Beyond Short Segments: A Comprehensive Multi-Modal Salience Prediction Dataset With Standard-Length 360-Degree Videos

Anonymous Author(s)

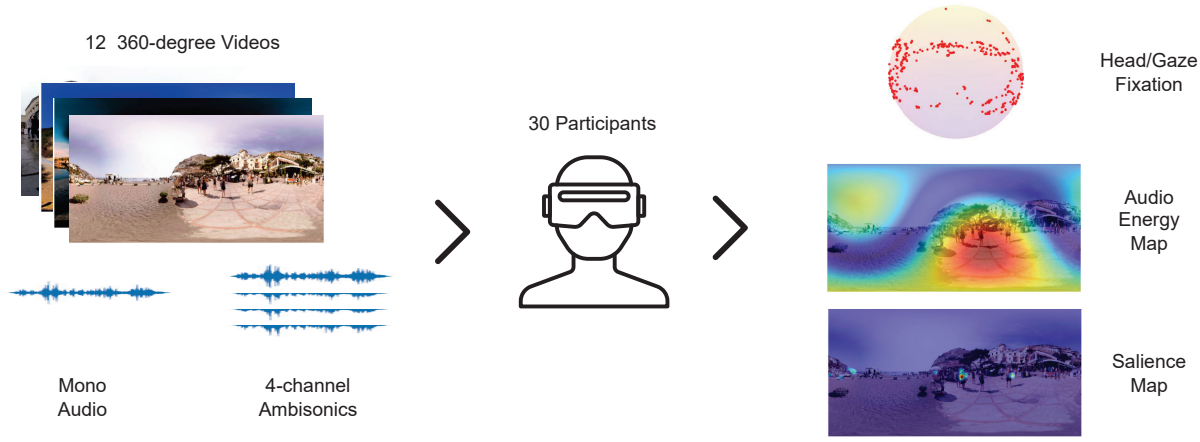


Figure 1: This work produces a novel multi-modal 360 degree video salience dataset with 12 standard-length video and collected observational data from 30 participants under both mono and ambisonic sound conditions.

ABSTRACT

Understanding user interactions in immersive 360-degree video environments is essential to optimizing both the user experience and transmission technology, such as advanced data compression and transmission over the Internet. Unlike flat videos, 360-degree videos enrich sensory experiences with complex ambisonic spatial information, posing unique challenges for multi-modal salience prediction. Recent research has introduced various 360-degree datasets equipped with ambisonic sound, but these primarily contain only short video segments, typically under 30 seconds. This limitation raises concerns about the generalizability of findings, particularly for the specific ambisonic features of 360-degree videos, leading to optimisation errors in the delivery. To overcome these challenges, we developed a comprehensive multi-modal, standard-length 360-degree video salience dataset and analyzed interactions from 30 participants under both mono and ambisonic audio settings. Our study reveals a detailed relationship between ambisonic audio distribution and viewer attention, underscoring the issues and complexities of applying leanings from short-segment to longer formats. Furthermore, we assess existing salience prediction models and introduce an efficient baseline model to evaluate the impact of different modal features in our dataset. Our findings indicate that ambisonic features, unlike visual and audio features, pose significant challenges when extrapolated from short video segments to longer video segments. The insights and the new dataset from this current study establish a more realistic benchmark for future research on multi-modal salience prediction in 360-degree videos. Our proposed dataset is available at https://anonymous.4open.science/r/ambi360dataset-F263(temporary anonymous link).

CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**; • **Human-centered computing** → **Virtual Reality**.

KEYWORDS

360 degree video, Ambisonic Sound, Salience Prediction, Multi-modal Learning

ACM Reference Format:

Anonymous Author(s). 2025. Beyond Short Segments: A Comprehensive Multi-Modal Salience Prediction Dataset With Standard-Length 360-Degree Videos. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

360-degree panoramic videos have become increasingly popular in virtual reality environments. Optimizing the transmission of these videos requires a deep understanding of user behavior, which is crucial for developing effective video compression and adaptive streaming techniques [8, 27]. For instance, predicting saliency maps

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and head movements can significantly enhance the processes [21, 25].

Previous research has predominantly focused on the visual aspects of 360-degree videos [23]. However, recent studies have started integrating both visual and auditory features to provide a more holistic view of user interaction within these immersive environments [3, 5]. The audio features in 360-degree videos, mainly using 4-channel B-format first-order ambisonic sound, differ markedly from traditional flat videos. This ambisonic audio transmits not only sound, but also embeds spatial information that indicates the direction of sound sources in the environment [31].

Sound features have been shown to effectively depict user behaviour in flat video studies [20], raising questions about the significance of spatial audio information in 360-degree videos. Despite the growing interest in spatial audio information, most research has focused on short video segments (less than 30 seconds) [3, 5], driven by the ease of collection and testing of short videos, a common practice for studying audio features in flat videos [20]. However, unlike audio features, the spatial information of ambisonic sound has a relatively weaker impact on user perception. It remains unclear whether this influence on user behavior is consistent between short videos and longer ones.

To address this issue, we introduce a novel dataset of 12 standard-length videos, ranging from 40 to 240 seconds, available in both ambisonic and mono audio formats. We collected head and eye tracking data from 30 participants to capture the 360-degree experience accurately and examine the influence of ambisonic sound on user behaviour over typical video durations. To our knowledge, this is the first dataset of its kind designed specifically for behavioural analysis in standard-length ambisonic 360-degree videos.

