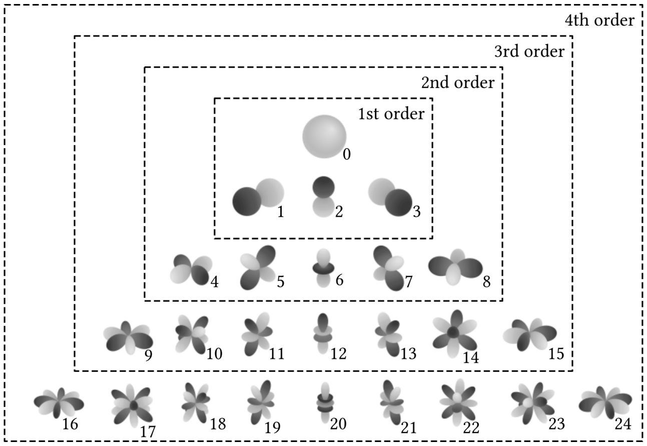

Figure 2: First four ambisonic orders and their audio channels with their corresponding spherical harmonics. Note that the higher the order, the more precise the information is.

Evaluation We have evaluated whether our upsampled 4OA is more accurate than its FOA counterpart. We resort to the QoEVAVE dataset [21], which contains ground-truth 4OA and FOA AEMs for twenty-six 360° videos. We have thus upsampled their FOA into 4OA and compared both the AEMs coming from the FOA and the ones extracted from our 4OA to the ground truth for all the videos in the dataset. Figure 3 shows the Root Mean Squared Error (RMSE) obtained, and Figure 4 depicts three qualitative comparisons of the three. The AEMs from our upsampled 4OA yield a smaller error, while also much better resembling the ground-truth AEMs, suggesting that our upsampling method can effectively generate much more precise AEMs, thus overcoming the common limitations of FOA.

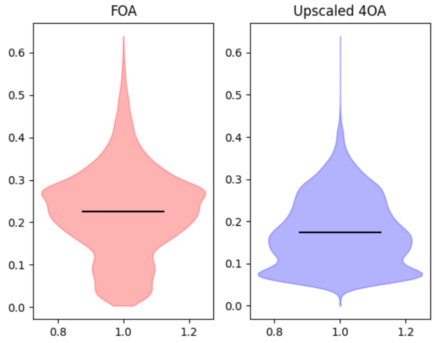

Figure 3: Qualitative evaluation of our upsampling strategy. The left violin plot shows the averaged RMSE when comparing the audio energy maps (AEMs) for ground-truth first-order ambisonics (FOA) and ground-truth fourth-order ambisonics (4OA), while the right violin plot shows the averaged RMSE when comparing the audio energy maps (AEMs) from our upsampled 4OA and ground-truth 4OA. Note how the latter yields a smaller error, thus better resembling the ground-truth data.

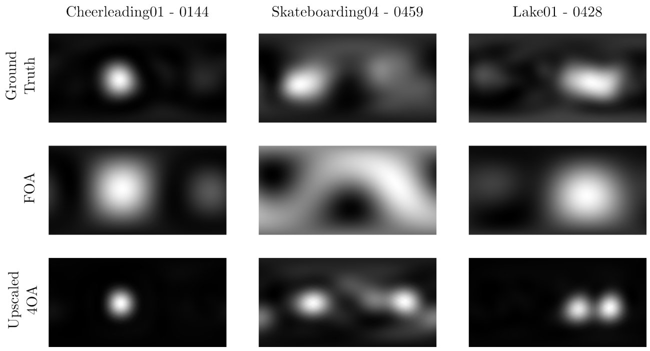

Figure 4: We show the audio energy maps (AEMs) for three frames randomly selected from three videos. From top to bottom: ground-truth 4OA, ground-truth FOA, upsampled 4OA. Note how our upsampled 4OA yields AEMs significantly more close to the ground-truth ones, surpassing FOA by a large margin, and suggesting that our upsampling strategy effectively improves the precision of AEMs.

S.2.1 Analysis of the Relationship Between AEMs and Fixations

To evaluate whether the proposed AEMs are, indeed, better representations of directional audio information for saliency prediction than the AEMs typically used [8, 26], we assess the degree of correlation between AEMs and gaze fixations. To this end, we use the AUC metric (see Section S.4) to measure to what extent fixation patterns align with conventional AEMs [17] and with our proposed AEMs. Higher AUC scores mean a greater alignment of ground truth fixations with the source of the sound. Figure 5 illustrates that our proposed AEMs achieve higher scores than those obtained from conventional AEMs and random AEMs.

