360+x: A Panoptic Multi-modal Scene Understanding Dataset

Hao Chen Yuqi Hou Chenyuan Qu Irene Testini Xiaohan Hong Jianbo Jiao
University of Birmingham
h.chen.12, j.jiao@bham.ac.uk

Abstract

Human perception of the world is shaped by a multitude of viewpoints and modalities. While many existing datasets focus on scene understanding from a certain perspective (e.g. egocentric or third-person views), our dataset offers a panoptic perspective (i.e. multiple viewpoints with multiple data modalities). Specifically, we encapsulate third-person panoramic and front views, as well as egocentric monocular/binocular views with rich modalities including video, multi-channel audio, directional audio, location data and textual scene descriptions within each scene captured, presenting comprehensive observation of the world. To the best of our knowledge, this is the first database that covers multiple viewpoints with multiple data modalities to mimic how daily information is accessed in the real world. Through our benchmark analysis, we presented 5 different scene understanding tasks on the proposed 360+x dataset to evaluate the impact and benefit of each data modality and perspective. Extensive experimental analysis reveals the effectiveness of each data modality and perspective in enhancing panoptic scene understanding. We hope the unique dataset could broaden the scope of comprehensive scene understanding and encourage the community to approach these problems from more diverse perspectives.

1. Introduction

Scene understanding is crucial for robotics and artificial intelligent systems to perceive the environment around them. As humans, we intuitively understand the world through primarily visual inputs, as well as auditory and other sensory inputs (*e.g.* touch and smell).

The community has made remarkable progress in mimicking human perception with contributions from various datasets and benchmarks [4, 5, 7, 9, 13, 15, 23]. These efforts have approached scene understanding from a diverse range of perspectives, such as normal frontal-view vision [5, 13, 23], panoramic view [22, 28], egocentric binocular/stereo view [20, 30], egocentric monocular view [4, 9], and audio [2, 7].

While there has been exciting progress in understanding scenes from a limited number of perspectives, it is notable that humans understand the world by incorporating a combination of viewpoints, in a holistic manner. This includes an egocentric view for activities we are involved in and a third-person view for activities we are observing. In addition to visual cues, we also rely on a range of modalities, including hearing and binaural audio delay, to fully comprehend our surroundings and track movements. Our prior knowledge of the scene, such as localisation information and scene descriptions, has also support our understanding of the environment (*e.g.* cafe in the city centre may be different from a similar cafe on a university campus).

Taking the above observations into consideration, a new dataset covering all these aforementioned aspects is presented in this work to provide a panoptic scene understanding, termed *360+x* dataset. This new dataset offers a diverse selection of perspectives, including a 360° panoramic view providing a complete panoptic view of the environment, and a third-person front view that highlights the region of interest that has the most movements in front of the camera. Additionally, we have included egocentric monocular and binocular videos to capture the first-person perspective of individuals in the environment. These viewpoints are complemented by aligned multi-channel audio with directional binaural delay information, as well as location information and scene descriptions as metadata. An illustration of the presented dataset collection system is shown in Figure 1.

Based on this newly collected dataset, we perform 5 visual-audio scene understanding tasks to analyse the contribution and effectiveness of each data viewpoint and modality. Particularly, we look at video classification, temporal action localisation, self-supervised representation learning, cross-modality retrieval and pre-training model migration for dataset adaptation with interesting findings and insights from extensive experimental analysis. The main contributions of this work are summarised as follows:

- We propose to our knowledge the first and probably the most authentic panoptic scene understanding dataset covering multiple viewpoints and data modalities *in the wild*.
- We perform extensive experimental analysis to validate