We further conducted a detailed analysis of how user behaviour varies with different sound types and video durations to explore whether the spatial information from ambisonic sound can be consistently applied across entire videos. We discuss how deep learning models can simultaneously process multi-modal information, including visual, audio, and ambisonic spatial data. Building on a comprehensive review of previous work, we developed a baseline model to assess whether ambisonic sound improves the prediction of user salience maps. Our focus was on evaluating whether the information learned from video segments can be generalised to entire videos.

In summary, our main contributions are as follows:

- We generated a standard-length 360-degree ambisonic video dataset, which included 12 videos from 40 to 240 seconds.
- We collected observational data from 30 participants with both head and gaze tracking data. Half of the observational data was collected under ambisonic sound conditions, while the other half was collected under mono sound conditions.
- Our in-depth analysis of longer video formats has yielded mixed results concerning the impact of ambisonic spatial information on participant behaviour. These findings contrast with earlier studies that were focused on shorter videos, suggesting that both the type and length of the videos significantly influence user experience. This disparity underscores the need to consider video duration and content characteristics when evaluating the effectiveness of ambisonic features in salience predictions.
- We established a multi-modal salience map prediction baseline model and demonstrated (using our dataset) that ambisonic spatial information cannot be effectively learned from video segments and generalized to entire videos.

2 RELATED WORK

2.1 360-Degree Video Dataset

// todo add mono + ambi reason + FOA

Optimizing the compression and transmission processes of 360-degree videos necessitates a deep understanding of user behaviour. Early studies primarily focused on visual features and their correlation with viewer salience. For example, the Salient360! Grand Challenges [10] propelled advancements in salience prediction by providing benchmark platforms and datasets, which facilitated more accurate predictions of viewer attention. Subsequent research expanded the collection of 360-degree content across various domains [9, 14, 23, 28], demonstrating the feasibility of salience map predictions in this medium. However, these datasets often lack ambisonic sound, omitting a crucial aspect of the immersive 360-degree video experience and potentially skewing analyses of the impact of audio on viewer attention. It is worth noting that xxxx .

The significant role of audio features in salience prediction has been established in studies on flat videos [11]. Addressing this, Zhang [26] introduced ASOD60K, an extensive audiovisual 360-degree dataset that includes visual, audio, and gaze tracking data. Nevertheless, the audio component was presented in mono, mirroring the format used in flat videos, which overlooks the spatial nuances of ambisonic sound. To rectify this weakness, Chao et al. [5] and Bernal-Berdun et al. [3] developed datasets that include ambisonic environments, making them comprehensive resources for studying salience in 360-degree videos so far. However, their focuses on short video segments raise questions about the generalizability of their findings to longer video formats. We believe that the analysis of standard-length 360-degree videos is needed.

2.2 Audio-Visual Salience Prediction in 360-Degree Videos

The introduction of various 360-degree video datasets has marked significant progress in research in this area. Initially, studies focused on visual modalities, with pioneers like Lebreton et al. [12] adapting heuristic models from 360-degree images to videos. Advances in deep learning, particularly through regression methods employed by Xu et al. [22] and Zhu et al. [30], extended these models to accommodate continuous video inputs. However, the traditional application of CNNs to equirectangular projections failed to address the inherent distortion in 360-degree video geometry. Researchers have since explored alternative formats like cube padding to mitigate these issues [7, 16] and developed specialized convolutional kernels for 360-degree content [28].

Despite these improvements, the absence of audio and ambisonic spatial information in many studies limits the overall predictive accuracy. Recognizing this limitation, Chao et al. [6] proposed a multi-modal salience prediction model that integrates visual, audio, and ambisonic spatial information, addressing distortions with cube

Table 1: Comparison of multi-modal 360 degree video dataset

Dataset	# Videos	Duration	Resolution	Mono Audio	Ambi Audio	# Observers	Head Data	Gaze Data	Public Available
360AVD [18]	256	10s	1K to 4K	✓	✗	-	✗	✗	✓
ASOD60K [26]	67	30s	4K	✓	✗	20(mono)	✓	✓	✓
D-SAV360 [3]	85	30s	4K	✗	✓	87(ambi)	✓	✓	✓
Li et al. [13]	46	15s	4K	✓	✓	15(mono)+15(ambi)	✓	✓	✗
SVGC-AVA [24]	57	28s	4K	✓	✓	21(mono)+21(ambi)	✓	✓	✗
Chao et al. [5]	15	25s	4K	✓	✓	15(mono)+15(ambi)	✓	✗	✓
Ours	12	**40s-240s**	4K	✓	✓	15(mono)+15(ambi)	✓	✓	✓

map projections and utilizing 3D Resnets to process multi-modal data [17]. Recent efforts by Zhu et al. [29] and Yang et al. [24] have further refined these models, enhancing their ability to leverage audio and spatial information. However, the scarcity of comprehensive multi-modal datasets and the reliance on short video segments continue to pose challenges in assessing the generalization of these findings to full-length videos.

In response, we present a novel dataset of standard-length 360-degree videos to evaluate the generalization capabilities of different modalities in deep learning models. This study aims to bridge the research gap by examining how effectively audio and ambisonic spatial information learned from short segments can be applied to longer video formats, thus, providing detailed insights to the field of salience prediction in immersive video environments using multi-modal information.