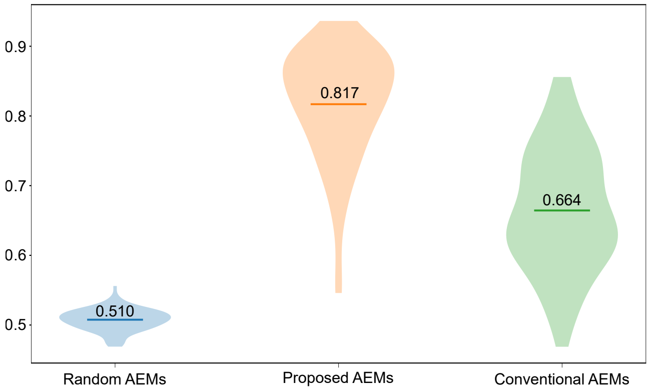

Figure 5: We have assessed the correlation between fixations and AEMs. The violin plots show AUC scores comparing ground-truth fixations from the D-SAV360 dataset with random (blue), conventional (orange), and our proposed (green) AEMs. Higher scores indicate better alignment between areas of high audio intensity and fixations. Our proposed AEMs achieve a higher score, the closest to 1, significantly outperforming random AEMs. This suggests that there exists a strong correlation between fixations and sound sources. On the other hand, conventional AEMs yield scores closer to random, which seems to highlight they are indeed limited for saliency prediction.

S.3 AViSal360: Implementation Details

Our model takes as input sequences of RGB frames, and their corresponding audio in first-order ambisonic (FOA) format. The RGB frame sequences have a total length of 20 frames, with a spatial resolution of 320×240, thus the input visual tensor has a total shape of $(3, 20, 320, 240)$. The visual tensor is fed to the visual encoder, which follows a ConvLSTM architecture, inspired by the implementation of SST-Sal [3]. It yields a visual feature vector of shape $(36, 120, 160)$ for each pass over each frame of the sequence.

Table 2: Summary of layers forming our semantic audio features encoder and our audiovisual fusion module. It represents the output size of each layer. bs denotes the batch size. We follow Sphernet implementation [10], thus all the spherical convolutional layers have a kernel size of 3×3.

Semantic Features Encoder	
Layers	Output Shape
Fully Connected	(bs, 512)
Batch Normalization	(bs, 512)
LeakyReLU	(bs, 512)
Fully Connected	(bs, 256)
Batch Normalization	(bs, 256)
LeakyReLU	(bs, 256)
Dropout (0.5)	(bs, 256)
Fully Connected	(bs, 64)
Batch Normalization	(bs, 64)
LeakyReLU	(bs, 64)
Fully Connected	(bs, 36)
Sigmoid	(bs, 36)

Audiovisual Fusion Module	
Layers	Output Shape
Spherical Convolution	(bs, 72, 120, 160)
Batch Normalization	(bs, 72, 120, 160)
LeakyReLU	(bs, 72, 120, 160)
Spherical Convolution	(bs, 36, 120, 160)
Batch Normalization	(bs, 36, 120, 160)
LeakyReLU	(bs, 36, 120, 160)

Then, we upsample the AEMs from the FOA to our proposed AEMs coming from fourth-order ambisonics (4OA) as stated in Section S.2. The sequence of our proposed AEMs is introduced after the visual encoder phase, thus we reshape them to match the spatial shape of the extracted visual features. Therefore, the AEM tensors corresponding to each frame of the sequence have a shape of $(1, 120, 160)$.

Finally, to obtain the audio semantic features, we convert the ambisonics to a mono audio track maintaining the sampling rate of 48,000 Hz. The tracks are split in sequences of 2.5 seconds before extracting the Log Mel spectrograms of 128 bins. The spectrograms are input to ImageBind-ViT [13] to obtain the audio semantic embeddings of length 1024. We further process these embeddings with four fully connected layers to extract the final audio semantic features of length 36. Detailed information is shown in the first column of Table 2. Other alternatives were explored for the extraction of audio features, from training CNNs on raw audio waveforms and log-mel spectrograms to employing pre-trained models like SoundNet [2]. However, ImageBind-ViT [13] achieved the highest performance, probably due to its superior audio semantic representation.