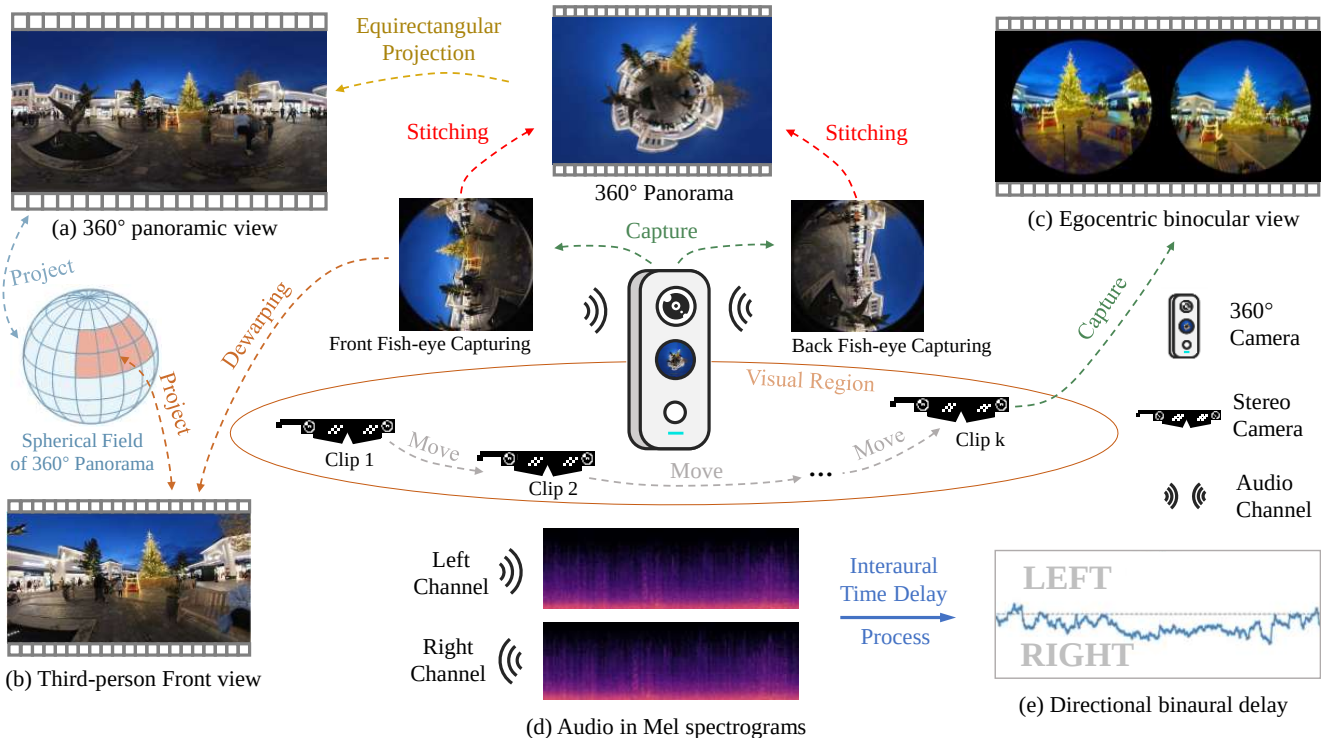


Figure 1. **Illustration of the proposed *360+x* dataset.** The 360° camera records fish-eye raw videos with front and back lenses. These videos are merged to create a spherical 360 panorama (middle-up figure, zoom in for details), which is then transformed to (a) 360 panoramic data using equirectangular projection. The (b) third-person front view is obtained by de-warping the rich movements region highlighted red in the spherical field of 360° Panorama (the middle-left figure). By wearing stereo cameras, the capturers record (c) egocentric clips while staying visible to the fixed 360 camera (central ellipse). (e) Directional audio time delay data is generated from left and right audio inputs (d) from the 360 camera by *interaural time delay* process [3]. This helps locate sound sources in the 360° panorama.

the effectiveness of the proposed dataset on different tasks from various perspectives and modalities.

- Interesting findings are derived from the analysis, suggesting the effectiveness of each viewpoint and data modality. Learning from this new dataset without supervision even shows a better performance than that from a model trained in a supervised manner.

2. Related Works

Video understanding and analysis. Video analysis has been extensively studied in the literature. Existing datasets such as UCF101 [23], ActivityNet [5], and Kinetics [13] have provided large-scale video data for action recognition and activity understanding tasks. However, these datasets often exhibit lower complexity compared to real-world scenes. Some datasets, like MultiThumos [31], aim to increase complexity but are limited to specific scenarios and annotated with domain-specific actions, deviating from real-life daily activities. In contrast, our dataset builds upon the activity labels from ActivityNet [5] and strives to capture data that closely simulates real-life scenarios. Apart from that, we also include multiple data viewpoints and modalities as compared to existing datasets.

Panoramic scene understanding. In recent years, panoramic scene understanding has gained significant attention due to its holistic reflection of the environment. Several datasets have been introduced to facilitate research in this area. For instance, the KITTI-360 [16] provides a collection of panoramic images for scene analysis. EGOK360 [1] has been introduced to address the need for video data with a panoramic view. Im2Pano3D [22] presents a panoramic dataset for indoor scenarios with semantic segmentation and focuses on the prediction from a partial observation. However, these datasets primarily focus on panoramic visual data while lacking the incorporation of other viewpoints (*e.g.* egocentric) and data modalities (*e.g.* audio), limiting their potential for comprehensive scene understanding and analysis.

Egocentric video analysis. Focusing on understanding scenes from a first-person perspective, existing datasets such as EPIC-Kitchens [4] and Ego4D [9] provide egocen-

tric video data collected during daily activities. They have contributed to research on activity recognition and object detection in egocentric scenes. Unlike these datasets focusing on egocentric views, our dataset also covers other viewpoints and modalities aiming at supporting scene understanding research in a more panoptic manner.

Visual-audio analysis. Integrating visual and audio information often enhances the performance of models in scene understanding tasks, as it provides richer contextual information. There are some existing datasets available to support research in audio-visual analysis, *e.g*. AVA [10], AudioSet [6] and VGGSound [2], to name a few. However, these datasets are lacking in multiple viewpoints and the directional property of audio signals, which are provided in the proposed new dataset.

3. 360+*x* Dataset

3.1. Data Acquisition and Alignment

Two main devices were used for data collection: the *Insta 360 One X2* and *Snapchat Spectacles 3* cameras. The first camera has two fish-eye cameras that collect 360° panoramic visual information in the scene with 5760×2880 resolution and a frame rate of 25 FPS. Additionally, directional audio was recorded using four microphones in directional audio mode. While the *Spectacles 3* has a stereo camera attached to glasses used to capture the egocentric binocular vision within the scene at a resolution of 2432×1216 and a frame rate of 60 FPS. The GPS information and weather information were also provided.