3 DATASET OVERVIEW

3.1 360-degree Video Data Collection

Prior to data collection, we established a benchmark for the typical length of 360-degree videos by leveraging the YouTube Search API [1]. We analyzed a sample of 1000 360-degree videos from YouTube, examining their metadata to ascertain their durations, as depicted in Figure 2. Our findings revealed that 360-degree videos generally exhibit a shorter, more concentrated length distribution compared to standard flat videos. This trend is primarily due to the increased likelihood of motion sickness in longer 360-degree video formats. However, the duration metrics, including a first quartile of 92 seconds, a median of 148 seconds, and a third quartile of 242 seconds, confirmed that most 360-degree videos substantially exceed 30 seconds in length. Consequently, for our dataset, we selected video lengths ranging from 40 to 240 seconds. Following this analysis, we randomly chose 12 videos from YouTube, all encoded in 4K resolution and featuring 4-channel B-format ambisonic sound, with frame rates between 25 and 30 fps. These selections cover a diverse range of environments, both indoor and outdoor. To quantify the visual diversity of our dataset, we computed Temporal Information (TI) and Spatial Information (SI) metrics, comparing these to those from a previous study by Chao et al. [5] as shown in Figure 3. SI assesses the complexity of each frame, while TI measures changes between consecutive frames. Our dataset demonstrated a comparable SI to the prior dataset, indicating similar content richness. However, the TI was lower in our dataset likely due to the longer average video lengths that tend to preserve more complete information, thus, exhibit reduced temporal variation.

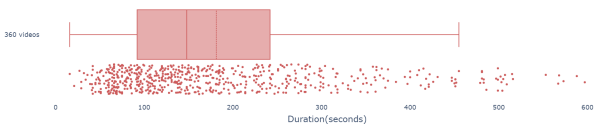


Figure 2: The duration distribution of 360-degree videos on the YouTube platform

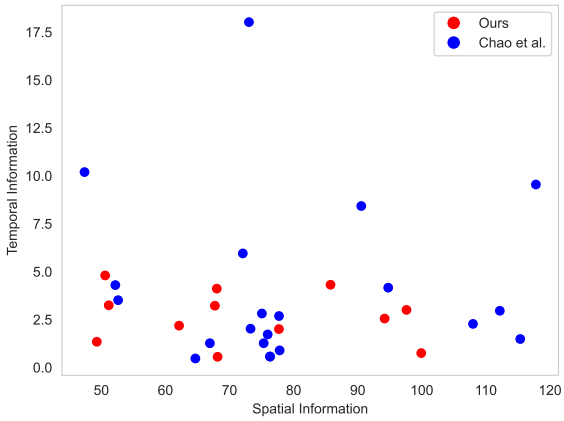


Figure 3: The comparison of Spatial Information(SI) and Temporal Information(TI) between our and Chao et al. [5] dataset

3.2 Participants Data Collection

We conducted experiments using the HTC Vive Pro Eye headset, which offers a visual angle of 110 degrees and a resolution of 1.5K per eye. The experimental environment was created in Unity. Due to thelimited support for ambisonic sound in Unity video player, we integrated the Oculus Native Spatializer plugin for high-quality ambisonic sound decoding. For data collection, the HTC head and eye tracking SDK was employed, selected for its rapid calibration capabilities and a 90Hz data capture rate, ensuring accurate tracking of participant movements.

Thirty participants were involved in the experiment, randomly assigned to view six mono and six ambisonic videos. This randomization ensured that each video was viewed by 15 participants in mono and another 15 in ambisonic sound settings. The demographic was composed of 19 males and 11 females, with 20% having no prior VR experience. Participants stood throughout the experiment, with eye-tracking calibration performed before each video. To mitigate fatigue, rest periods ranging from 2 to 5 minutes were allotted between videos.

3.3 Fixation and Salience Map Generation

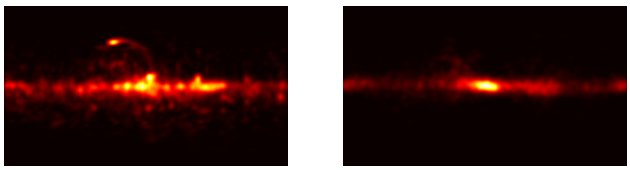
Following the methodologies from prior research [3, 5], a fixation is defined as a gaze that remains nearly stationary (movement under 3 degrees) for at least 200 milliseconds. To process gaze data, we initially applied the Identification by Dispersion-Threshold (I-DT) method to filter out saccade movements [4]. Eye fixations were then clustered using DBSCAN [19] for each frame across different participants, allowing us to generate individual and aggregated fixation data for both the ambisonic and mono sound groups.

Saliency maps were created from these fixation maps by counting the number of fixations per pixel for each frame (every 33ms). A Gaussian filter with a standard deviation of 5 degrees was applied to these counts to smooth the data and produce a continuous representation of areas most engaged by observers.

