AViSal360: Audiovisual Saliency Prediction for 360° Video

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

Universidad de Zaragoza, I3A

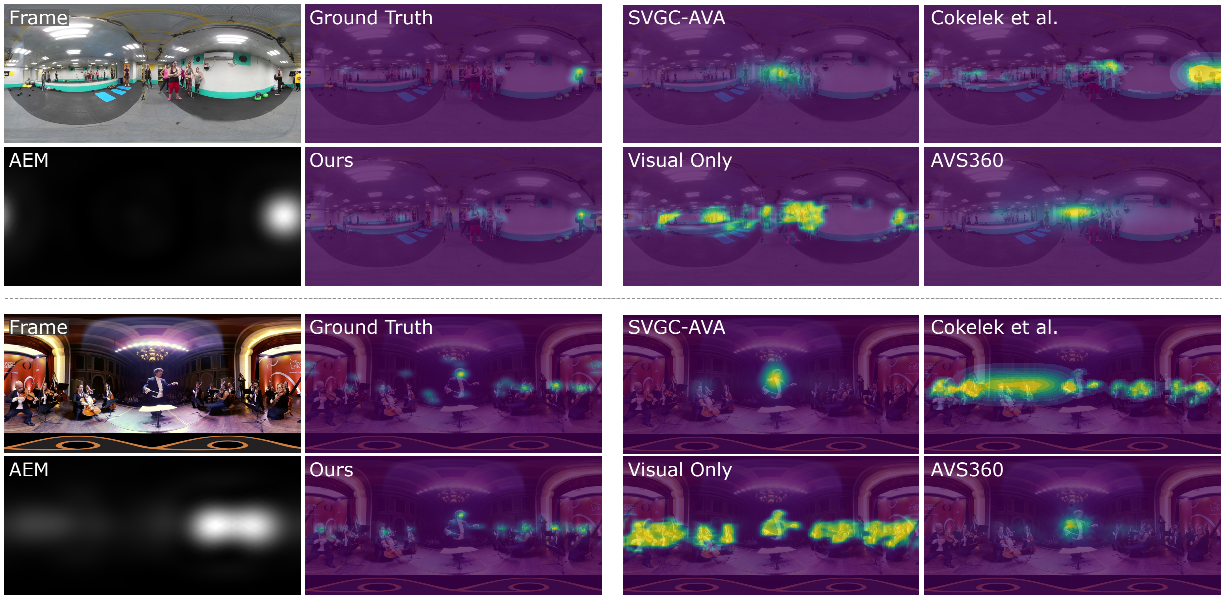

Figure 1: We introduce AViSal360, a novel audiovisual saliency prediction model for 360º video. By leveraging spatial and semantic audio information, in combination with an audiovisual saliency loss function, our model produces informed and accurate predictions. For two scenes (*gym*, *top*, and *orchestra*, *bottom*), we show: an input RGB frame; the corresponding audio energy map (AEM), representing audio spatial localization and computed as proposed in the present work; the ground truth saliency map; and the saliency predicted by our method. For comparison purposes, we also include the saliency map computed with our method without audio information (visual-only), and the resulting saliency map from three different state-of-the-art methods: AVS360 [10], Cokelek et al.'s method [13] combined with the recent SST-Sal [4], and SVGC-AVA [57]. Our model is able to disambiguate between multiple visually-salient regions (e.g., people) in cases in which saliency is conditioned by auditory information, predicting more accurately the ground truth. The *gym* and *orchestra* videos are *video_5018* and *video_5009* in the D-SAV360 dataset [5], respectively.

Abstract

Saliency prediction in 360° video plays an important role in modeling visual attention, and can be leveraged for content creation, compression techniques, or quality assessment methods, among others. Visual attention in immersive environments depends not only on visual input, but also on inputs from other sensory modalities, primarily audio. Despite this, only a minority of saliency prediction models have incorporated auditory inputs, and much remains to be explored about what auditory information is relevant and how to integrate it in the prediction. In this work, we propose an audiovisual saliency model for 360° video content, AViSal360. Our model integrates both spatialized and semantic audio information, together with visual inputs. We perform exhaustive comparisons to demonstrate both the actual relevance of auditory information in saliency prediction, and the superior performance of our model when compared to previous approaches.

Index Terms: Audiovisual Saliency, Visual Behavior, Ambisonic Sound, 360° Videos.

1 Introduction

Understanding and modeling visual attention in dynamic 360° content can play a key role in different stages of the content creation and visualization pipeline in virtual reality (VR), from content design, to gaze-contingent rendering, compression algorithms, or quality assessment techniques. While certain principles and techniques can certainly be borrowed from attention modeling in conventional 2D media, immersive environments have intrinsic characteristics that make them different, mainly the ability of the user to control the camera, and the placement of content *surrounding* the viewer, i.e., with content outside the field of view (FoV) of the user at any time instant. Since the study by Sitzmann et al. [46], one of the first to thoroughly analyze viewing patterns in 360° content, a number of works have focused on building computational models of visual attention, both in 360° images [2, 31] and video [16, 4]. These works often predict attention in the form of *saliency maps*, i.e., spatial representations of how conspicuous each point in a scene is [4, 50, 11].

Most of the approaches so far have focused solely on visual input, yet multimodality plays a key role in virtual reality. Input stimuli in VR are typically not only visual, but also involve other sensory modalities, such as auditory, proprioceptive, or olfactory inputs [30, 18]. Besides vision, auditory input is arguably the most common, and it can have a large impact on visual attention [9, 23, 34]. Examples of this are shown in Figure 1, where, for

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

two different scenes, we can see how audio clearly affects saliency: The presence of sound coming from specific regions of the scene (illustrated by the audio energy maps, AEMs) guides visual attention towards those areas, and away from others which could also be potentially salient, as shown by the ground-truth saliency maps. Besides, and as mentioned before, the existence of content outside the FoV of the user increases the relevance of sound, since it can provide information on events or objects outside this FoV. This can also make audio a powerful cue to guide attention in VR, an outstanding and much-researched problem [41, 44].

Interactions between visual and auditory modalities, however, can be complex. Empirical research has shown, e.g., that the inclusion of audio stimuli can make visual stimuli more salient, even if not co-located [52]. At the same time, spatially incongruent auditory stimuli outside the field of view have been shown to cause visual degradation in the viewer [29]. On a higher, more applied level, audio has also been shown to impact the perceived rendering quality [32, 35]. Despite the multiple advances in psychology and neuroscience (see, e.g., the review by Fu et al. [19]), the complexity of the interactions and the lack of a unified theory of attention have motivated the use of data-driven techniques for computational modeling of audiovisual saliency [10, 13, 57, 61].

In this paper, we present AViSal360, a robust and accurate saliency prediction model for audiovisual 360° video. Our model takes as input 360° RGB frames with equirectangular projection and their corresponding audio in first-order ambisonics format[1], and learns to predict per-frame saliency. This is done through two separate branches that extract relevant features from the visual and auditory inputs, respectively. These are then fused into audiovisual features, and fed into a decoding stage for the generation of the final saliency map. Motivated by studies indicating that audio information is often not effectively leveraged by audiovisual saliency prediction approaches [1], we propose four key aspects that improve audio processing and consideration, contributing to our model's success.