Once we obtained the raw data, we aligned the different viewpoints and modalities through a specific process. The initial raw footage captured by the two fish-eye cameras on the 360 camera was in the form of two circular videos, which were then stitched and de-warped into a spherical panorama. This panorama can be projected into an equirectangular format to produce a panoramic video. However, this direct compression of the spherical view into a rectangular format can introduce unnatural distortions.

In order to provide a more natural and informative view, we inversely project a rectangular region into equirectangular space and use it to crop the spherical panorama. We use optical flow to determine the crop region with the most motion activity in the spherical panorama field. This crop region is then projected back to rectangular, resulting in an informative video view with minimal distortions.

Egocentric binocular videos, as shown in Figure 1 (c), were captured ranging from approximately 30 seconds to 1 minute in duration for each clip. A total of 1 to 5 stereo clips were recorded, scattered throughout the duration of the average 6 mins 360° video. In addition to stereo videos, we also provide a higher-resolution version (1580×1580) of monocular egocentric videos.

The audio recordings were temporally aligned with their corresponding videos with left/right channel modality. The four-channel audios with the 360° panoramic video are provided as well for further exploration. Moreover, we also provide the directional information of the audio which was presented using the estimated interaural time delay of the sound obtained from the method introduced in [3].

Given the possibility of occlusions in regions visible to the egocentric camera but not to the panoramic camera, we ensured during data collection that the cameras were positioned in close proximity. This setup, with clear mutual visibility, allowed both cameras to capture a similar overall scene.

3.2. Scene Selection

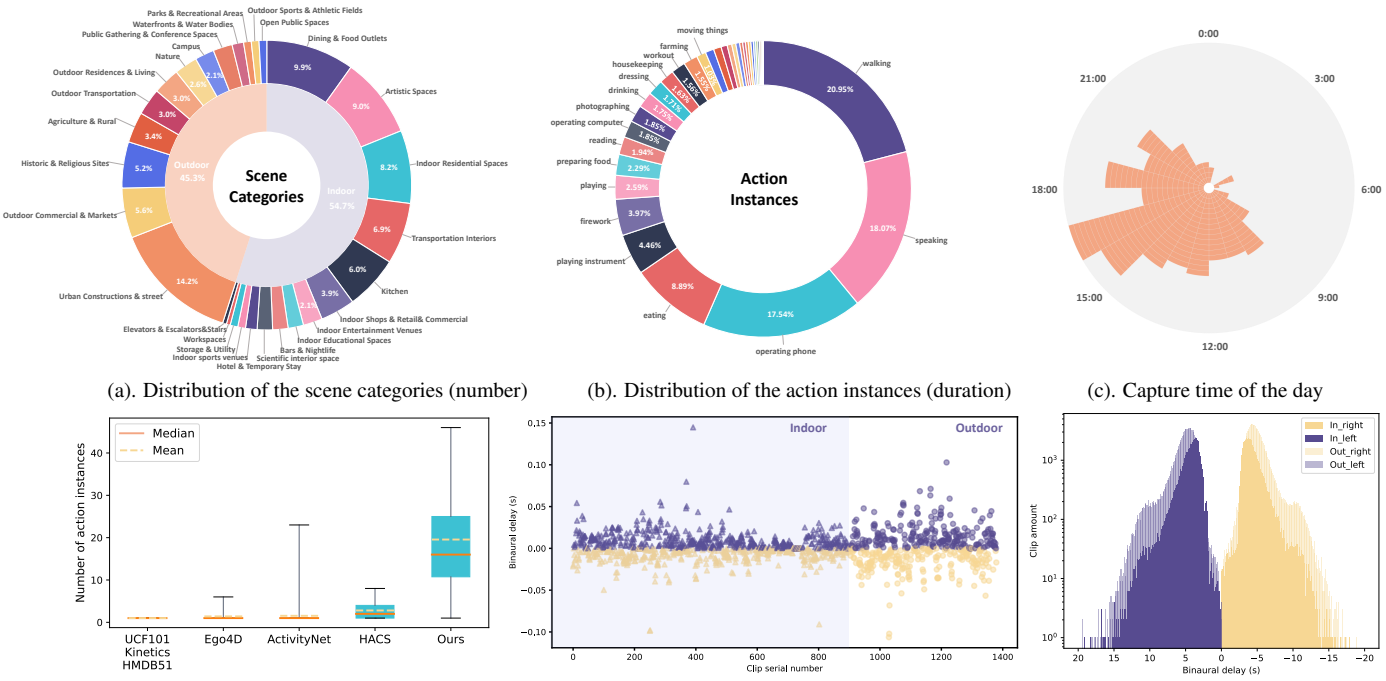
To enhance the scene coverage and foster multi-modal collaborative learning, we implemented a strategic selection process for the captured scenes. This process was guided by three primary criteria:

i) Scene categories must be carefully crafted to be comprehensive, yet concise, while also being authoritative and reflective of everyday life. The location where a scene unfolds plays a crucial role in providing essential environmental context to the activities within it [17]. Distinct scenes can impart unique meanings or emotional nuances to identical events. For instance, the act of chatting could convey divergent implications in a school setting as compared to a home environment. Such nuances are critical as they offer deeper insights into the contextual interpretation of behaviours and interactions in varied settings.

ii) The data should ideally span a wide array of weather and lighting conditions. This criterion aims to ensure the inclusion of both indoor and outdoor activities under various environmental scenarios. Such diversity is important in accurately representing the multifaceted nature of daily life and the various conditions in which these activities occur.

iii) Our third criterion is the inclusion of scenarios rich in activities, particularly those where multiple activities co-occur. It is essential for the dataset to not only visually represent these activities but also to capture the corresponding auditory elements. We try to include diverse and distinctive sound sources within each scene. The goal is to present the complexity and realism of real-world environments as much as possible, marked by simultaneous and various actions and behaviours.