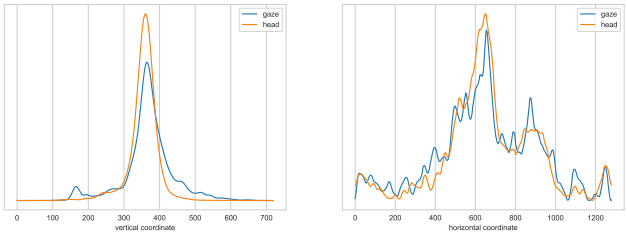
4 DATASET ANALYSIS

4.1 Head and Gaze tracking data

In prior VR research, head orientation data has been commonly used as a proxy for estimating gaze direction when calculating salience maps, largely due to its easier collection process [5, 14]. However, the accuracy of salience maps derived from head orientation relative to those obtained from direct gaze tracking in VR settings is not often examined. In our study, we concurrently gathered both head and gaze tracking data, allowing for a direct comparison between salience maps generated from each type of data.



(a) Average salience map generated by Gaze(left) and Head(right) data



(b) Salience distribution on Vertical(left) and Horizontal(right) directions

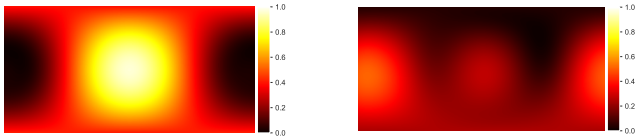
Figure 4: Average salience map and distribution

Figure 4a illustrate the average differences in the distribution of saliency maps based on head orientation and gaze data. Further analysis of these differences in the vertical and horizontal planes was conducted using max pooling, as shown in Figure 4b. Our findings indicate that while the saliency maps from both data sources align closely in the horizontal direction, significant discrepancies arise in the vertical distribution. The gaze-based saliency maps displayed a broader vertical spread, which can be attributed to the physical characteristics of VR headsets.

The weight and design of VR headsets tend to encourage users to maintain a balanced head position, preferring to use their eyes rather than their heads to look upwards or downwards. This behavioral tendency means that head-based salience maps often fail to capture true areas of interest along the vertical axis. Consequently, gaze tracking appears to be a more precise method for identifying eye fixations and, by extension, salient regions within VR environments. Additionally, our analysis confirmed the presence of an equator bias within our dataset, a phenomenon noted in previous research with very similar outcomes [3, 5, 24].

4.2 Ambisonic Audio Data

Ambisonic audio data in our study is encoded in a 4-channel B-format, comprising components W, X, Y, and Z. The W channel represents the omnidirectional audio signal, while X, Y, and Z channels encode the spatial information of sound sources. Following the methodology proposed by Morgado et al. [15], we decoded this 4-channel ambisonic audio into a single-channel audio signal to isolate the audio feature, and generated a 2D Audio Energy Map (AEM) to represent the spatial characteristics of the sound. An example of AEM is illustrated in Figure 5.



(a) Example of High entropy AEM

(b) Example of Low entropy AEM

Figure 5: Audio Energy Map(AEM) Example

To analyze the ambisonic spatial information comprehensively, we computed the entropy of the AEM for each video frame. AEM entropy, akin to the entropy of a saliency map, quantifies the complexity and diversity of the data based on Shannon entropy from information theory. High entropy, as depicted in Figure 5a, signifies concentrated audio sources with clear positional information. Conversely, low entropy, shown in Figure 5b, indicates a more scattered distribution of sound sources, suggesting the absence of a dominant audio source.

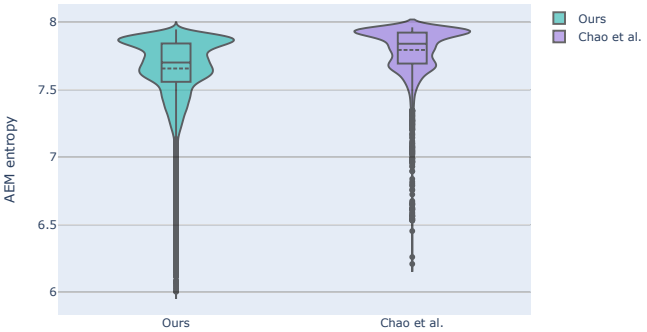


Figure 6: The comparison of AEM entropy violin plot between our and Chao et al. [5] dataset

Our analysis, visualized in Figure 6, compares the distribution of AEM entropy in our standard-length dataset against previous short-segment datasets. We found that standard-length videos typically exhibit a higher occurrence of low-entropy frames compared to the 30-second segments, indicating a broader and more diverse range of sound source distributions. This results in both median and mean entropy values being lower in our dataset, reflecting more realistic audio environments where multiple sound sources may be present simultaneously or intermittently, often without a distinct, dominant source.

4.3 Ambisonic spatial information and user attention

To investigate the correlation between audio spatial information and user attention, we adapted the Normalised Scanpath Salience (NSS) metric, which is traditionally used in saliency map prediction. This adaptation allows us to measure the correlation between participants' eye fixation positions and the Audio Energy Map (AEM), assessing the impact of audio spatial information on attention. Due to the inherent latency in human responses to audio cues, we modified the NSS calculation to include a temporal window, accommodating the delayed reaction time of participants following auditory stimuli. This is expressed in the adapted NSS Equation 1.

$$NSS = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=t}^{t+\Delta} Z_{AEM}(x_i^j) \tag{1}$$

where $Z_{AEM}(x)$ represent standardized AEM map,$\sum_{i=1}^{N}$ denotes the sum over all fixation points, and $\sum_{j=t}^{t+\delta}$ represents the sum over a temporal window from time t to $t+\Delta$. This extension accounted for potential behavioral delays, thus enhancing our ability to analyze the temporal correlation between AEM and user attention. In our calculations, we set Δ to 300 ms based on typical human reaction times to auditory stimuli.