First, previous methods [57, 10, 61] have taken into account spatial audio location through so-called *audio energy maps* (a representation of sound spatial location, examples can be seen in the insets in Figure 1, left column), computed from first-order ambisonics [37]. Instead, we upscale the audio signal to fourth-order ambisonics, and employ a Minimum Variance Distortionless Response (MVDR) method [8] that maximizes power in the signal direction, effectively reducing noise and interference in the audio energy map. Second, while previous approaches directly rely on the raw audio signal or its spectrogram [10, 13], or extract features from it using a classification network [61], we leverage audio embeddings from a single unified embedding space that binds multiple sensory inputs together, offering a more reliable semantic representation [20]. Third, we employ a spherical convolutional long-short term memory (ConvLSTM) network not only in the visual encoding stage but also in the audiovisual *decoding* stage, allowing us to more effectively capture the intricate spatio-temporal dependencies between audio and visual compared to conventional 3D convolutions. Finally, to help the model learn the relevance of audio and its location, we employ a loss function that includes not only a traditional data term based on the Kullback-Leibler divergence between ground truth and predicted saliency, but also an additional term that enforces similarity between the predicted saliency and the audio energy map. This term is based on an analysis of the correlation between gaze fixations and audio energy maps.

We compare AViSal360 to state-of-the-art audiovisual video saliency prediction approaches, and show that our model consistently outperforms them. Besides, we have conducted thorough ablation studies, and assessed the relevance of audio information on the predictions, both in our and other methods. Our method is able to effectively leverage audio information, and combine it with visual features to generate reliable saliency maps in a wide variety of scenarios. This cannot only further our understanding of visual attention in audiovisual content but can also be leveraged in applications and techniques that ultimately lead to more engaging and appealing immersive experiences. Our code and model are available on the project page at https://graphics.unizar.es/projects/AViSal360_2024.

2 Related Work

2.1 Visual Saliency Prediction

Since the seminal work of Itti et al. [25] in the late 90s, a vast body of literature has been interested in studying and modeling saliency from visual inputs. The first works were built upon heuristic-based approaches, generally leveraging low-level features of the content (see, e.g., the review from Itti [24]). Later, with the proliferation of data-driven methods, deep learning-based approaches took over also in saliency prediction, and is and has been a very active area of research, whether in conventional 2D images [51, 48], video [49, 36], or 360° static content [38, 46]. Here, we focus on visual saliency prediction on 360° video and refer the interested reader to the thorough review of Borji [6] for the other modalities.

Dynamic content offers cues such as the movement of the elements (i.e., *bottom-up* attention) or the plot in a narrative sequence (i.e., *top-down* attention), which can affect viewers' attention and cause the saliency of each frame to be influenced by previous frames. One of the most common approaches to address this is the use of *long short-term memory* (LSTM) architectures [12, 16, 27], one particular type of recurrent neural network (RNN) that can retain temporal information and leverage it to perform posterior predictions. These works combine LSTMs with the well-established convolutional neural networks (CNNs), to enable the learning of spatio-temporal features that can predict saliency. Xu et al. [56] further proposed a saliency prediction network tailored to the particularities of 360° content, including spherical convolutions, pooling, and loss function. Later, Bernal-Berdun et al. presented SST-Sal [4], improving upon previous approaches by introducing a new spherical loss function based on Kullback-Leibler divergence, and a new paradigm of spatio-temporal feature extraction.

Visual saliency models, however, cannot account for the importance of audio when driving attention by themselves (see Figure 2, top right). To tackle this, our model is built over the architecture of SST-Sal for the visual feature extraction branch. Those visual features, together with our audio feature extraction and audiovisual fusion and decoding, allow us to overcome this limitation (Figure 2, bottom right).

2.2 Audiovisual Saliency Prediction

Conventional 2D Content Predicting saliency in audiovisual content is an outstanding problem that, within the last few years, has been mostly tackled by leveraging deep learning techniques. Tavakoli et al. [48] presented a two-branch model where they used 3D ResNets [22] to extract both visual and audio features, and Shunyu et al. [58] proposed a multi-scale spatial fusion to combine audio and visual features. Tsiami et al. [51] introduced STAViS, which combined 3D ResNets with attention modules for visual processing, and used SoundNet [3], a convolutional neural network pre-trained for sound classification, for audio processing. Jain et al. [26] introduced ViNet following a similar approach, although using a model pre-trained on action recognition tasks. Later, Zhu et al. [62, 60] presented LAVS, which combined a pre-trained VGG16 and LSTMs for visual processing, while using separable CNNs [43] for audio processing (i.e., CNNs able to process both time and frequency domains). Further approaches have been explored for audio

[1] First-order ambisonics is a common encoding for spatialized audio; for further explanation on it, please refer to Section S.2 in the supplementary material.

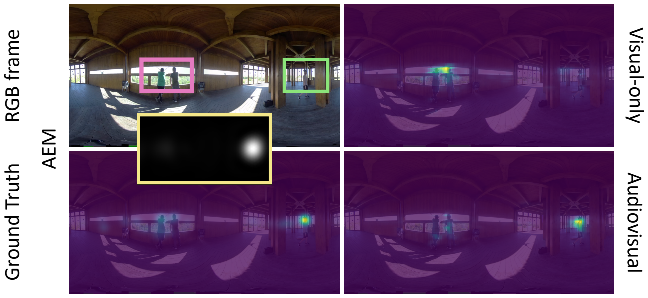

Figure 2: Limitations of visual-only saliency prediction models in contexts where audio plays a critical role (*video_0004* in the dataset). *Left column:* Sample RGB frame (*top*) and corresponding ground truth saliency (*bottom*). Despite the presence of a visually salient region consisting of two people talking (pink box), the ground truth indicates that attention is mainly drawn towards a third person creating noise (green box). Sound location is represented in the audio energy map (yellow). Visual-only models, such as SST-Sal [4] (*top right*), fail to account for this shift in attention due to the lack of auditory data integration. Our audiovisual model (*bottom right*) accurately predicts this shift, incorporating both visual and auditory cues effectively.

and visual feature fusion. While these works highlight the potential of data-driven techniques for the prediction of audiovisual saliency and an adequate representation of audio features, they are not directly applicable to saliency prediction in 360° content. VR incorporates an important attentional cue not present in traditional content, *spatial audio*. This new ability to locate sounds in 360° highly impacts visual attention, while models intended for traditional content do not contemplate it.