As a final step, the visual features, AEMs, and audio semantic features are concatenated before input into the *Audiovisual Fusion* module, forming a tensor of shape $(73, 120, 260)$. The Audiovisual Fusion module is built over two spherical convolutions with a kernel size of 3·3, the detailed layers are provided in the second column of Table 2. The output of the Audiovisual Fusion module, with a shape of $(36, 120, 160)$, is provided to the audiovisual decoder to extract the final saliency map of shape $(1, 320, 240)$. Note that only the final saliency map obtained after processing the whole sequence of 20 frames is used as the final prediction [3]. Once the whole sequence has been processed, the cell and hidden states of the encoder and decoder (formed by spherical ConvLSTMs) are reset.

S.3.1 Spherical ConvLSTMs

We follow the implementation of Bernal-Berdun et al. [3] for the spherical ConvLSTM employed in both our encoder and decoder. Convolutional LSTMs (ConvLSTMs) [20] replace the fully connected layers of traditional LSTMs with convolutional operations that additionally account for the spatial relationships of sequential data. Furthermore, spherical ConvLSTMs rely on spherical convolutions [10], which employ a distorted kernel to compute real pixels neighboring in spherical space, thus handling spatial features over time while accounting for the unique particularities of equirectangular representation. Our spherical ConvLSTMs can be expressed as follows:

$$\begin{aligned}
i_t &= \sigma\left(W_i * [e_t, h_{t-1}] + b_i\right) \\
l_t &= \sigma\left(W_l * [e_t, h_{t-1}] + b_l\right) \\
o_t &= \sigma\left(W_o * [e_t, h_{t-1}] + b_o\right) \\
g_t &= \tanh\left(W_g * [e_t, h_{t-1}] + b_g\right) \\
c_t &= l_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh\left(c_t\right),
\end{aligned} \tag{1}$$

where c and h are the cell and hidden states as in traditional LSTMs, $*$ spherical convolutions with 3×3 kernel size, $\odot$ the Hadamard product, and σ a sigmoid activation function. W_i, W_l, W_o, W_g and b_i, b_l, b_0, b_g are the learned weights and biases for the four spherical convolutions, whereas $[e_t, h_{t-1}]$ represents the concatenation of the frame with the hidden state at time t. The output of the spherical ConvLSTM is the last value of h_t.

S.3.2 Audiovisual Loss Function

Our novel audiovisual loss function can be computed as follows:

$$\mathscr{L}_{\text{AV}} = \gamma\, KLD(Q,P) + (1-\gamma)\ KLD(AEM,P), \tag{2}$$

where P and Q are the predicted and ground truth saliency maps, AEM is the audio energy map, and γ is a weighting parameter fixed at 0.75. We compute the Kullback-Leibler Divergence (KLD) terms as:

$$KLD(A,B) = \sum_{i,j} w_{i,j}\ A_{i,j}\ \log\left(\varepsilon + \frac{A_{i,j}}{\varepsilon + B_{i,j}}\right), \tag{3}$$

where

$$w_{i,j} = \Delta\varphi \frac{\sin(\Delta\theta j) + \sin(\Delta\theta(j+1))}{2\Delta\theta}, \tag{4}$$

$$\Delta\theta = \frac{\pi}{H}; \Delta\varphi = \frac{2\pi}{W}; i \in [0,W]; j \in [0,H]; \varepsilon = 10^{-8}. \tag{5}$$

A and B represent the evaluated saliency maps of shape $W \times H$, and $w_{i,j}$, is a spherical weighting that acknowledges the distortions introduced by the equirectangular projection, ensuring that the contribution of each pixel (i,j) is proportional to its solid angle [3, 25].

S.4 Metrics for Saliency Map Evaluation

To evaluate our model and compare it to the state of the art, we resort to the well-established metrics analyzed by Bylinskii et al. [5]. As they state in their work, there is no definite metric to evaluate saliency, since each one offers different information about the goodness of the predictions. In our main paper, we resort to Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), Similarity (SIM), and Root Mean Square Error (RMSE) to evaluate our model. Here, we also show results with the Area Under the ROC Curve (AUC) and Earth Mover's Distance (EMD) (which have a limited penalty imposed on false positives), and the Kullback-Leibler Divergence (KLD), which is employed during training of most approaches. We further outline the key features of each of these metrics in the following.