It is worth noting that our dataset was collected across several countries, including the United Kingdom (*e.g*. London, Birmingham, Cardiff and Jersey), France (Paris), Spain (*e.g*. Oviedo and Picos de Europa), China (*e.g*. Guangzhou and Shenzhen), and Japan (*e.g*. Kyoto and Osaka). During the data collection, the *360 Camera* was placed statically to record the scene, while a capturer wearing the *Spectacles* glasses recorded first-person interactions with the scene.



(a). Distribution of the scene categories (number)

(b). Distribution of the action instances (duration)

(c). Capture time of the day

(d). Distribution comparison of the number of action instances per video.

(e). Binaural delay per clip

(f). Overall binaural delay histogram

Figure 2. **Dataset statistics analysis**, on the distributions of (a) the scene category, (b) temporal action instance duration, (c) capturing time, and (d) number of actions per video, (e) binaural delay per clip, (f) the overall clips binaural delay.

Sensitive data handling. Our dataset was collected in a real-world setting and may contain sensitive personal information (*e.g.* human faces). To ensure ethical and responsible research, the video capture was conducted with proper consent. Additionally, we have taken measures to protect privacy by anonymising the data. This includes applying a face detection mechanism to outline predicted face locations in each frame and applying blurring filters to maintain meaningful details while ensuring information security. More detailed information on our privacy protection measures can be found in the supplementary material.

3.3. Data Annotation

Metadata and scene label. The *360+x* dataset provides a diverse range of classification labels for various scenes, encapsulating a total of 28 scene categories – 15 designated for indoor scenes and 13 for outdoor scenes, as illustrated in Figure 2(a). To establish comprehensive and authoritative scene categories that reflect daily life, we referred to the Places Database [34], which is derived from WordNet [18], as our primary basis. We then leverage the sophisticated semantic analysis capabilities of large language models, to conduct a thorough filtering and classification of a multitude of everyday scenes. This curation resulted in a refined set of 28 scene categories, each symbolising aspects of daily life. Simultaneously, the recordings concentrate on capturing common occurrences within conventional settings, providing a realistic depiction of everyday life. Detailed descriptions defining each category, along with discussions regarding these constraints and potential sampling biases, are presented in the supplementary material.

Temporal segmentation label. We also provide temporal segment labelling for the understanding of activities in the shooting scenes. We follow the activity hierarchy standard defined by ActivityNet [5], which provides a comprehensive categorisation of human activities, consisting of seven top-level categories: *Personal Care, Eating and Drinking, Household, Caring and Helping, Working, Socialising and Leisure, and Sports and Exercises*. To capture the diversity and granularity of activities within each category, we defined a total of 38 action instances, covering specific actions and behaviours. To ensure high-quality annotations, the temporal segmentation labelling was annotated by three experienced annotators. Each annotator independently annotated the temporal segments corresponding to the activities in the videos. To obtain a consensus, we merged the individual annotations and resolved any discrepancies according to discussion and consensus among the annotators.

3.4. Dataset Statistics and Analysis

Overview. Existing publicly available datasets primarily focus on visual unimodality [4, 5, 13, 15, 23]. In contrast,

Table 1. **Dataset comparison.** Ego: Egocentric, V: Video, A: Audio, A+V: Audio-visual events.

Dataset	Video Viewpoints				Other Modalities			Statistics			Attributions	
	Third-person Front View	360° Panoramic	Ego Monocular	Ego Binocular	Normal Audio	Directional Binaural Delay	GPS Info	Avg Duration	Total Duration(s)	Frames Count(K)	Annotations Source	Multiple Events
UCF101 [23]	✓	✗	✗	✗	✓	✗	✗	7.21 s	96,000	2,400	V	✗
Kinetics [13]	✓	✗	✗	✗	✗	✗	✗	10 s	2,998,800	74,970	V	✗
HMDB51 [14]	✓	✗	✗	✗	✗	✗	✗	3 s	21,426	643	V	✗
ActivityNet [5]	✓	✗	✗	✗	✗	✗	✗	2 min	2,332,800	11,664	V	✓
EPIC-Kitchens [4]	✗	✗	✓	✗	✓	✗	✗	7.6 min	198,000	11,500	V	✗
Ego4D [9]	✗	✗	✓	✗	✓	✗	✓	8 min	13,212,000	-	A+V	✓
360+x (Ours)	✓	✓	✓	✓	✓	✓	✓	6.2 min	244,000	8,579	A+V	✓

our dataset introduces a novel approach by collecting different views or modalities, as presented in Table 1, including 360° panoramic video, third-person front view video, egocentric monocular video, egocentric binocular video, normal audio, directional binaural delay, location and textual scene description. This diverse range of modalities provides multiple dimensions and clues for understanding and analysing complex scenes. Our dataset comprises 28 scene categories, captured from 15 indoor scenes and 13 outdoor scenes, totalling 2,152 videos, capturing a wide range of environments. Among these, 464 videos were captured using the 360 cameras, while the remaining 1,688 were recorded with the Spectacles cameras. To facilitate analysis and accessibility, these extended videos have been segmented into 1,380 shorter clips, each spanning approximately 10 seconds. In summary, these clips accumulate to a total duration of approximately 244,000 seconds (around 67.78 hours), and the total number of frames is 8,579K.