Traditional methods calculate a single NSS value for short video segments, which suffices for brief analyses. However, our focus on standard-length videos necessitates a more detailed approach. We computed the NSS for each 30-second window within the videos to evaluate the temporal variability of attention response to ambisonic spatial information.

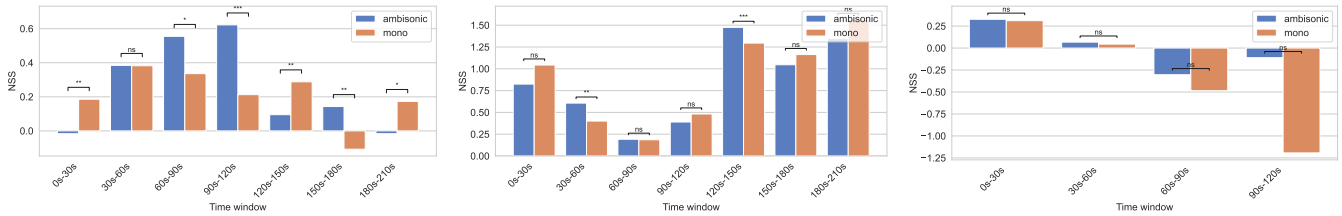
Three exemplar cases, shown in Figure 7, illustrate our analysis. For each time window, we averaged the NSS scores for all frames to obtain the window's mean score. The significance of differences between NSS scores from ambisonic and mono audio settings was tested using the non-parametric Wilcoxon signed-rank test after confirming data non-normality with the Shapiro-Wilk test. In the figure, pairs marked 'ns' represent no significant statistical difference, * indicates $.01 < p < .05$, ** indicates $.001 < p < .01$, and *** represents $p < .001$.

Our results, as described in Figure 7, indicate variable impacts of ambisonic spatial information on user behavior across different video segments. For instance, significant impacts on user attention are noted during specific intervals (e.g., 60s to 120s in Figure 7a), while other intervals show negligible or no significant effects. A similar pattern is observed in Figure 7b, where ambisonic spatial information shows significant impact from 0s to 30s and from 150s to 180s, but not in other segments. In contrast, Figure 7c illustrates a scenario where ambisonic spatial information does not significantly influence user attention at any point.

Moreover, the NSS values suggest varying degrees of alignment between the AEM and the saliency maps across different video segments. Positive values indicate a strong correlation where user attention aligns with sound source directions, while negative values suggest attention diversion from these areas.

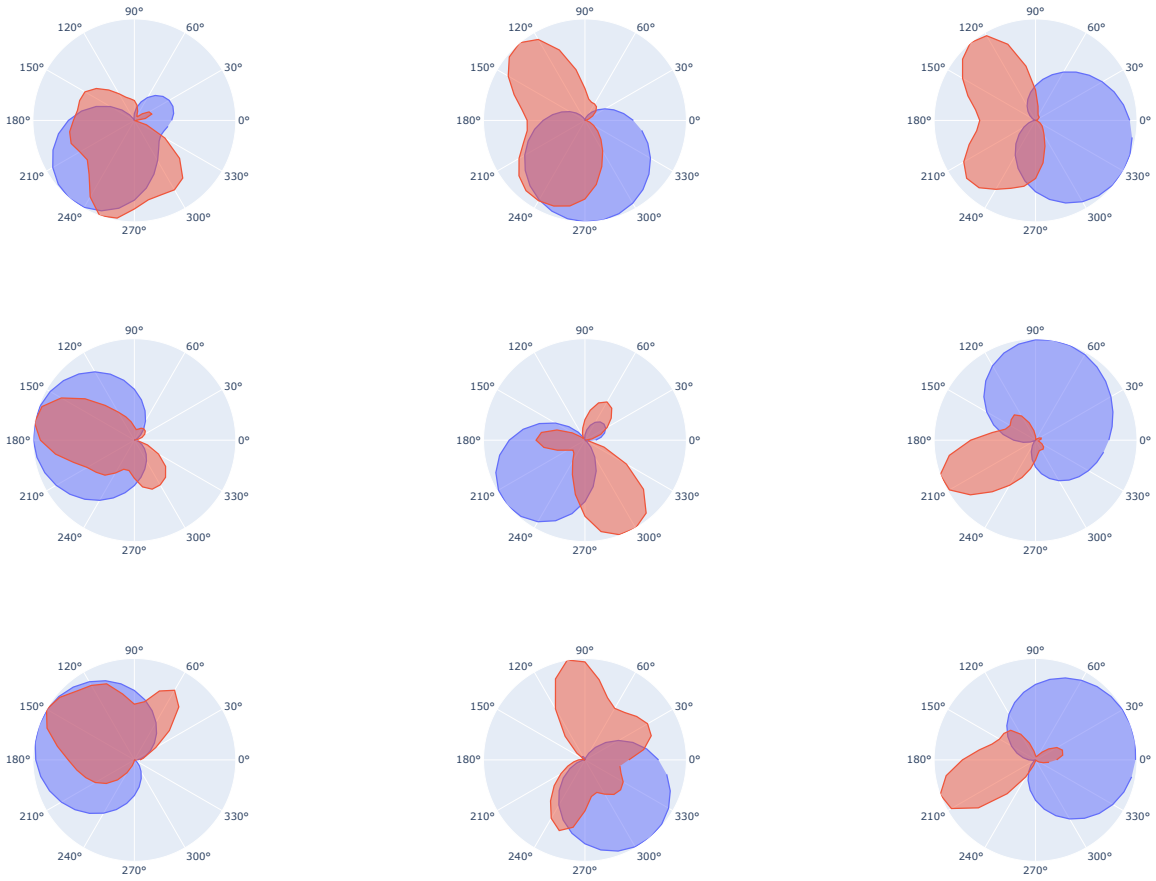
To further elucidate the relationship between AEM and user attention, we projected the 2D distribution map onto a circular horizontal plane, clearly visualized the directional relationship between the AEM and saliency maps(Figure 8). We demonstrate three different types of distributions to highlight the significant differences in the alignment between the AEM and saliency map distributions across different video segments:

