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

Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, Jianbo Jiao

The Machine Intelligence + x Group, University of Birmingham

CVPR SEATTLE, WA JUNE 17-21, 2024

Paper | Project | Data

https://x360dataset.github.io/

https://cvpr.thecvf.com/virtual/2024/oral/32053

https://cvpr.thecvf.com/virtual/2024/poster/31150

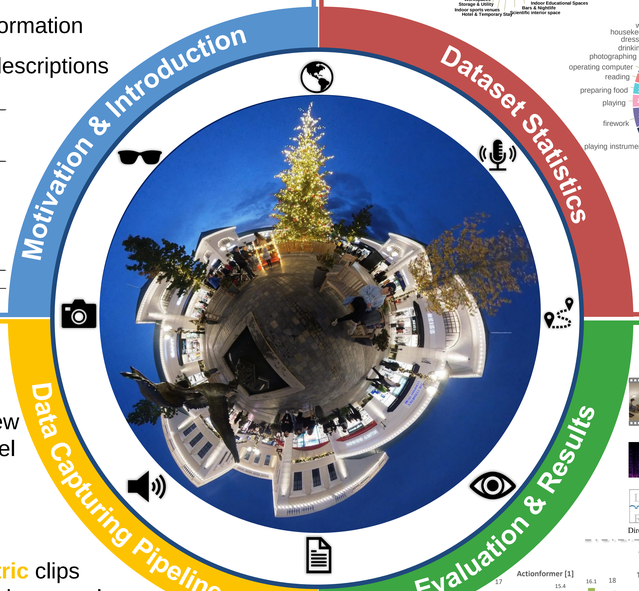
Why? Human perception of the world is shaped by a multitude of viewpoints and modalities.

What? ***360+x*** dataset offers a panoptic perspective, presenting comprehensive observation of the world.

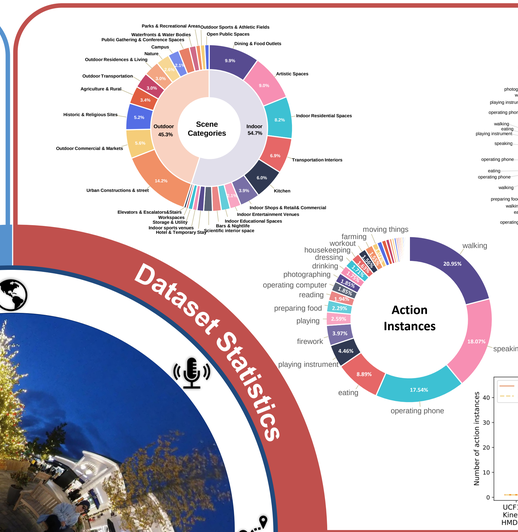
- **Third-person** panoramic views
- Third-person **front** views
- **Egocentric** monocular views
- Egocentric **binocular** views
- Multi-channel **audio**
- **Directional** binaural delay
- **Location** information
- **Text** scene descriptions

Dataset	Video Viewpoints: Third-person Front View	360° Panoramic	Ego Monocular	Ego Binocular	Other Modalities: Normal Audio	Directional Binaural	GPS Info	Statistics: Avg Duration	Total Duration(s)	Frames Count(K)	Attributions: Annotations Source	Multiple Events
UCF101	✓	X	X	X	✓	X	X	7.21 s	96000	2400	V	X
Kinetics	✓	X	X	X	X	X	X	10 s	2998800	74970	V	X
HMDB51	✓	X	X	X	X	X	X	3 s	21426	643	V	X
ActivityNet	✓	X	X	X	X	X	X	2 min	2332800	11664	V	✓
EPIC-Kitchens	X	X	✓	X	✓	X	X	7.6 min	198000	11500	V	X
Ego4D	X	X	✓	X	✓	X	✓	8 min	13212000	-	A+V	✓
360+x (Ours)	✓	✓	✓	✓	✓	✓	✓	6.2 min	244000	8579	A+V	✓