Figure 2(a) presents the distribution of video counts across each of the 28 scene categories. Our dataset is characterised by a balanced distribution of data across these scenes. Notably, it diverges from conventional databases like UCF101 [23], Kinetics [13], HMDB [15], and ActivityNet [5], particularly in terms of average video duration, which is approximately 6.2 minutes. This longer duration is crucial for maintaining the integrity and coherence of actions within each scene, allowing for a comprehensive temporal analysis of the activities.

Temporal segment label. The annotations of temporal segment labels in our dataset contribute to the fine-grained analysis of activities. We defined 38 action instances representing specific actions and behaviours. The length of each segment labelled with a specific activity varies across the dataset, as depicted in Figure 2(b). Note we acknowledge the significance of audio in accurately identifying certain actions, such as 'coughing' or 'clapping'. Therefore, our dataset combines audio information to enhance accuracy in action recognition [4, 5, 13, 15, 23], as shown in Table 1.

Comparative complexity. Due to its realistic scene simulation, our dataset offers more complexity compared to previous datasets. This complexity arises from the diverse range of activities and interactions captured, resulting in a more challenging and realistic setting for scene understanding and activity recognition. As shown in Figure 2(d), most existing datasets, such as UCF101 [23], Kinetics [13], and HMDB51 [14], typically consist of one action instance per video. While datasets like Ego4D [9] and ActivityNet [5] have large volumes and broad coverage, they often contain a limited number of action instances per individual video. The HACS dataset [33] contains more multiple action instances per video but still pales in comparison to the richness of our dataset. Our dataset surpasses these existing datasets in terms of the number of action instances per video, showcasing the extensive variety of activities captured. The improved complexity and richness of our dataset enable follow-up research to explore and develop more robust algorithms, pushing the boundaries of scene understanding in real-world contexts.

Data distribution. We have ensured a balanced distribution across various dimensions, including scene categories, action instances, binaural delay, *etc*. Figure 2(a) depicts the scene number distribution across 28 scene categories, demonstrating a comprehensive coverage of scene categories. Notably, the dataset achieves an almost equal proportion of indoor and outdoor scenes, accounting for 54.7% and 45.3% respectively. Our dataset allows each scene to conclude multiple diverse action instances naturally, and also enables different scenes to share common action instances. As illustrated in Figure 2(b), the distribution of action duration shows our dataset has captured extensive and realistic human behaviours across natural scenes. One interesting observation from our dataset is the high-frequency occurrence of action 'operating phone', which contributes 17.54% of the whole duration, providing a reflection of mobile usage in common life. Additionally, the dataset offers valuable directional audio to supplement visual understanding. Figure 2(e) illustrates the diversity of binaural delay for each clip. The positive point means the audio is directed towards the left direction while negative the right. Figure 2(f) represents the balanced clip histogram distribution of binaural delay. Notably, data capture time follows natural human activities, as shown in Figure 2(c), human activities

throughout the day are mainly concentrated during the daytime (more in the afternoon and evening). Therefore, the presented *360+x* dataset covers broad modalities and diversity with an authentic distribution from different perspectives, mimicking real daily life.

4. Benchmark and Experiments

To establish a comprehensive benchmark for the presented *360+x* dataset, we choose five visual understanding tasks to delve into the exploration of multiple viewpoints and modalities usage, including video scene classification, temporal action localisation, cross-modality retrieval, self-supervised representation learning and dataset adaptation.

Remark: Unless specifically stated otherwise, the experiments on *360+x* will utilise three views: the 360° view, egocentric binocular view, and the third-person front view.

4.1. Experimental Setting

Models. We employed a consistent set of model backbones across different tasks to minimise model interference, except for *temporal action localisation* task (detailed in section 4.3). We followed the commonly used setup and selected the widely adopted video-specific backbone I3D [13] as our video model. To handle audio-related aspects, we chose the VGGish [12] as our audio model. Additionally, for directional binaural feature extraction, we utilised the ResNet-18 model [11]. A fully connected layer will be placed after the backbones to perform the final task prediction for each specific task based on backbone output features. It is important to note that a simple concatenation of all modalities features can diminish the potential information derived from multi-modality [26]. Therefore, instead of solely concatenating modality features, we leverage a hierarchical attention mechanism for multi-modality integration. In this approach, the directional binaural feature serves as an attention query to direct focused attention towards the audio feature, enabling it to encapsulate the directional information into the audio feature. At the same time, the audio feature is also leveraged by acting as a query itself, enabling it to attentively interact with the video feature. This mechanism allows for creating a synergistic representation of the underlying data that integrates the features of all modalities. For more in-depth information, please refer to the supplementary material.