- `Figure 8a`: Overlapping distribution, where there is a high degree of correspondence between the AEM and saliency maps, indicating aligned user attention with audio spatial cues.



(a) Example Sample 1 (b) Example Sample 2 (c) Example Sample 3

Figure 7: Comparison of NSS scores between ambisonic and mono sound groups with a statistically significant difference test.



(a) Overlapping distribution (b) Mixed distribution (c) Divergence distribution

Figure 8: Three different types of distributions in each column(Red refers to salience distribution and Blue refers to AEM distribution)

- `Figure 8b`: Mixed distribution, showing partial alignment, with areas of both high and low correlation between the audio cues and user attention.
- `Figure 8c`: Divergence distribution, where the AEM and saliency maps are largely uncorrelated, highlighting segments where spatial audio cues do not effectively guide user attention.

In conclusion, the influence of ambisonic spatial information on user behavior exhibits notable inconsistencies across standard-length video segments. This variability, alongside the diverse alignment of AEM with the salience map, contrasts with earlier findings from short video segment studies and underscores the complexity of assessing audio spatial impacts in more prolonged and immersive settings. This necessitates a nuanced approach to analyzing how temporal dynamics influence the relationship between spatial audio information and viewer attention.

5 MULTI-MODAL USER SALIENCE PREDICTION IN 360 DEGREE VIDEO

5.1 Baseline Model

For our experimental model, we focus on evaluating the generalization capabilities of features across videos of various lengths, which is critical for practical applications. We devise a baseline model for the purpose of evaluating xxxxx using xxxxx.

Our baseline model employs a lightweight 3D ResNet [17] for video feature extraction and SoundNet [2] for audio features, chosen for their efficiency and robust performance.

Our approach is different to prior work in multi-modal 360-degree salience prediction. Despite various methodologies being used, as described in Section 2.2, prior work generally adapts a three-stream architecture to handle different data modalities separately. The three-stream architecture processes video, audio, and ambisonic spatial information independently through dedicated feature extractor modules. This approach ensures that the unique characteristics of each modality are captured effectively and the extracted features are then amalgamated using advanced multi-modal fusion techniques, which may include methods such as concatenation [6], attention-based mechanisms [24], or weighted fusion [29]. The integration of multimodalities is crucial for leveraging the complementary nature of the modalities to enhance prediction accuracy.

Similarly, in our framework the fusion of these features is performed using a weighted strategy [29], aiming to optimize the contribution of each modality based on its relevance and impact on the prediction task. The fused features are then decoded into salience maps using a 1x1 convolution, ensuring precise and refined predictions.

To validate the effectiveness of our baseline model, we conducted comparative evaluations against prominent existing models using the well-known dataset from [5]. The results, summarized in Table **??**, demonstrate that our model achieves performance comparable to the state-of-the-art. These findings underscore the efficacy of our model in integrating multi-modal information and lay a strong foundation for further exploration on our extended dataset. Our approach is not only practical but also scale-able, making it suitable for broader applications in salience prediction research.

5.2 Experiment Setup

5.2.1 Dataset settings. To better compare with previous datasets and test the generalisation capabilities of different modality features, we constructed two distinct training and testing sets:

- `From Segment to Full Video:` Given that our videos are of varying lengths, we divided each video into 30-second segments. The first 30 seconds of each video were used as the training dataset, while the remaining segments longer than 30 seconds were used as the testing dataset. This setup allowed us to test whether features from different modalities learned from a segment of a video could generalize to the entire video.
- `From Full Video to Others` We selected 4 full length videos as the training data and the remaining 8 videos as the testing data. This method of data partitioning is consistent with previous datasets, allowing us to test whether features learned from full videos can generalise to other full videos.

5.2.2 Implementation Details. To minimize variables and enhance reproducibility, we strictly followed the implementation details from previous works [6, 29]. We constructed the baseline model using PyTorch and employed spherical KL divergence as the loss function.The 3d-Resnet and Soundnet component are both pre-trained the same as DAVE and Soundnet project separately. The AEM was pre-calculated at the same resolution as the video frames. We used the Adam optimizer to train the entire network and dynamically adjusted the learning rate based on the cosine annealing restart scheduler for 20 epochs. All experiments were conducted on a single Nvidia RTX 3090 GPU.