- **AUC:** The Area Under the ROC Curve primarily reflects the presence of highly valued salient areas. This ROC curve is generated by applying increasing thresholds to the predicted saliency map values and then assessing, for each threshold, the number of fixations within salient regions surpassing the threshold. Consequently, a larger AUC indicates a greater number of fixations falling within these salient regions. Given its reliance on high-valued predictions, AUC is less sensitive to low-valued false positives.

- **NSS:** Normalized Scanpath Saliency assesses the alignment between predicted saliency maps and actual eye fixations. This metric computes the average normalized saliency specifically at the locations of ground truth fixations. Importantly, NSS remains unaffected by parameters employed in generating the ground truth saliency maps, such as the sigma of the Gaussian kernel applied. NSS penalizes false positives as they do not contribute to the average value and reduce the values associated with true positives due to normalization.

- **CC:** Pearson's Correlation Coefficient evaluates the correlation between the distributions of two saliency maps, yielding a value of one when the distributions are identical and minus one when they are opposite. Because of its symmetrical nature, it equally penalizes false positives and false negatives. Figure 6 illustrates this behavior, where B2 saliency maps present fewer false positives, achieving a higher CC score, while in the last row, where both false positives and false negatives increase, the score notably decreases compared to the B2 maps in first and second rows.

- **SIM:** The similarity metric is obtained by summing up the minimum value between the saliency maps at each pixel. Consequently, the presence of a false positive in the predicted saliency map doesn't result in a penalty, as the metric will consistently select the minimum value at each pixel, which in this case corresponds to that of the ground truth saliency map.

- **KLD:** The Kullback-Leibler Divergence metric quantifies the difference between two probability distributions by measuring the amount of information lost when one distribution is used to approximate the other. Due to its sensitivity to zero values, KLD harshly penalizes a sparse set of predictions.

- **EMD:** Earth Mover's Distance quantifies the spatial disparity between two saliency maps by determining the minimum effort needed to transform one distribution into another. Particularly, saliency maps with dispersed density across a larger area tend to yield higher EMD values, indicating worse scores, as redistributing this excess density to match the ground truth map requires more effort. Moreover, EMD imposes a slight penalty on false positives that are spatially close to the ground truth.

- **RMSE:** The Root Mean Square Error metric gauges the disparity between predicted and ground truth saliency maps. It calculates the square root of the average of the squared differences between corresponding pixel values in both maps. Unlike metrics focused on area-based evaluations, RMSE directly measures the magnitude of the discrepancies across the entire map. Consequently, it is sensitive to both high and low-valued errors. In scenarios where false positives or false negatives occur, even at lower magnitudes, RMSE can be affected, impacting its overall value.

We also illustrate in Figure 6 how each metric might yield different results within different pairs of saliency maps.

S.5 Extended Evaluation and Comparative with Previous Works

In this section, we offer an extended discussion and performance evaluation of the state-of-the-art models for audiovisual saliency prediction over 360° video [8, 9, 26]. Table 3 presents quantitative results obtained across all metrics outlined in Section S.4. For each metric, we performed a Wilcoxon signed-rank test between the scores of our model and each of the state-of-the-art models [8, 9, 26]. The results can be found in Table 4, showing a statistical difference for all the metrics.