Training and verification setup. For each temporal action localisation model, we follow their original training settings. For I3D, VGGish, and ResNet-18 networks, the training settings are 200 epochs with the parameters described in [19]. The training process utilises the AdamW optimiser with a learning rate of 10^{-5} and a decay rate of 0.1 at the 80th and 120th epochs. We also apply data augmentation techniques such as rotation, scaling, and colour jittering. The dataset was divided into training, validation, and test sets, following an 80/10/10 split. To ensure a balanced representation of scene categories, the examples were stratified probabilistically across the sets.

Table 2. Video classification performance across different views (Ego: egocentric binocular view, Front: third-person front view, and 360°: 360° view) and data modalities (V: Video, A: Audio, D: Directional binaural delay). Reported in Avg. Prec. (%).

Selected Views	Modalities		
	V	V + A	V + A + D
Egocentric Only	51.95 (±0.0)	55.24 (±0.0)	58.92 (±0.0)
Front Only	54.05 (+2.1)	65.33 (+10.1)	67.19 (+8.3)
360° Only	56.33 (+4.4)	67.14 (+11.9)	70.95 (+12.0)
360° + Egocentric	58.99 (+7.0)	70.48 (+15.2)	72.11 (+13.2)
360° + Front	59.70 (+7.8)	75.06 (+19.8)	77.69 (+18.8)
360° + Front + Ego	**63.73** (+11.8)	**77.32** (+22.1)	**80.62** (+21.7)

4.2. Video Scene Classification

Video scene classification assigns scene labels to videos based on their frames, enabling analysis of visual content and determining the subject matter.

Single view vs. multi-view. First, we are interested in the influence of different combinations of video views on the classification performance. The results, representing each combination, are summarised in Table 2. The results for single views are presented in the first three rows, indicating that using a single 360° panoramic view outperforms using either an egocentric binocular view or a third-person front view only. When employing multiple views, it is noted that better performance can be achieved compared to using a single view. Specifically, utilising all three views leads to the best performance. Such a performance can be attributed to the fact that although these three views describe the same scene, each different view offers a unique perspective that contributes to a more comprehensive understanding of the scene, resulting in improved performance.

Single-modality vs. multi-modality and more. We further investigate the impact of modalities on the model's performance. Various combinations of modalities are analysed, and the results are summarised in Table 2 on a column-wise basis. In particular, the first column represents the visual modality alone, the second column combines video with audio, and the last column incorporates visual, audio, and directional binaural information modalities.

The inclusion of additional modalities leads to average precision improvements. For example, when all three views are utilised, incorporating more modalities results in improvements of *13.59%* and *16.89%*, respectively. This underscores the benefits of leveraging multiple modalities for a more comprehensive understanding of the scene and enhancing overall performance.

Table 3. Temporal action localisation results. Baseline extractors are used in [2, 21, 24, 32]. The mAP@σ represents the mean average precision (%) at a threshold of σ. The best performance is achieved by employing *V+A+D* modalities with extractors pre-trained on *360+x*.

Extractors	Modalities	Actionformer [32] mAP @0.5	Actionformer [32] mAP @0.75	Actionformer [32] mAP @0.95	Actionformer [32] Avg.	TemporalMaxer [24] mAP @0.5	TemporalMaxer [24] mAP @0.75	TemporalMaxer [24] mAP @0.95	TemporalMaxer [24] Avg.	TriDet [21] mAP @0.5	TriDet [21] mAP @0.75	TriDet [21] mAP @0.95	TriDet [21] Avg.
Baseline Extractors	V	11.9 (±0.0)	7.8 (±0.0)	3.3 (±0.0)	7.7 (±0.0)	13.1 (±0.0)	8.8 (±0.0)	3.7 (±0.0)	8.6 (±0.0)	16.7 (±0.0)	10.1 (±0.0)	4.8 (±0.0)	10.5 (±0.0)
	V + A	19.1 (+7.2)	11.3 (+3.5)	4.2 (+0.9)	11.5 (+3.8)	21.0 (+7.9)	14.8 (+6.0)	5.6 (+1.9)	13.8 (+5.2)	23.6 (+6.9)	17.2 (+7.1)	6.4 (+1.6)	15.7 (+5.2)
Pre-trained on *360+x*	V	16.4 (+4.5)	9.8 (+2.0)	3.9 (+0.6)	10.0 (+2.3)	20.4 (+7.3)	14.3 (+5.5)	5.2 (+1.5)	13.3 (+4.7)	21.1 (+4.4)	15.3 (+5.2)	5.5 (+0.7)	14.0 (+3.5)
	V + A	23.6 (+11.7)	16.9 (+9.1)	5.7 (+2.4)	15.4 (+7.7)	25.8 (+12.7)	18.0 (+9.2)	6.4 (+2.7)	16.7 (+8.1)	26.4 (+8.7)	18.5 (+8.4)	6.9 (+2.1)	17.3 (+6.8)
	V + A + D	**24.9** (+13.0)	**17.4** (+9.6)	**6.1** (+2.8)	**16.1** (+8.4)	**26.6** (+13.5)	**18.3** (+9.5)	**6.5** (+2.8)	**17.1** (+8.5)	**27.1** (+10.4)	**18.7** (+8.6)	**7.0** (+2.2)	**17.6** (+7.1)

4.3. Temporal Action Localisation

Temporal Action Localisation (TAL) is a video understanding task that involves the dense identification and temporal segmentation of activities within a video stream over a specific time period. Current TAL approaches typically employ a two-stage paradigm [27, 32]. The first stage extracts features from the entire video, and the second stage predicts temporal segmentation based on these features.