5.3 Experiment Result

To understand the generalization capabilities of different modality features under various dataset settings, we tested the baseline model in three scenarios: using only visual features, using both visual and audio features, and using visual, audio, and ambisonic spatial features. We evaluated the prediction accuracy of salience maps using four typical metrics: Normalized Scanpath Salience (NSS), Correlation Coefficient (CC), Kullback-Leibler Divergence (KLD), and the Area Under the Receiver Operating Characteristic Curve (AUC-J). Higher scores for NSS, CC, and AUC-J, and lower scores for KLD, indicate more accurate predictions. The results of all tests are presented in Table 2.

Our results show that it is evident that the inclusion of audio features significantly enhances salience prediction across various dataset settings. Models trained on 30-second segments performed well on standard-length videos, demonstrating that audio features possess strong generalization capabilities, allowing them to learn from segments and generalize to other parts. This finding aligns with observations from flat audiovisual salience prediction tasks.

However, the ambisonic spatial features did not improve salience prediction results on our dataset. Notably, features learned from 30-second segments significantly reduced the accuracy of salience prediction on full videos. Similarly, features learned from full videos did not positively impact the prediction of other full videos. This outcome is consistent with our discussion in Section 4.3, where we explored the correlation between AEM and the salience map. The AEM does not consistently align with the salience map distribution, and this complex relationship is challenging to capture from 30-second segments. Additionally, this relationship is difficult to generalize across different standard-length videos. These conclusions diverge from those obtained with previous datasets.

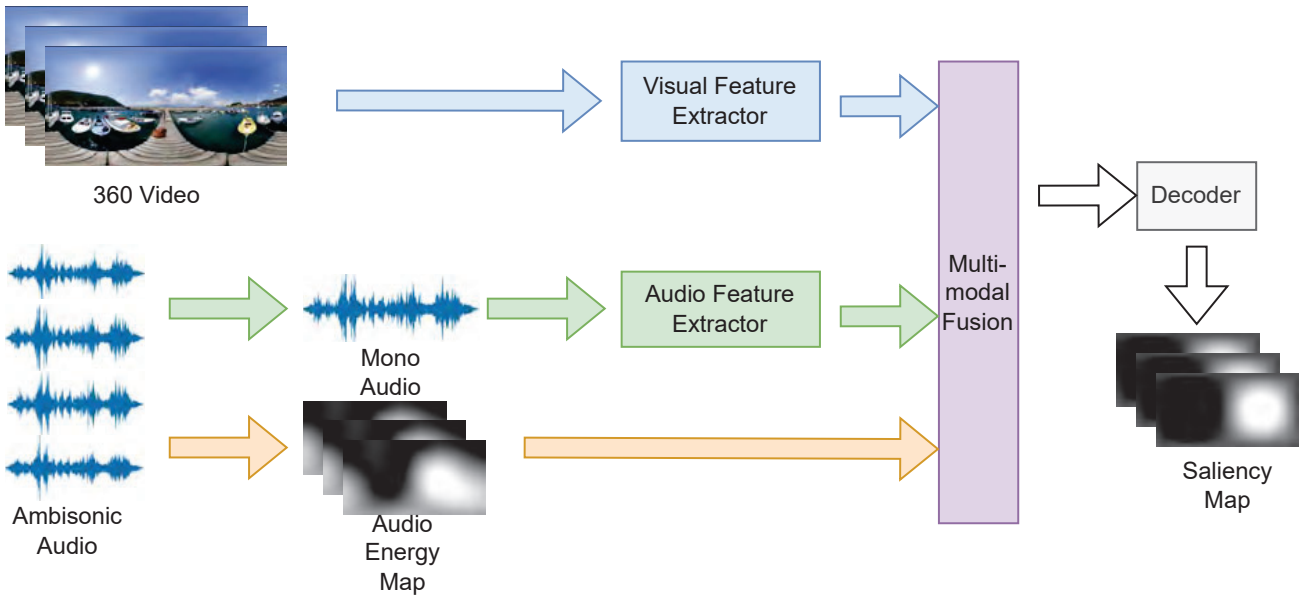


Figure 9: Basic architecture of multi-modal Salience Prediction in 360-Degree Video

Table 2: The impact of different modality features on salience prediction across various training sets on our dataset

Training set	Visual	Audio	Audio Spatial	NSS↑	CC↑	AUC-J↑	KLD↓
30s Segment	✓			1.673	0.559	0.824	3.043
30s Segment	✓	✓		**2.313** +38%	**0.692** +24%	**0.903** +9%	**2.060** +32%
30s Segment	✓	✓	✓	2.036 -12%	0.657 -5%	0.872 -3%	2.373 -15%
Full video	✓			1.557	0.489	0.785	3.445
Full video	✓	✓		**2.038** +31%	0.660 +35%	**0.841** +7%	**2.199** +36%
Full video	✓	✓	✓	1.957 -4%	**0.664** +0.6%	0.835 -1%	2.352 -7%

We attribute this discrepancy to our dataset, which more closely resembles real-world standard-length videos, encompassing more complex relationships between AEM and user behaviours.

Effectively utilizing ambisonic spatial features for predicting user behaviour remains a challenging problem. Despite their potential, our findings indicate that ambisonic spatial features do not consistently enhance salience prediction accuracy and may even reduce prediction performance when derived from short segments. This underscores the need for further research to develop methods that can better leverage the spatial audio information provided by ambisonic sound to accurately predict user attention in 360-degree videos.