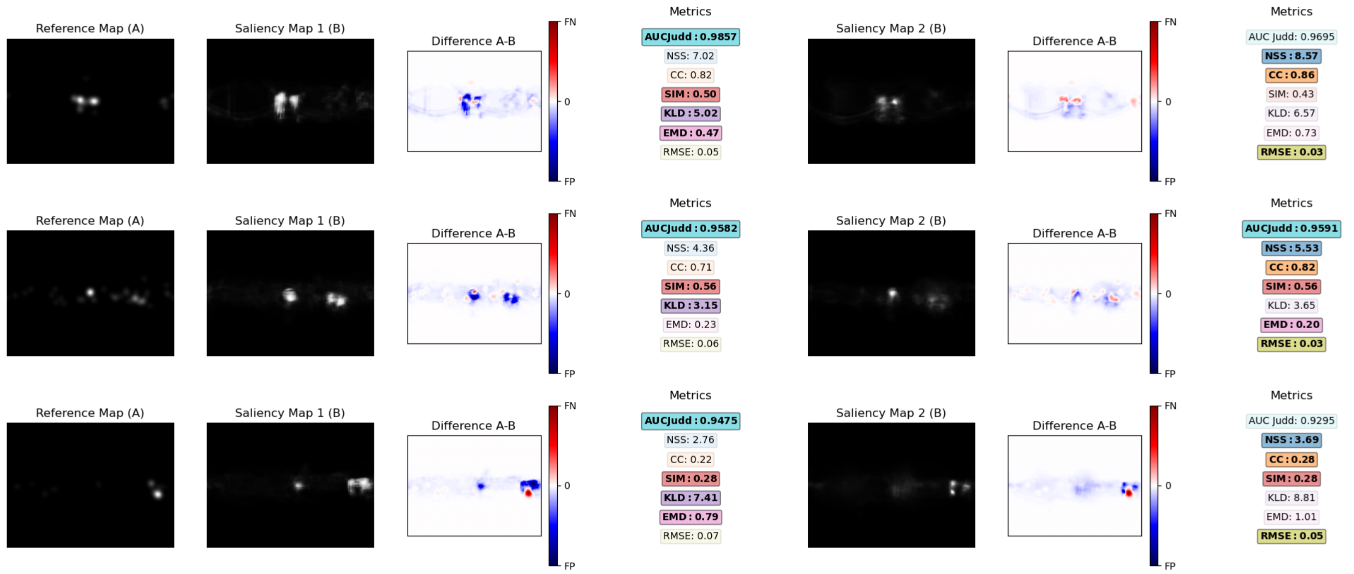

Figure 6: Each row illustrates a comparative analysis wherein, for a given reference saliency map A (first column), we calculate AUC_Judd, NSS, CC, SIM, KLD, EMD, and RMSE (see Section S.4) metrics for two distinct saliency maps, B1 (second column) and B2 (fourth column). Additionally, we illustrate the difference between each saliency map and the reference, with reddish colors indicating false negatives (FN) and bluish colors representing false positives (FP). The whiter the difference, the more similar the pair of maps are. The best metric between maps B1 and B2 is highlighted in boldface. While saliency maps B2 exhibit qualitative similarity to the reference map compared to B1, certain metrics (e.g., KLD, EMD, or AUC_Judd) do not accurately reflect this similarity due to their low sensitivity to false positives.

Inspired by Agrawal et al. [1], we aim to perform thorough analyses to assess the impact of the different audio cues on the final saliency predictions. To this end, we have introduced random audio inputs to AViSal360, AVS360 [8], and SVGC-AVA [26], both in terms of directional features (i.e., AEMs) and semantic features (as a waveform for AVS360 and ImageBind-ViT embeddings for AViSal360). For this comparative analysis, we exclude Cokelek et al.'s approach due to its non-data-driven nature. The results of these analyses can be found in Table 5.

Our results show that prior state-of-the-art models exhibit consistent or similar results with the introduction of random audio information to the ones obtained with correct audio, suggesting that their approaches are likely overlooking audio features, thus reducing their performance in cases where visual and auditory salient regions are not similar. Differently, our model experiences a notable drop in performance, which suggests that our model is indeed leveraging such audio features to provide more robust and accurate predictions.

As shown in Figure 8, AViSal360 can generally leverage audio semantic features to discern when the audio direction is relevant. This minimizes the negative effects introduced by reverberation or noises like wind in the AEMs, focusing on visually relevant features instead. Regardless, AViSal360's predictions on these cases degrade when no element is clearly visually salient (see Figure 7 of the main paper). The small number of videos in D-SAV360 presenting these conditions (no visually salient elements and noise/reverberation) difficult to learn these specific scenarios during training. This limitation could be mitigated with a larger and more diverse dataset.

S.6 Drawbacks of Optical Flow for Saliency Prediction

As discussed in Section 3 in the main document, our model builds upon a visual feature extraction module similar to SST-Sal [3], the state-of-the-art in visual saliency prediction. Importantly, different from them, we exclude the use of optical flow, as we have found two main occasions where its inclusion may hinder our model's ability to predict accurate saliency.

SST-Sal resorts to optical flow, which estimates the pixel movement across consecutive frames, as a measure of the movement of objects within a static scene. Nonetheless, when applied to videos captured with a moving camera, it does not serve to distinguish which objects are in motion, as the scene itself shifts along with them. The optical flow SST-Sal relies on is computed using the deep learning model RAFT [23]. We have empirically observed that this model tends to produce noisy optical flow estimates in scenes where there is either no movement or where the motion is subtle.