360° Content Critical differences in data representation and model architecture hinder the direct application of traditional 2D models to 360° content. The projection of a 360° environment onto a 2D plane introduces distortions or discontinuities that need to be carefully considered in model design. Furthermore, in 360° content, viewers control their point of view, and both visual and auditory stimuli surround them, potentially including conspicuous elements outside their field of view. Visual behavior in immersive environments exhibits different trends and biases compared to conventional 2D ones, with e.g., 360° content showing an equator bias rather than a center bias [46, 5]. Consequently, audiovisual saliency prediction in 360° content has emerged as a distinct subfield, leveraging insights from methods designed for traditional 2D content, but requiring careful consideration and adaptation to these unique challenges. Chao et al. [10] presented the first approach in this direction, although they only used spatialized sound (in the form of *audio energy maps*, AEMs) in inference time. Additionally, their methodology relied on an equator center bias that did not account for the longitudinal continuity of equirectangular content. Subsequent studies have also revealed an equatorial bias rather than a central one along a specific longitude [5]. Later, Cokelek et al. [13] introduced a method where they leverage any pre-trained visual saliency prediction module and fuse its output with a heuristic approach to extract spatial salient audio information. However, this approach only extracts the most salient sound on the scene, which may be overly simplistic in the case of multiple relevant audio sources. Moreover, since audio and visual branches are independent, inherent relations between sound and video are likely to be missed. Zhu et al. [61] have lately presented a model that combines the use of SoundNet [3] and AEMs for audio, and 3D LSTMs for video. SoundNet, an audio classification model, can extract semantic information from audio through transfer learning from a vision classification model. However, this reliance on visual classification labels is likely to introduce biases that affect the learned audio features; in other words, the learned audio representations may miss audio features that would be relevant for saliency prediction (but are not for visual classification). Finally, Yang et al. [57] recently used graph convolutional networks in their approach, albeit they did not consider any temporal or semantic audio information. In contrast with these methods, we leverage both spatial and semantic audio information (with a state-of-the-art audio representation) and take into consideration the temporal dimension of both the audio and video channels.

Datasets of Viewing Behavior in Audiovisual 360° Content Datasets featuring head and gaze data from viewers for *visual-only* content are relatively abundant, both in images [46, 40, 17] and in video [39, 15, 28, 59, 55]. However, datasets for *audiovisual* content are scarce or exhibit limitations. In this work, we leverage the largest audiovisual 360° video dataset to date, D-SAV360 [5], currently the only publicly available dataset containing gaze data for 360° videos with spatialized audio. More information on this and other existing datasets can be found in Section S.7 of the supplementary material.

3 A Model for Audiovisual Saliency Prediction

Our model for audiovisual saliency prediction takes as input a sequence of 360° RGB frames with equirectangular projection and the corresponding audio in first-order ambisonic format. We make this choice because these are the most common formats for 360° video and associated spatialized audio. The output of the model is a saliency map per input RGB frame. The model has an encoder-decoder architecture, with separate branches for encoding the audio and visual inputs. The audio and visual features extracted in the encoding branches are then fused, and fed into the decoding stage. An overview of this is shown in Figure 3, while the next subsections (Sections 3.1 to 3.3) explain each of the stages in detail, and motivate our design choices. This is followed by a description of our loss function (Section 3.4) and training details (Section 3.5).

3.1 Audio Features

Our prediction model takes into account *both* accurate spatial location of the audio sources (directional features), and the nature of this audio information (semantic features). For the former, instead of relying on the audio energy maps commonly used in previous works [10, 57], we process the audio input to improve sound localization. For the latter, some approaches have tried to use the raw audio signal or its spectrogram [10], or features extracted from it with a classification network [61]; instead, we extract semantic features with a state-of-the-art, modality-binding model.

Directional Features Spatial location of the auditory information is encoded in directional features. As mentioned above, it is common to find spatialized audio in first-order ambisonic format. From these, so-called audio energy maps (AEMs) can be computed, which depict the spatial distribution of incoming sound energy from the 360° soundscape, in equirectangular projection. AEMs computed from first-order ambisonics [37], commonly used in audiovisual saliency prediction [10, 57], cannot represent high spatial frequencies. This leads to a lack of precision when localizing sound sources that can potentially have an effect on saliency prediction, particularly when multiple nearby sound sources are present. To enhance precision and resolve possible ambiguities between sound sources, we upscale the audio to fourth-order ambisonics using an ambisonic upsampler [33], before computing the corresponding AEMs. Following the upsampling, we employ the Minimum Variance Distortionless Response (MVDR) method [8] to generate the AEMs. Unlike earlier methodologies [37], MVDR is a signal processing algorithm that minimizes total power at the receiver while preserving power in the signal direction, effectively reducing interferences, thereby reducing interferences and producing clearer AEMs. Figure 4 shows an example of this process,

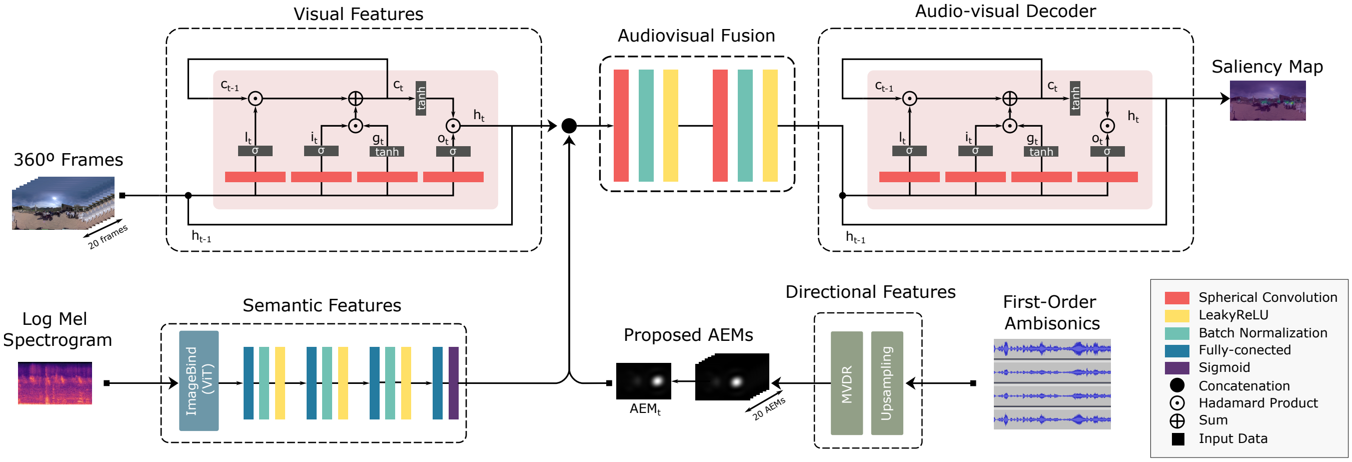

Figure 3: AViSal360's architecture. Our model adopts an encoder-decoder architecture with distinct branches for visual and audio feature extraction. Features are then concatenated and input into an Audiovisual Fusion module. This module, formed by spherical convolutional layers, processes the features to produce a cohesive audiovisual representation, which will then be processed by the decoder to generate the corresponding saliency map. Both the visual encoder and the audiovisual decoder utilize spherical ConvLSTMs, to facilitate learning of the spatial and temporal characteristics of the data. The audio encoder consists of two main components: one extracts semantic audio features using ImageBind-ViT [20], followed by embedding post-processing, while the other upscales ambisonic audio to fourth order and employs MVDR to enable precise AEM extraction. Further details about the model can be found in Section 3 and in the supplementary material.