6 LIMITATION

Our study utilized a dataset comprising 12 standard length videos, with a total video frame exceeding that of previous datasets consisting of <30s video segments. While this is a significant improvement, it remains limited compared to flat video datasets, particularly those used for training large-scale pre-trained models. The relatively short number of video may not provide sufficient data to fully capture the potential advantages of ambisonic spatial information for salience prediction.

7 CONCLUSION

In this paper, we introduce a novel dataset specifically designed for the study of multi-modal 360-degree video salience prediction, with a particular focus on the generalization capabilities of different modals on standard length 360 videos. Our analysis demonstrated that the incorporation of ambisonic spatial information presents unique challenges and does not consistently improve prediction performance. Our findings suggest that, while ambisonic spatial features have the potential to contribute to user behaviour prediction, their effective integration requires further investigation. Future work should aim to address these limitations by developing larger datasets and more sophisticated models to better leverage the rich information provided by different modalities.

REFERENCES

[1] [n. d.]. Search: list YouTube Data API Google for Developers — developers.google.com. https://developers.google.com/youtube/v3/docs/search/list. [Accessed 22-01-2024].

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems* 29 (2016).

[3] Edurne Bernal-Berdun, Daniel Martin, Sandra Malpica, Pedro J Perez, Diego Gutierrez, Belen Masia, and Ana Serrano. 2023. D-SAV360: A Dataset of Gaze Scanpaths on 360° Ambisonic Videos. *IEEE Transactions on Visualization and Computer Graphics* (2023).

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

[5] Fang-Yi Chao, Cagri Ozcinar, Chen Wang, Emin Zerman, Lu Zhang, Wassim Hamidouche, Olivier Deforges, and Aljosa Smolic. 2020. Audio-visual perception of omnidirectional video for virtual reality applications. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.

[6] Fang-Yi Chao, Cagri Ozcinar, Lu Zhang, Wassim Hamidouche, Olivier Deforges, and Aljosa Smolic. 2020. Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 355–358.

[7] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. 2018. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1420–1429.

[8] Lovish Chopra, Sarthak Chakraborty, Abhijit Mondal, and Sandip Chakraborty. 2021. Parima: Viewport adaptive 360-degree video streaming. In *Proceedings of the Web Conference 2021*. 2379–2391.

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

[10] Jesús Gutiérrez, Erwan J David, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360 contents. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–3.

[11] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. 2021. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3520–3527.

[12] Pierre Lebreton, Stephan Fremerey, and Alexander Raake. 2018. V-BMS360: A video extention to the BMS360 image saliency model. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–4.

[13] Jie Li, Guangtao Zhai, Yucheng Zhu, Jun Zhou, and Xiao-Ping Zhang. 2022. How sound affects visual attention in omnidirectional videos. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3066–3070.

[14] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 211–216.

[15] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. 2018. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems* 31 (2018).

[16] Minglang Qiao, Mai Xu, Zulin Wang, and Ali Borji. 2020. Viewport-dependent saliency prediction in 360 video. *IEEE Transactions on Multimedia* 23 (2020), 748–760.

[17] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.

[18] Aakanksha Rana, Cagri Ozcinar, and Aljosa Smolic. 2019. Towards generating ambisonics using audio-visual cue for virtual reality. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012–2016.

[19] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* 42, 3 (2017), 1–21.

[20] Hamed R Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. 2019. Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693* (2019).

[21] Shibo Wang, Shusen Yang, Hailiang Li, Xiaodan Zhang, Chen Zhou, Chenren Xu, Feng Qian, Nanbin Wang, and Zongben Xu. 2022. SalientVR: Saliency-driven mobile 360-degree video streaming with gaze information. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 542–555.

[22] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. 2018. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence* 41, 11 (2018), 2693–2708.

[23] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. 2018. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5333–5342.

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

[25] Abid Yaqoob, Ting Bi, and Gabriel-Miro Muntean. 2020. A survey on adaptive 360 video streaming: Solutions, challenges and opportunities. *IEEE Communications Surveys & Tutorials* 22, 4 (2020), 2801–2838.

[26] Yi Zhang. 2021. Asod60k: An audio-induced salient object detection dataset for panoramic videos. *arXiv preprint arXiv:2107.11629* (2021).

[27] Yuanxing Zhang, Pengyu Zhao, Kaigui Bian, Yunxin Liu, Lingyang Song, and Xiaoming Li. 2019. DRL360: 360-degree video streaming with deep reinforcement learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1252–1260.

[28] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. 2018. Saliency detection in 360 videos. In *Proceedings of the European conference on computer vision (ECCV)*. 488–503.

[29] Dandan Zhu, Kaiwei Zhang, Nana Zhang, Qiangqiang Zhou, Xiongkuo Min, Guangtao Zhai, and Xiaokang Yang. 2023. Unified Audio-visual Saliency Model for Omnidirectional Videos with Spatial Audio. *IEEE Transactions on Multimedia* (2023).

[30] Yucheng Zhu, Guangtao Zhai, Xiongkuo Min, and Jiantao Zhou. 2020. Learning a deep agent to predict head movement in 360-degree images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 4 (2020), 1–23.

[31] Franz Zotter, Matthias Frank, Franz Zotter, and Matthias Frank. 2019. XY, MS, and first-order Ambisonics. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (2019), 1–22.