We have explored the impact of those cases in SST-Sal. We have found that in both cases (see Figure 7), and as briefly discussed by the authors, predictions are hindered and the model's accuracy

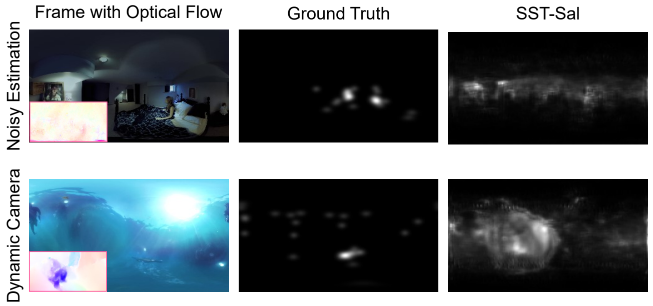

Figure 7: We have evaluated the impact of noisy optical flow estimations in the state-of-the-art visual saliency predictor SST-Sal [3] for cases with a complex scene with noisy optical flow (top) and cases where the camera is dynamic (bottom). For both, we show an RGB frame together with an inset with the optical flow estimation (left), and the ground-truth (center) and predicted (right) saliency. The inaccurate optical flow estimation in both cases significantly hinders SST-Sal's performance. Therefore, we decide to exclude optical flow from our visual feature extraction module, alleviating this limitation and favoring AViSal360's generalizability.

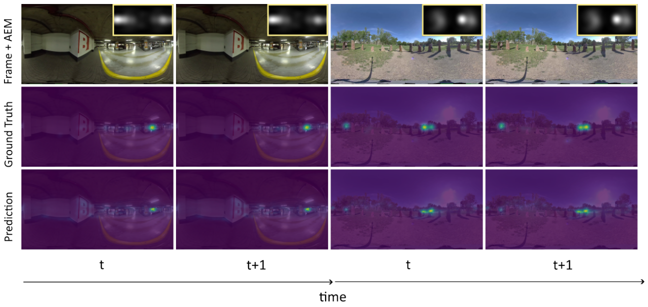

Figure 8: Additional qualitative results obtained from AViSal360 for two different videos (from left to right: *video_0001*, and *video_0011* in the D-SAV360 dataset). Both present scenarios where the captured audio is noisy and could hinder the prediction. In the first video (left), a strong case of reverberation is presented due to the architecture of the place, but the model pays no attention to the echo. For the second video, outdoor noise captured between the two mainly salient spots leads to a suboptimal audio representation. However, our proposed model avoids being mistakenly driven by the noise.

significantly drops. We therefore decided not to feed optical flow into our visual feature extraction module, ensuring that our model is not tied to only videos with static cameras and does not depend on having accurate optical flows.

S.7 Datasets of Viewing Behavior in 360° Content

With the proliferation of VR systems, we are witnessing an increased interest in better understanding the intricacies of viewing behavior in virtual environments. Datasets featuring head and gaze data from viewers for *visual-only* content are relatively abundant, both in static 360° images [12, 19, 22] and in 360° video [11, 14, 18, 24, 28]. However, datasets for *audiovisual* content are more scarce or exhibit limitations. Zhang et al. [27] collected around 1,000 scanpaths from 20 different viewers, but their sound was not spatialized, limiting the real impact of sound. Chao et al. [7] presented video with ambisonic (i.e., spatialized) audio, but only recorded head data (no gaze). In this work, we leverage the largest audiovisual 360° video dataset to date, D-SAV360 [4], which is currently the only publicly available dataset containing gaze data for 360° videos with spatialized audio. D-SAV360 features varied video content, and a diverse set of viewers, including a balanced number of female and male participants with varying levels of familiarity with VR.

S.8 D-SAV360 Videos for Audiovisual Evaluation

The D-SAV360 dataset [4] presents 87 videos exhibiting varying degrees of audio presence and its influence on visual attention. To capture a more representative scenario where audio significantly drives attention, we selected a subset of videos where audio plays a pivotal role. This subset was chosen based on the CC scores between the AEMs and the ground truth saliency, selecting the videos whose CC score is above a threshold set at 0.25. This subset includes 29 videos, enabling us to assess the impact of audio without dimming its influence among videos where it may be less relevant. The video IDs included in this subset are as follows: 5026, 0021, 5017, 0017, 0012, 0028, 1011, 1001, 1012, 5018, 5019, 1008, 5010, 5009, 0002, 1004, 2017, 0004, 1018, 0014, 0027, 2008, 0016, 0007, 0005, 0030, 5037, 1005, and 0013.