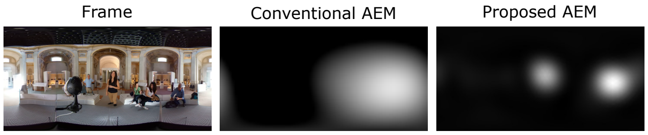

Figure 4: *Left:* RGB frame depicting a scene where audio originates from two key figures: a person singing in the center and another seated to the right, speaking (*video_5035*). Conventional AEMs (*center*) provide a basic representation of sound locations using first-order ambisonics. In contrast, our proposed approach (*right*) utilizes upsampled fourth-order ambisonics combined with the Minimum Variance Distortionless Response (MVDR) method, yielding more precise spatial locations of the sound sources.

where our approach for computing the final AEMs resolves the potential ambiguity between the different people (typically visually salient) present. More implementation details can be found in Section S.2 of the supplementary material. To test whether our proposed AEMs would, as hypothesized, be better representations of audio spatial location for saliency prediction purposes, we computed to what extent gaze fixations fell within areas of high energy in the AEMs. The results, showing that indeed the number was significantly higher in the case of the proposed AEMs, can also be found in Section S.2.1 of the supplementary material.

These proposed AEMs thus serve as a representation of audio directional features and will be concatenated together with the semantic audio features (explained next) and the visual features (Section 3.2) before decoding them into saliency predictions, as illustrated in Figure 3.

Semantic Features The nature of the audio information also plays a role in visual attention, since not all sounds carry equal significance. For instance, conversations tend to catch our attention, whereas sounds like birdsong may go unnoticed. Therefore, providing our model with audio semantic information can help it learn which sounds are salient. We explored several alternatives for audio semantic extraction (more details can be found in Section S.3 of the supplementary material), finally leading to the use of ImageBind-ViT [20]. ImageBind-ViT learns a single unified embedding space that binds multiple sensory inputs together, including audio and image data. The fact that the learned audio embeddings have been shown to work well for multiple zero-shot tasks, and their alignment to their corresponding image embeddings, support their suitability as semantic features for our task.

As mentioned above, visual features are concatenated with audio directional and semantic features, and fed into the decoding stage, explained in the next subsections.

3.2 Visual Features

Our visual encoder is built on a convolutional long short-term memory (ConvLSTM) architecture with spherical convolutions, which has shown to be successful for visual saliency prediction [4]. The use of LSTMs allows to capture and process temporal features from sequential data, inferring temporal relationships between them. ConvLSTMs [45], in particular, replace the fully-connected layers of traditional LSTMs with convolutional operations that account for the spatial relationships of sequential data. Finally, spherical convolutions, introduced by SphereNet [14], employ a distorted kernel to compute neighboring pixels in spherical space, accounting for the particularities of the equirectangular representation. The equations describing spherical ConvLSTMs can be found in Section S.3.1 of the supplementary material.

While our visual feature extraction is similar to that of SST-Sal [4], we exclude the use of optical flow (which is an input to their model). This allows us to lift one of their assumptions, that of having a static camera. We can thus handle videos in which optical flow may be misleading for saliency estimation, including cases in which the optical flow estimation is inaccurate, and cases in which both the camera and the scene are moving. An illustration of these cases can be found in Section S.6 of the supplementary material.

3.3 Audiovisual Fusion and Decoding

We leverage the previously extracted visual and audio features to derive audiovisual features, which are then used to predict the final saliency map. Initially, all feature vectors are introduced to an *Audiovisual Fusion module* to facilitate the learning of relationships among visual, directional audio, and semantic audio features. The joint inclusion of these features is important due to their intertwined nature. For instance, localizing semantic audio features within the image relies heavily on the information provided by directional audio features. The module comprises two spherical convolutions with LeakyReLU activations; further implementation details are provided in Figure 3, as well as in Section S.3 and Table 2 of the supplementary material. Then, the audiovisual features inferred by the Audiovisual Fusion module are decoded into the saliency map through another spherical ConvLSTM with specifications identical to the encoder. Unlike other audiovisual 360° saliency prediction approaches, our decoder also relies on ConvLSTMs, to account for the spatiotemporal characteristics of the features and the previous history at each time instant. This allows our model to establish temporal dependencies between visual and audio features.

3.4 Audiovisual Loss Function

Our loss function is based on the well-established Kullback-Leibler divergence (KLD) metric, enforcing similarity between two probability distributions, and is composed of two terms.

The first term is the traditional KL-divergence between the predicted saliency map P and the ground-truth saliency Q, $KLD(Q,P)$. In order to strengthen the influence of audio in the loss function score, helping the model learn salient regions when audio is the main feature driving attention, we add a second term that enforces similarity between the predicted saliency map P and the audio energy map (AEM), which can be regarded as a probability distribution map. This term is further supported by the observation that a large number of gaze fixations fall upon high density areas in the AEM (see Section 3.1). The relative contribution of both terms is weighted by a parameter γ, yielding our final audiovisual loss:

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

where:

$$KLD(A,B) = \sum_{i,j} w_{i,j}\ A_{i,j}\ \log\left(\varepsilon + \frac{A_{i,j}}{\varepsilon + B_{i,j}}\right), \tag{2}$$

with A and B representing two probability density functions as 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 [56, 4]. Implementation details of the spherical weighting can be found in Section S.3.2 of the supplementary material.

3.5 Training and Implementation Details

Our model is trained with the D-SAV360 dataset [5], since it is the largest currently available dataset containing 360° videos with ambisonic audio and gaze data. The 360° videos encompass varied content with indoor, outdoor, complex, and simple scenes of diverse topics (e.g., sports, music, or lectures). Videos were down-sampled from 60 fps to 8 fps and reshaped to a 320 × 240 resolution to reduce memory, computation, and processing requirements. Each video is then divided into subsequences of 20 frames, corresponding to 2.5 s. The ambisonic audio is also divided into segments of equal duration. D-SAV360 provides the saliency maps obtained from the recorded eye fixations of 87 participants.

AViSal360 has a size of 3,057 MB and the training took 5 h on a Nvidia RTX 3090 with 26 GB. We used the following hyperparameters for training: a stochastic gradient descent optimizer with a momentum of 0.9, a learning rate of 0.8, a batch size of 20 frames, and a total of 120 epochs. The value of γ in our loss function (Eq. 1) was empirically set to 0.75. To obtain a saliency map for each frame, we pass through the network the entire sequence of 20 previous frames (and corresponding audio information), providing the network with previous history, and achieving seamless and more consistent predictions. The average inference time was 118.36 ms (STD = 0.36 ms) with sequences of 20 frames on a GPU with the aforementioned specifications. Additional implementation and model details can be found in Section S.3 of the supplementary material.