Feature extractors. Baseline extractors are widely utilised for various datasets, *e.g*. ActivityNet [5] and Ego4D [9], on the TAL task. The baseline video features are obtained from an I3D model pre-trained on the Kinetics400 dataset [13]. The baseline audio features are derived from the pre-classification layer following activation of the VGGish model, pre-trained on AudioSet [7]. There is no baseline extractor for *directional binaural delay* feature, so the V+A+D modality was not included accordingly. For a fair comparison, we reused our video classification models in section 4.2 as *Pre-trained on 360+x* extractors, following the same baseline extraction setup for both video and audio features. Additionally, the ResNet-18 feature extractor was used for *directional binaural delay* feature extraction.

Experimental results. We provide a concise overview of the performance comparison for various temporal action localisation methods, including ActionFormer [32], TriDet [21] and TemporalMaxer [24], between the baseline extractors and our *Pre-trained on 360+x* extractors. The summarised results are presented in Table 3, from which we can see that the introduction of additional modalities, *i.e*. audio and direction binaural delay, has a prominent positive impact on the TAL task, leading to performance improvements for both sets of extractors. This result highlights the importance of leveraging multiple modalities in enhancing the accuracy and effectiveness of temporal activity localisation techniques. Using our custom extractors can provide additional improvements, as the baseline extractors may not be optimised for our specific binocular or 360° views.

4.4. Cross-modality Retrieval

In this context, we focus on a series of retrieval tasks that across modalities including audio, video and directional binaural delay. In a modality-specific retrieval scenario, the query modality (Q) serves as the input for retrieving the key modality (K) in the Q-to-K retrieval task. The performance evaluation metric Rθ represents the recall at ranks θ.

Table 4. Q-to-Video retrieval results. The superscript* indicates modalities are co-trained. Recall reported with rank in $\{1, 5, 10\}$.

Query Modality	R1 (%)	R5 (%)	R10 (%)
A	39.14 (±0.0)	62.76 (±0.0)	79.21 (±0.0)
A + D	44.30 (+5.16)	66.92 (+4.16)	84.78 (+5.57)
(A + D)*	**55.88** (+16.74)	**72.53** (+9.77)	**86.6** (+7.39)

Q-to-Video retrieval results. Table 4 illustrates the retrieval results for the Query modality retrieve videos. In this table, *A+D* denotes a set of independently trained audio and directional binaural features employed as query features. Moreover, *(A+D)** signifies the collaborative training of these features instead of treating them independently. The inter-modality retrieval results shown in Table 4 clearly show the modality compliance quality of the *360+x* dataset. Besides Q-to-Video retrieval, we also performed Q-to-Audio and Q-to-Directional binaural delay experiments, details can be found in the supplementary material.

4.5. Self-supervised Representation Learning

Experiment setup. In this section, we investigated the impact of different self-supervised learning (SSL) methods using two engaging video pretext tasks: video pace (VP) prediction [25] and clip order (CO) shuffle prediction [29]. The VP task challenges the model to determine the pace of a video, while the CO task asks the model to rearrange shuffled video clips into their correct chronological order. The original VP and CO primarily concentrated on video data, but to capitalise on the advantages of multi-modality, we expanded these approaches to include audio and directional binaural delay modalities. This extension was done to align modality with the temporal coherence and dynamics observed in the video. For more comprehensive explanations, please refer to the supplementary material.

Experimental results. We first examined the impact of self-supervised learning models for video classification. Ta-

Table 5. Following original setup of THUMOS14 dataset [8], our dataset adaptation task uses video modality only.

Feature Extractor	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP0.7	Avg.
Kinetics400 (*Pre-train*)	83.7 (±0.0)	80.2 (±0.0)	72.8 (±0.0)	62.4 (±0.0)	47.4 (±0.0)	69.5 (±0.0)
360+x (*Pre-train*)	84.5 (+0.8)	81.0 (+0.8)	73.4 (+0.6)	65.9 (+3.5)	54.6 (+7.2)	71.9 (+2.4)
Kinetics400 (*Pre-train*) and 360+x (*Fine-tune*)	**85.3** (+1.6)	**81.8** (+1.6)	**74.9** (+2.1)	**68.1** (+5.7)	**58.2** (+10.8)	**73.7** (+4.2)

Table 6. Models with different pre-train methods were fine-tuned and tested on video classification. The experiments use all three video views. Reported in Avg. Prec. (%).

Pre-train Method	Modalities V	V + A	V + A + D
From Scratch	63.73 (±0.0)	77.32 (±0.0)	80.62 (±0.0)
Video Pace [25]	69.27 (+5.5)	79.56 (+2.2)	81.97 (+1.3)
Clip Order [29]	69.91 (+6.2)	80.14 (+2.8)	82.18 (+1.6)
VP [25] + CO [29]	**76.84** (+13.1)	**82.66** (+5.3)	**83.32** (+2.7)

Table 7. Comparison between supervised pre-trained extractors with SSL pre-trained counterparts on TAL task. The experiments use all three video views with modalities (V+A+D).