Motivation & Introduction

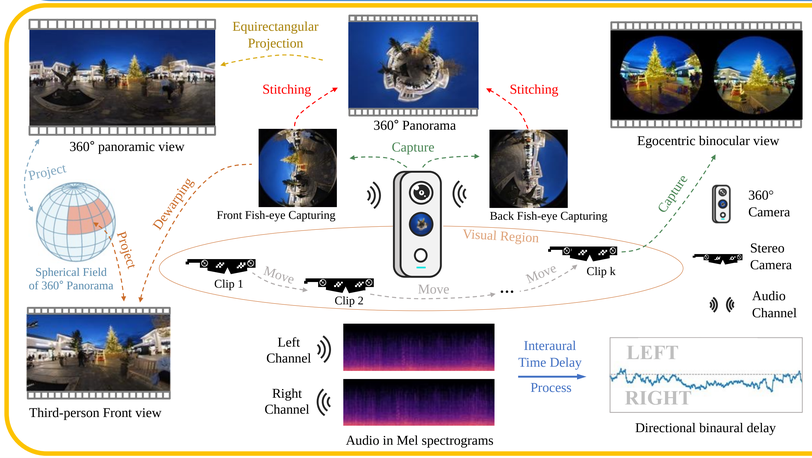


Dataset Statistics



- **28** scene categories
- **38** action instances
- **2,152** videos representing **232** data examples, with **464** videos captured using the 360 camera and the remaining **1,688** recorded with the Spectacles camera.

Data Capturing Pipeline

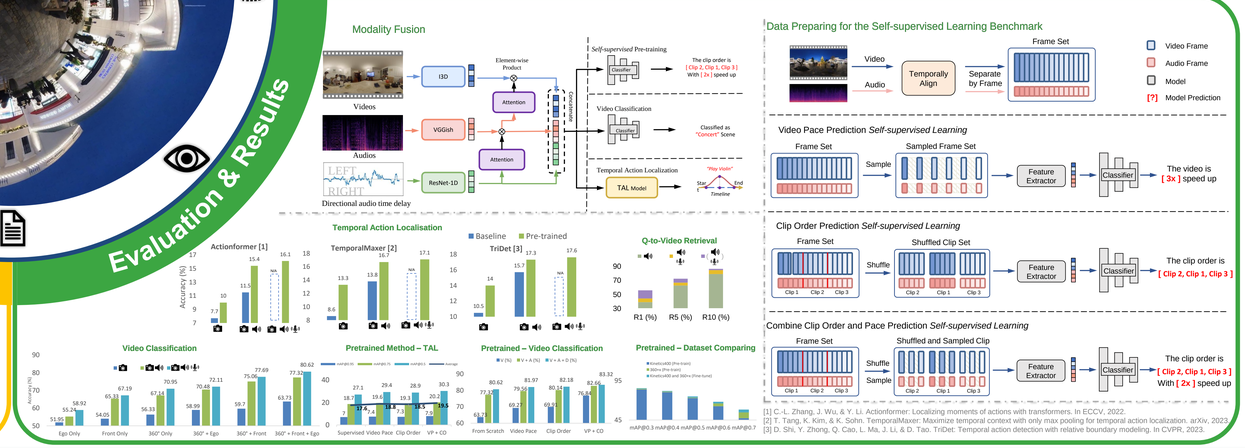


360◦ camera records fisheye **third-person** view with multi-channel **audio**.

Stereo camera records **egocentric** clips into binocular and monocular views with directional binaural **delay**.

Temporal action annotations, **scene** labels, **text** scene descriptions and **location** information will be manually annotated.

Evaluation & Results



[1] C.-L. Zhang, J. Wu, & Y. Li. Actionformer: Localizing moments of actions with transformers. In ECCV, 2022.
[2] T. Tang, K. Kim, & K. Sohn. TemporalMaxer: Maximize temporal context with only max pooling for temporal action localization. arXiv, 2023.
[3] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, & D. Tao. TriDet: Temporal action detection with relative boundary modeling. In CVPR, 2023.

Applications: Multimodal Scene understanding, Video Captioning, 3D Scene Reconstruction, Visual Tracking, AR/VR Generation, Humans: Body, Pose, Gesture ⋯