4 Evaluation

In this section, we perform a thorough evaluation of AViSal360 using a k-fold cross-validation ($k = 5$) approach on the D-SAV360 dataset [5]. While this is done to mitigate the risks of overfitting and provide results in a wide variety of scenes, please note that the model never sees the scenes used for testing during training. Section 4.1 introduces the metrics we employ. Then, Section 4.2 presents and discusses results from our model, and Section 4.3 compares it to the state-of-the-art models on audiovisual saliency prediction. Finally, in Section 4.4 we conduct exhaustive ablation studies to endorse our design decisions.

4.1 Metrics

There exist many different metrics to evaluate saliency prediction models [7]. Each of them is more sensitive to some particular cases (e.g., having more true negatives or more false positives), or resorts to different criteria for the evaluation (e.g., distribution-based metrics vs. location-based metrics). We conduct our evaluations on a subset of four of such well-established saliency metrics: linear correlation coefficient (CC), similarity (SIM), normalized scanpath saliency (NSS), and root mean squared error (RMSE). We nevertheless compute additional metrics and discuss the suitability of all of them for the task at hand in Section S.4 of the supplementary material. We additionally refer the reader to the study by Bylinskii et al. [7] for an in-depth analysis of the particularities of each metric. In our work, we compute each of the metrics following the implementation proposed by Gutierrez et al. [21], where each pixel of the saliency maps is weighted by the sine of its latitude, therefore accounting for the distortion present in 360° content.

4.2 Results

We have first evaluated the quality of the audiovisual saliency maps predicted with AViSal360 with respect to the ground-truth ones. Figure 5 depicts four sample videos from the D-SAV360 dataset [5]; for each of them we show a sequence of two RGB frames, their corresponding audio energy maps (AEMs), and both the ground-truth and the predicted saliency maps. Each sequence spans 0.53 seconds of video to show temporal consistency. The first example (top left) showcases how our model properly leverages information from the AEMs: The scene is composed of two visually salient clusters of people, yet only the right-most one is emitting sound. Our model accurately focuses on it, significantly reducing the predicted attention over the cluster that emits no sound. In the second example (top right), our model, while attending to audio cues, also considers motion: The video shows an indoor cafeteria with moving people. AViSal360 can direct its attention to them momentarily, even if they are not the main sound source, closely mimicking the behavior captured in the ground truth. Our third example (bottom left) highlights our model's ability to discern the importance of sounds: In this scene, two people are speaking (left) in a tunnel with high reverberation (right). While the AEM points to two sources of sound, our model can distinguish which of them is actually emitting sound, directing attention to them, closely resembling the ground truth. The fourth case shows how our model successfully focuses on the most salient visual stimuli by taking into account the audio information. The attention is correctly directed to the group of people speaking, ignoring the bright-colored kayak that would be labeled salient by models focusing on low-level saliency. These four cases show both outdoor and indoor scenarios, and where audio played different roles, showcasing our model's versatility to adapt to each of them. Please refer to the project page for a web-based browser showing additional qualitative results from the D-SAV360 dataset.

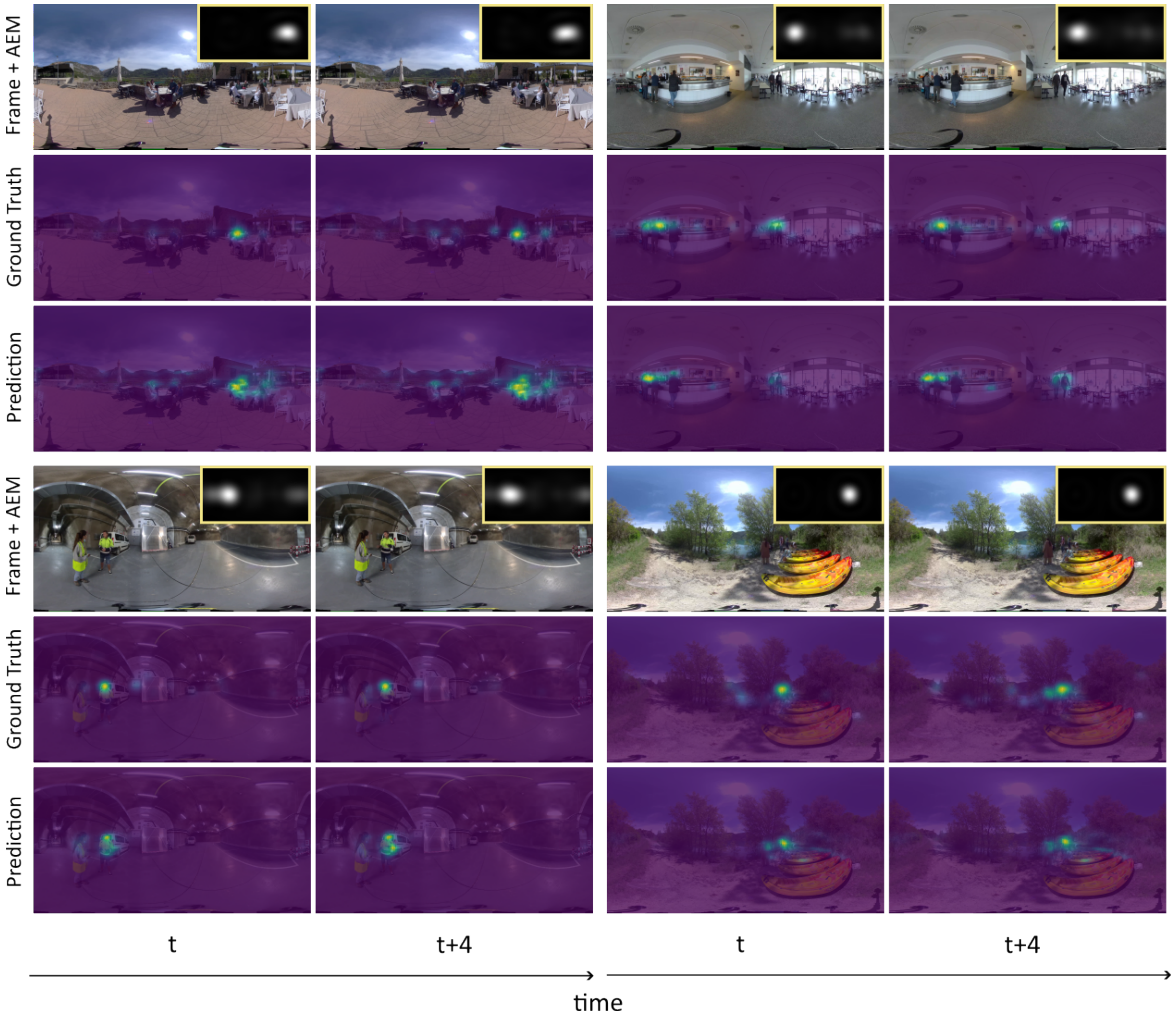

Figure 5: Qualitative results from AViSal360 for four different videos (in reading order: *video_0012*, *video_1001*, *video_1018*, and *video_0014* in the D-SAV360 dataset). We show a sequence of two RGB frames (spanning 0.53 seconds), their corresponding AEMs as insets, and both the ground-truth and predicted saliency maps. *Top left:* This example shows our model's ability to leverage the information from the AEMs to discern which visually salient features are relevant. *Top right:* Our model, while attending to audio, is also able to focus on other cues such as motion. *Bottom left:* AViSal360 can discern salient audio from reverberation or noise, focusing only on the former. *Bottom right:* Our model is not mistakenly diverted by low-level salient features, such as the bright-colored kayak, focusing on the actual salient area where a group of people are talking. Please refer to the project page for qualitative results on the whole dataset.