Pre-train Method	mAP @0.5	mAP @0.75	mAP @0.95	Avg.
Supervised	27.1 (±0.0)	18.7 (±0.0)	7.0 (±0.0)	17.6 (±0.0)
Video Pace [25]	29.4 (+2.3)	19.6 (+0.9)	7.4 (+0.4)	18.8 (+1.2)
Clip Order [29]	28.9 (+1.8)	19.3 (+0.6)	7.3 (+0.3)	18.5 (+0.9)
VP [25] + CO [29]	**30.3** (+3.2)	**20.2** (+1.5)	**7.9** (+0.9)	**19.5** (+1.9)

ble 6 demonstrates the consistent precision gains achieved by utilising SSL pre-trained models. Notably, leveraging both video pace and clip order SSL techniques resulted in an average performance improvement of $\sim$ *7%*.

We proceeded to perform experiments using SSL pre-trained models as feature extractors for the temporal action localisation task incorporating all three modalities (V+A+D) with the TriDet framework [21]. Since a training-from-scratch model cannot serve as the first-stage extractor, we employed the supervised extractors from section 4.2 as a comparison. The summarised results in Table 7 indicate that pre-training with video pace (VP) or clip order (CO) individually leads to an average performance improvement of $\sim$ *1.2%* and $\sim$ *0.9%* respectively on average, compared to the supervised baseline. The combination of both SSL methods yields the highest performance gain of $\sim$ *1.9%*.

4.6. Pre-training Model for Dataset Adaptation

Here we explore the effectiveness of applying models pre-trained on the proposed *360+x* dataset for adaptation to other datasets, such as the THUMOS14 dataset [8]. In this section, we utilise the TriDet framework [21] to perform TAL task and follow the original modality setup in THUMOS14 to employ only video modality for the comparison.

The performance of this experiment, specifically the mean average precision (mAP) scores covering IoU thresholds from 0.3 to 0.7, are presented in Table 5. As outlined by the results, exclusive reliance on *360+x* video data for training showcases the potential for enhanced performance as compared to training solely based on the Kinetics400 dataset [13]. Remarkably, this performance improvement becomes more prominent at higher IoU thresholds. The utmost optimal performance, however, emerges through a two-step approach, commencing with pre-training on the Kinetics400 dataset followed by fine-tuning on the *360+x* dataset with an average $\sim$ *4.2%* improvement compared to solely Kinetics400 pre-trained extractor. This finding showcases that the employment of the *360+x* dataset for feature extractor training can be beneficial for dataset adaptation in sub-stream tasks. Additional experimental results about the integration with more datasets like the EPIC-Kitchens [4] dataset are provided in the supplementary materials.

5. Conclusions

In this work, we studied the problem of panoptic scene understanding and presented, to our knowledge, the first-of-its-kind dataset – *360+x* to support the study. The proposed *360+x* is a large-scale multi-modal dataset that consists of several different viewpoints (*e.g.* egocentric, third-person-view, and panoramic view) and covers various real-world activities in real daily life. With the most possibly available perspectives describing a real-world scene, *360+x* aims to support the research in understanding the world around us in a way that humans understand (and even beyond). Additionally, we also presented a benchmark study of several scene understanding tasks based on this newly collected dataset, with a comparison to other existing datasets. Extensive experimental analysis validated the effectiveness of each of the perspectives within our dataset, and also suggested interesting insights, confirming that with more viewpoints or data modalities, the understanding of a scene could be more comprehensive. Surprisingly, models trained without manual annotation (*i.e.* self-supervised learning) on our dataset even perform better than those trained with human annotations in a fully supervised manner. We hope this new dataset could bring in new directions towards scene understanding and look forward to the research on them.

References

[1] Keshav Bhandari, Mario A DeLaGarza, Ziliang Zong, Hugo Latapie, and Yan Yan. Egok360: A 360 egocentric kinetic human activity video dataset. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 266–270. IEEE, 2020. 2

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 3, 7

[3] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. In *European Conference on Computer Vision*, pages 489–508. Springer, 2022. 2, 3

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 2, 4, 5, 8

[5] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 1, 2, 4, 5, 7

[6] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 3

[7] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 776–780. IEEE Press, 2017. 1, 7

[8] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes. *http://www.thumos.info*, 2015. 8

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 5, 7

[10] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 4, 5, 6, 7, 8

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 5

[15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1, 4, 5

[16] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[17] Benjamin R Meagher. Ecologizing social psychology: The physical environment as a necessary constituent of social processes. *Personality and social psychology review*, 24(1): 3–23, 2020. 3

[18] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4

[19] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[20] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 1

[21] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 7, 8

[22] Shuran Song, Andy Zeng, Angel X Chang, Manolis Savva, Silvio Savarese, and Thomas Funkhouser. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3847–3856, 2018. 1, 2

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2, 4, 5

[24] Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023. 7

[25] Jiangliu Wang, Jianbo Jiao, and Yunhui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, 2020. 7, 8

[26] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. 6

[27] Xiang Wang, Zhiwu Qing, Ziyuan Huang, Yutong Feng, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Changxin Gao, and Nong Sang. Proposal relation network for temporal action detection. *arXiv preprint arXiv:2106.11812*, 2021. 7

[28] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012. 1

[29] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 7, 8

[30] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 1

[31] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126:375–389, 2018. 2

[32] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 7

[33] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 5

[34] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 4