References

[1] R. Agrawal, S. Jyoti, R. Girmaji, S. Sivaprasad, and V. Gandhi. Does Audio Help in Deep Audio-Visual Saliency Prediction Models? ICMI '22, p. 48–56, 2022.

[2] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[3] E. Bernal-Berdun, D. Martin, D. Gutierrez, and B. Masia. SST-Sal: A spherical spatio-temporal approach for saliency prediction in 360º videos. *Computers & Graphics*, 2022.

[4] E. Bernal-Berdun, D. Martin, S. Malpica, P. J. Perez, D. Gutierrez, B. Masia, and A. Serrano. D-sav360: A dataset of gaze scanpaths on 360° ambisonic videos. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[5] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[6] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969. doi: 10.1109/PROC.1969.7278

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

[8] 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 *International Conference on Visual Communications and Image Processing (VCIP)*, pp. 355–358. IEEE, 2020.

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

[10] B. Coors, A. P. Condurache, and A. Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[11] X. Corbillon, F. De Simone, and G. Simon. 360-degree video head movement dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 199–204, 2017.

[12] A. De Abreu, C. Ozcinar, and A. Smolic. Look around you: Saliency maps for omnidirectional images in vr applications. In *International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, 2017.

[13] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *ArXiV*, 2023.

[14] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu. 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 211–216, 2017.

[15] L. McCormack, S. Delkaris-Manias, and V. Pulkki. Parametric acoustic camera for real-time sound capture, analysis and tracking. In *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17), Edinburgh, UK*, pp. 5–9, 2017.

[16] L. McCormack and A. Politis. Sparta & compass: Real-time implementations of linear and parametric spatial audio reproduction and processing methods. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[17] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang. Self-supervised generation of spatial audio for 360°video. In *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[18] C. Ozcinar and A. Smolic. Visual attention in omnidirectional video for virtual reality applications. In *International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2018.

[19] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Proc. of ACM Multimedia Systems Conference*, pp. 205–210, 2017.

[20] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. p. 802 – 810, 2015.

[21] A. Singla, T. Robotham, W. Menz, A. Raake, and E. A. P. Habets. Saliency of omnidirectional videos with different audio presentations: Analyses and dataset. In *15th International Conference on Quality of Multimedia Experience*, pp. 1–6. Ghent, Belgium, 2023. doi: 10.1109/QoMEX58391.2023.10178588

[22] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in VR: How do people explore virtual environments? *IEEE Trans. on Visualization and Computer Graphics*, 24(4):1633–1642, 2018.

[23] Z. Teed and J. Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[24] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze prediction in dynamic 360° immersive videos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 5333–5342, 2018.

[25] Y. Xu, Z. Zhang, and S. Gao. Spherical DNNs and Their Applications in 360º Images and Videos. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2021.

[26] Q. Yang, Y. Li, C. Li, H. Wang, S. Yan, L. Wei, W. Dai, J. Zou, H. Xiong, and P. Frossard. Svgc-ava: 360-degree video saliency prediction with spherical vector-based graph convolution and audio-visual attention. *IEEE Transactions on Multimedia*, 2023.

[27] Y. Zhang, F.-Y. Chao, and L. Zhang. ASOD60K: An audio-induced salient object detection dataset for panoramic videos. *ArXiV*, 2021.

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

Table 3: Extended quantitative comparison of the proposed AViSal360 to the state-of-the-art audiovisual saliency models: AVS360, SVGC-AVA, and Cokelek's proposal for audio inclusion fused with SST-Sal. The values presented for each model are the average mean scores for each of the videos in the D-SAV360 dataset, the average standard deviation is shown between parenthesis, and the confidence intervals between brackets. Upward and downward arrows for each metric indicate whether higher or lower values represent better performance. Bold text indicates the best model. Our model AViSal360 archives the best score among the audiovisual state-of-the-art models.