4.3 Comparison to Other Approaches

We compare AViSal360 to the three models that represent the state of the art of audiovisual saliency prediction for 360° video: AVS360 [10], SVGC-AVA [57], and Cokelek et al.'s proposal for audio inclusion [13] fused with the state-of-the-art model in visual saliency prediction, SST-Sal [4]. These models have already been shown to outperform visual-only and non-360° approaches thanks to the inclusion of spatial audio. We used each work's publicly available pre-trained model for our comparisons and evaluated their performance using the metrics described in Section 4.1.

Quantitative results can be found in Table 1, which shows how AViSal360 outperforms previous approaches by a large margin. To evaluate whether the improvement of our method over previous ones is statistically significant, we conduct a Wilcoxon signed-rank test. This test shows that the differences are indeed statistically significant in all cases ($p < 0.001$). Full results of the test, together with 95% confidence intervals for the means, can be found in Section S.5 of the supplementary material. AVS360 does not include AEMs in the training phase and Cokelek et al.'s model includes audio only as a bias during inference time, which we hypothesize hinders the models' ability to establish connections between audio and saliency depending on the context of the scene, ultimately impacting their achieved metrics. On the other hand, SVGC-AVA does not consider the semantics of audio, which may impact its ability to properly focus on the salient region. We have further evaluated the influence of audio on such models and found that using random audio inputs does not notably affect their final predictions, suggesting the models may not be fully leveraging spatial and semantic audio information. Further details on this study can be found in Section S.5 of the supplementary material. AViSal360 includes both audio spatial and semantic information during training, which makes it capable of learning meaningful connections between visual features, audio features, and saliency, overcoming previous works' limitations, and yielding superior metrics.

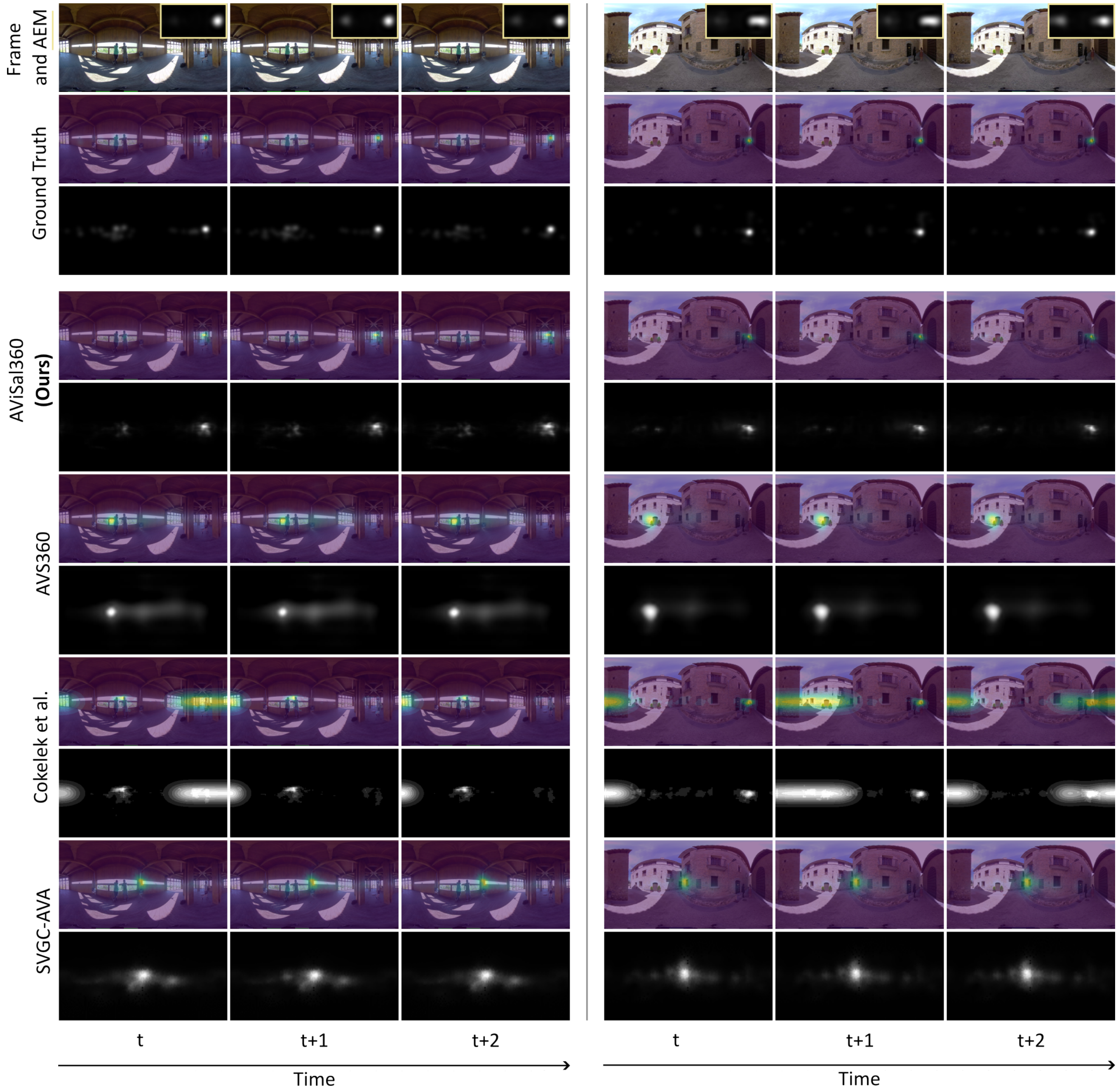

Figure 6: Qualitative comparisons of AViSal360 to AVS360 [10], Cokelek et al. with SST-Sal [13, 4], and SVGC-AVA [57] for two videos (from left to right: *video_0004*, and *video_0013* in the dataset). We show the sequence of RGB frames, their corresponding AEMs, ground-truth saliency maps, and the prediction with each method. AViSal360's predicted saliency better resembles the ground-truth, focusing on the actual salient elements, while the rest of the approaches do not fully reflect the influence of audio cues or introduce biases that yield inaccurate predictions.

We additionally show qualitative results of our comparisons in Figure 6. For two different videos from the D-SAV360 dataset, we show three consecutive RGB frames, their corresponding AEMs, the ground-truth saliency maps, and the predictions for AViSal360 and the three aforementioned state-of-the-art works. The left video depicts a case where there exist two visually salient regions (i.e., two people to the left, and one person to the right), but only the rightmost one is emitting sound, thus being the salient one. The right video shows a case where two people are emitting sound in the rightmost part of the scene, while some ambient noise sounds in the left part. AVS360 is strongly influenced by visually salient stimuli, missing the actual salient regions. Cokelek et al.'s approach applies a bias to their prediction based on spatial audio, which can lead to consecutive predictions being inconsistent, and a lack of spatial accuracy in the prediction. SVGC-AVA presents a strong center bias, which can cause it to miss the actual salient regions. In contrast, AViSal360 consistently focuses on the actual salient regions, correctly considering the relative importance of audio, and better resembling the ground truth.

Table 1: Quantitative evaluation. Comparison of our proposed AViSal360 to state-of-the-art audiovisual saliency models. For each model, we show average mean scores for each of the videos in the D-SAV360 dataset (and average standard deviation in brackets). Arrows indicate whether higher or lower values represent better performance; bold text indicates the best result. Our model achieves the best score, and differences with respect to the other methods are statistically significant with $p < 0.001$ (Wilcoxon signed-rank test, full results in Section S.5 of the supplementary material).

Model	CC ↑	NSS ↑	SIM ↑	RMSE ↓
Cokelek et al. + SST-Sal	0.356 (0.105)	2.414 (0.815)	0.294 (0.066)	0.141 (0.030)
AVS360	0.252 (0.100)	1.544 (0.665)	0.236 (0.054)	0.121 (0.020)
SVGC-AVA	0.248 (0.091)	1.505 (0.580)	0.244 (0.051)	0.094 (0.007)
AViSal360 (Ours)	**0.462** **(0.116)**	**3.543** **(1.196)**	**0.339** **(0.059)**	**0.068** **(0.015)**

4.4 Ablation Studies

We conducted several ablation studies to justify our design choices. We summarize them here, while detailed quantitative and qualitative results, along with an extended discussion, are available in Section S.1 of the supplementary material. First, we compared our proposed AEMs with the traditional ones [37], used in previous audiovisual saliency models [57, 9], showing how our enhanced AEMs led to improved saliency predictions. Next, we assessed the impact of including audio branches and our proposed training loss, by training our model without audio inputs (i.e., removing the audio branches and using a traditional KLD loss). We observed a drop in performance, both in the quantitative results (Table 1 in the supplementary) and in the qualitative comparisons (see Figure 1 here, and Figure 1 in the supplementary): Without auditory information, the model identifies all individuals and motion in the scene as salient, whereas our model can distinguish truly salient elements. Additionally, inspired by Agrawal et al. [1], we further tested the impact of each of the audio branches by introducing random audio information. The resulting decrease in accuracy highlights the model's reliance on precise directional and semantic audio information. We also conduct this evaluation for the state-of-the-art models AVS360 and SVGC-AVA (see Section S.5 of the supplementary material).

5 Discussion

While our model demonstrates a significant advancement in audiovisual saliency prediction for 360° video, the vast diversity of possible interactions between visual and auditory elements, each with its relative importance, and of semantic contexts, including e.g. narrative and non-narrative scenarios, constitutes a big challenge. The large dataset we employ enables good performance in a wide variety of scenarios (see video results in the project page), but increasing our understanding of how auditory and visual input combine and integrate to capture human attention in immersive environments can help improve generality [47, 19]. Further, while we use a state-of-the-art model to provide a suitable representation of semantic audio information [20], incorporating newer, more robust models as they become available could improve performance. Similarly, the data we use is collected from a large and diverse population sample (87 participants, balanced between female and male), but the collection of datasets with larger populations could be leveraged by our model and help better model inter-user variability.

We have observed that visual features typically focus on smaller, more specific regions, while spatial audio features extend over broader areas. Our Audiovisual Fusion module addresses these inconsistencies by learning the relationships between both visual and auditory distributions. It yields high saliency values in areas where visual and spatial audio features overlap, while diminishing emphasis in other visual areas where there is no sound source. This ensures that, despite the broader spatial coverage of audio, AViSal360 preserves the specificity of visual regions. However, some cases are particularly challenging for our method. Scenarios where the direction of sound is unclear, such as those with strong ambient noise (e.g., wind or background sounds) or reverberation. An example of this can be seen in Figure 7, which corresponds to a video recorded in a rooftop with strong wind. The model's pre-trained semantic audio branch, designed to match audio cues with corresponding visual data, may struggle with these scenarios, in which there are strong ambient noises that do not have a direct visual counterpart, together with no highly salient visual elements. Nevertheless, for mild cases of reverberation or echoes, AViSal360 still achieves a good performance if visual regions of interest are present in the scene, e.g., bottom-left video in Figure 5. In this scenario, the prediction correctly focuses on the people at the left of the frame, disregarding the reverberation coming from the right side, where no visual sources are evident. We show more examples of these challenging scenarios in Figure 8 of the supplementary material.

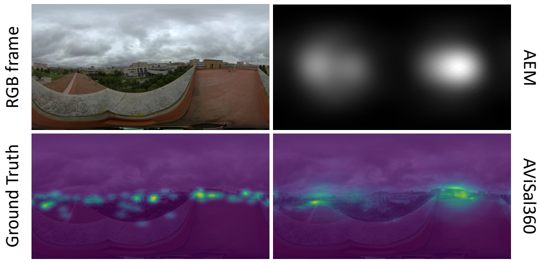

Figure 7: Limitations. In videos with strong ambient sounds and no clear visually salient regions, such as this windy rooftop (*video_1008*), our model primarily relies on directional audio cues and struggles to represent the ground truth saliency, which is distributed across the scene. Please see the text for more details.

As a saliency prediction model designed for audiovisual 360° videos, AViSal360 can enable a variety of applications. For instance, it can be leveraged by compression methods for immersive video [54], by informing algorithms of regions where more attention will be placed. Saliency prediction is also a valuable tool for content generation, contributing to tasks such as editing, content composition, cut alignment in cinematographic content, or thumbnail generation [42, 46, 31, 53]. Incorporating auditory input to provide robust audiovisual saliency prediction can thus be a relevant asset for the development of engaging virtual experiences.

Conclusion In this work, we have presented AViSal360, a novel audiovisual saliency prediction model specifically tailored for 360° content with ambisonic audio. In addition to the visual input, AViSal360 effectively leverages both spatial and semantic audio information, which are known to impact visual attention, achieving more accurate saliency predictions and surpassing the state of the art. To support future research and applications, both the trained model and code are publicly available at `https://graphics.unizar.es/projects/AViSal360_2024`.