Model	CC ↑	NSS ↑	SIM ↑	RMSE ↓	AUC ↑	EMD ↓	KLD ↓
Cokelek et al. + SST-Sal	0.356 (0.105) [0.336,0.376]	2.414 (0.815) [2.212,2.615]	0.294 (0.066) [0.281,0.307]	0.141 (0.030) [0.137,0.145]	0.834 (0.06) [0.821,0.847]	1.453 (0.620) [1.347,1.559]	8.269 (1.728) [7.913,8.625]
AVS360	0.252 (0.100) [0.224,0.280]	1.544 (0.665) [1.351,1.737]	0.236 (0.054) [0.221,0.252]	0.121 (0.020) [0.113,0.129]	0.793 (0.062) [0.774,0.812]	1.503 (0.351) [1.378,1.627]	9.343 (1.430) [8.936,9.343]
SVGC-AVA	0.248 (0.091) [0.224,0.271]	1.505 (0.580) [1.343,1.668]	0.244 (0.051) [0.231,0.258]	0.094 (0.007) [0.090,0.097]	0.881 (0.027) [0.872,0.890]	1.345 (0.309) [1.241,1.449]	8.804 (1.202) [8.467,9.141]
AViSal360	**0.462** **(0.116)** **[0.435,0.490]**	**3.543** **(1.196)** **[3.164,3.922]**	**0.339** **(0.059)** **[0.325,0.353]**	**0.068** **(0.015)** **[0.063,0.072]**	**0.911** **(0.028)** **[0.901,0.921]**	**0.978** **(0.344)** **[0.908,1.049]**	**7.155** **(1.274)** **[6.840,7.469]**

Table 4: Wilcoxon signed-rank test comparison between our model AViSal360 and each state-of-the-art audiovisual saliency model: AVS360, SVGC-AVA, and Cokelek's approach for audio inclusion fused with SST-Sal. For each metric, we present the associated p-value, z-value, and effect size *r*. Results support that the performance improvement exhibited by AViSal360 is statistically significant.

Model		CC	NSS	SIM	RMSE	AUC	EMD	KLD
Cokelek et al. + SST-Sal	p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	z-value	-7.153	-7.412	-6.549	-8.008	-8.003	-7.421	-6.719
	r-value	0.053	0.037	0.091	<0.001	<0.001	0.037	0.080
AVS360	p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	z-value	-7.368	-7.605	-7.469	-7.767	-7.960	-6.943	-7.232
	r-value	0.040	0.025	0.034	0.015	0.003	0.066	0.048
SVGC-AVA	p-value	<0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	z-value	-7.442	-7.718	-7.333	-6.855	-6.443	-5.725	-6.535
	r-value	0.035	0.018	0.042	0.072	0.098	0.143	0.092

Table 5: Random Inputs. Quantitative evaluation while providing random audio input to our model and the state-of-the-art models AVS360 and SVGC-AVA. The values presented for each model are the average mean scores for each of the videos in the D-SAV360 dataset, and the average standard deviation is shown between brackets. Upward and downward arrows for each metric indicate whether higher or lower values represent better performance. AVS360 and SVGC-AVA achieve similar performance when random audio is provided, which suggests a poor use of the audio information.

	Random AEM				Random Semantic				Correct Inputs			
Model	CC ↑	NSS ↑	SIM ↑	RMSE ↓	CC ↑	NSS ↑	SIM ↑	RMSE ↓	CC ↑	NSS ↑	SIM ↑	RMSE ↓
AVS360	0.248 (0.099)	1.521 (0.656)	0.234 (0.054)	0.123 (0.018)	0.252 (0.100)	**1.547** **(0.663)**	**0.245** **(0.054)**	0.122 (0.017)	0.252 (0.100)	1.544 (0.665)	0.236 (0.054)	**0.121** **(0.016)**
SVGC-AVA	0.248 (0.091)	1.505 (0.580)	**0.245** **(0.051)**	0.094 (0.007)	-	-	-	-	0.248 (0.091)	1.505 (0.580)	0.244 (0.051)	0.094 (0.007)
AViSal360	0.431 (0.121)	3.302 (1.190)	0.311 (0.061)	0.070 (0.016)	0.451 (0.115)	3.418 (1.160)	0.332 (0.059)	0.071 (0.016)	**0.462** **(0.116)**	**3.543** **(1.196)**	**0.339** **(0.059)**	**0.068** **(0.015)**