Acknowledgments

We thank the members of the Graphics and Imaging Lab and Alfonso Ortega for their support and collaboration. This work was funded by the Spanish Agencia Estatal de Investigación (Project PID2022-141539NB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, EU), and J. Pina was supported by an FPI predoctoral grant. D. Martin and E. Bernal-Berdun were also supported by Gobierno de Aragon predoctoral grants (2020–2024 and 2021–2025).

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, 8

[2] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proc. International Conference on Computer Vision (ICCV) Workshops*, pp. 2331–2338, 2017. 1

[3] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 2, 3

[4] 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. 1, 2, 3, 4, 5, 6, 7

[5] 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. 1, 3, 5

[6] A. Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):679–700, 2019. 2

[7] 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. 5

[8] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969. doi: 10.1109/PROC.1969.7278 2, 3

[9] 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. 1, 8

[10] 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. 1, 2, 3, 6, 7

[11] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 1420–1429, 2018. 1

[12] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun. Cube padding for weakly-supervised saliency prediction in 360° videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1420–1429, 2018. 2

[13] 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. 1, 2, 3, 6, 7

[14] 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. 4

[15] 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. 3

[16] Y. Dahou, M. Tliba, K. McGuinness, and N. O'Connor. ATSal: An attention based architecture for saliency prediction in 360 videos. *Lecture Notes in Computer Science*, pp. 305–320, 2020. 1, 2

[17] 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. 3

[18] E. Doukakis, K. Debattista, T. Bashford-Rogers, A. Dhokia, A. Asadipour, A. Chalmers, and C. Harvey. Audio-visual-olfactory resource allocation for tri-modal virtual environments. *IEEE transactions on visualization and computer graphics*, 25(5):1865–1875, 2019. 1

[19] D. Fu, C. Weber, G. Yang, M. Kerzel, W. Nan, P. Barros, H. Wu, X. Liu, and S. Wermter. What can computational models learn from human selective attention? a review from an audiovisual unimodal and crossmodal perspective. *Frontiers in integrative neuroscience*, 14:10, 2020. 2, 8

[20] 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. 2, 4, 8

[21] J. Gutiérrez, E. David, Y. Rai, and P. Le Callet. Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images. *Signal Processing: Image Communication*, 69:35–42, 2018. 5

[22] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proc. International Conference on Computer Vision (ICCV) Workshops*, pp. 3154–3160, 2017. 2

[23] C. Harvey, K. Debattista, T. Bashford-Rogers, and A. Chalmers. Multi-modal perception for selective rendering. In *Computer Graphics Forum*, vol. 36, pp. 172–183. Wiley Online Library, 2017. 1

[24] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 2

[25] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 2

[26] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3520–3527. IEEE, 2021. 2

[27] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O'Connor, X. G. i Nieto, and K. McGuinness. Simple vs complex temporal recurrences for video saliency prediction. In *BMVC*, 2019. 2

[28] 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. 3

[29] S. Malpica, A. Serrano, D. Gutierrez, and B. Masia. Auditory stimuli degrade visual performance in virtual reality. *Scientific Reports (Nature Publishing Group)*, 2020. 2

[30] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. Multimodality in VR: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–36, 2022. 1

[31] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia. ScanGAN360: A Generative Model of Realistic Scanpaths for 360° Images. *IEEE Trans. on Visualization and Computer Graphics*, 28(5):2003–2013, 2022. 1, 8

[32] G. Mastoropoulou, K. Debattista, A. Chalmers, and T. Troscianko. Auditory bias of visual attention for perceptually-guided selective rendering of animations. In *Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East Asia*, pp. 363–369, 2005. 2

[33] 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. 3

[34] X. Min, G. Zhai, Z. Gao, C. Hu, and X. Yang. Sound influences visual attention discriminately in videos. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 153–158, 2014. 1

[35] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing*, 29, 2020. 2

[36] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Trans. on Image Processing*, 29:3805–3819, 2020. 2

[37] 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. 2, 3, 8

[38] A. Nguyen, Z. Yan, and K. Nahrstedt. Your attention is unique: De-

tecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1190–1198, 2018. 2

[39] 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. 3

[40] 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. 3

[41] S. Rothe and H. Hußmann. Guiding the viewer in cinematic virtual reality by diegetic cues. In *Augmented Reality, Virtual Reality, and Computer Graphics*, pp. 101–117, 2018. 2

[42] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia. Movie editing and cognitive event segmentation in virtual reality video. *ACM Trans. on Graphics*, 36(4), 2017. 8

[43] J. Sharma, O.-C. Granmo, and M. Goodwin. Environment sound classification using multiple feature channels and attention based deep convolutional neural network. 2020. 2

[44] A. Sheikh, A. Brown, Z. Watson, and M. Evans. Directing attention in 360-degree video. *IET Conference Proceedings*, pp. 29 (9 .)–29 (9 .)(1), January 2016. 2

[45] 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. 4

[46] 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. 1, 2, 3, 8

[47] C. Spence and C. Frings. Multisensory feature integration in (and out) of the focus of spatial attention. *Attention, Perception, & Psychophysics*, 82:363–376, 2020. 8

[48] H. R. Tavakoli, A. Borji, J. Kannala, and E. Rahtu. Deep audio-visual saliency: Baseline model and data. In *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–5, 2020. 2

[49] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *ArXiV*, 2019. 2

[50] M. Tliba, M. Sayah, and Y. A. D. Djilali. 2d-based saliency prediction framework for omnidirectional–360° video. In *11th International Conference of Pattern Recognition Systems (ICPRS 2021)*, vol. 2021, pp. 31–37, 2021. 1

[51] A. Tsiami, P. Koutras, and P. Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4766–4776, 2020. 2

[52] E. Van der Burg, C. N. Olivers, A. W. Bronkhorst, and J. Theeuwes. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1053, 2008. 2

[53] A. Vermast and W. Hürst. Introducing 3d thumbnails to access 360-degree videos in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2547–2556, 2023. 8

[54] M. Xu, C. Li, S. Zhang, and P. Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020. 8

[55] 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. 3

[56] 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. 2, 5

[57] 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. 1, 2, 3, 6, 7, 8

[58] S. Yao, X. Min, and G. Zhai. Deep audio-visual fusion neural network for saliency estimation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1604–1608, 2021. 2

[59] 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. 3

[60] D. Zhu, X. Shao, Q. Zhou, X. Min, G. Zhai, and X. Yang. A novel lightweight audio-visual saliency model for videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(4), 2023. 2

[61] D. Zhu, K. Zhang, N. Zhang, Q. Zhou, X. Min, G. Zhai, and X. Yang. Unified audio-visual saliency model for omnidirectional videos with spatial audio. *IEEE Transactions on Multimedia*, 26:764–775, 2024. 2, 3

[62] D. Zhu, D. Zhao, X. Min, T. Han, Q. Zhou, S. Yu, Y. Chen, G. Zhai, and X. Yang. Lavs: A lightweight audio-visual saliency prediction model. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021. 2