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

(Supplementary material)

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

This document provides supplementary materials for the main paper. Specifically, Section 2 describes the data organization in detail. Section 4 discusses the ethical usage of the dataset and the author's statement. Section 5 analyzes the limitations of this work and the directions for future work. Section 6 and Section 7 present the future work and the social impact of this work, respectively. Finally, Section 8 provides more implementation details of the benchmarks and the computational resources required.

Website. The dataset/benchmark and model weights are publicly available on the project website at https://x360dataset.github.io and the code repo at https://github.com/x360dataset/x360dataset-dev.

License. The *360+x* dataset is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License. To view a copy, visit here.

Author statement. The authors declare that they are solely responsible for any violation of rights, ethical issues, or legal disputes arising from their work and that they have obtained the necessary permissions and licenses for the data used in their research.

2 *360+x* Dataset Organization

Data alignment. For each video segment, we provide a comprehensive set of data, including 360° panoramic video, ecocentric binocular video, audio files, and the original files labelled by all annotators, accompanied by metadata such as weather conditions, time periods, and GPS information. However, due to the distinct sources of 360° panoramic video and first-person video, and the limitations of long-duration capture with Snapchat Spectacles 3 used for ecocentric binocular video, we captured multiple ecocentric binocular videos within the intervals of the 360° video and performed alignment to ensure synchronization. This integration allows us to present a more holistic view of the captured scenes, combining the immersive perspective of 360° video with the detailed egocentric view from the binocular video.

Pre-extracted features. In addition, we offer pre-extracted features of the data. This includes frame-by-frame images, Mel spectrograms, and directional audio. These pre-extracted features provide a convenient and efficient way for researchers to access and analyze the data without the need for extensive processing. The frame-by-frame images offer a visual representation of the video

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

content, allowing for visual analysis and feature extraction. The Mel spectrograms capture the frequency components of the audio signals, enabling researchers to perform audio-based analysis and modelling. Furthermore, directional audio provides a spatial representation of the audio sources, allowing for the exploration of sound localization and spatial audio processing. Along with these pre-extracted features, we also provide accompanying script files that describe the actions and events occurring in each video segment, facilitating the understanding and annotation of the data. These pre-extracted features and script files serve as valuable resources for researchers to conduct various experiments and develop novel algorithms for tasks such as video analysis, audio processing, and multimodal fusion.

3 Dataset Documentation

The structure of our dataset, as depicted in the diagram, showcases the organization of the data:

```
└─ Inside
   └─ scene name
      └─ scenario name
         ├─ 360
         │  ├─ 360_panorama_cut
         │  │  └─ 360_panorama_cut_{frame}.jpg
         │  ├─ front_view_cut
         │  │  └─ front_view_cut_{frame}.jpg
         │  ├─ vis
         │  ├─ 360_panorama.mp4
         │  └─ front_view.mp4
         ├─ Egocentric
         │  ├─ clip 1
         │  │  ├─ audios
         │  │  ├─ binocular_cut
         │  │  │  └─ binocular_cut_{frame}.jpg
         │  │  ├─ binocular.mp4
         │  │  ├─ binocular.wav
         │  │  ├─ cropped.mp4
         │  │  └─ monocular.mp4
         │  └─ ...
         └─ label.json
```

The dataset is structured into different sections based on the location, starting with the “Inside" category. Each scene within this category is further organized by specific scene names, and within each scene, different scenarios are distinguished by their respective names. For each scenario, two main subdirectories exist: “360" and “Egocentric", representing the two distinct video perspectives.

Under the “360” directory, the “360_panorama_cut” and “front_view_cut” folders contain frame-by-frame images extracted from the 360 panoramic and front-view videos, respectively, with filenames denoted as “360_panorama_cut_frame.jpg” and “front_view_cut_frame.jpg.” The “vis” folder contains additional visualizations or processed images related to the 360° video. The actual videos are stored as “360_panorama.mp4” (original unaltered version) and “360_panorama_distorted.mp4” (distorted version). Additionally, the “front_view.mp4” file represents the front-view video captured alongside the 360 footage.

Within the “Egocentric” directory, individual clips are organized numerically, such as “clip 1”, “clip 2”, and so on. Each clip folder contains an “audios” subdirectory, storing audio-related files, while the “binocular_cut” folder includes frame-by-frame images extracted from the binocular view video, denoted as “binocular_cut_frame.jpg”. The “binocular.mp4” file represents the original binocular video, accompanied by “binocular.wav”, which stores the corresponding audio. The “cropped.mp4” file denotes a cropped version of the binocular video, while the “monocular.mp4” file represents the monocular view video.

Finally, the “label.json” file contains the annotations and labels associated with the videos, providing detailed information about the actions and events occurring within the specific video segments.

This structured organization of the dataset allows for easy navigation and access to the different video perspectives and associated metadata, facilitating research and analysis in various domains.

Accessibility. Large-scale data collection can pose an obstacle for researchers due to hardware limitations such as storage and computing resources. To mitigate this issue, we have implemented several measures["features"?]. These features include directional-audio time delay, audio, and visual features. They are used in the temporal action localization benchmark in the main paper. This provides an easy way to kick off downstream tasks.

Additionally, we provide partitioned data for standardized mini-sets. This allows for a quick overview and experimentation before fully exposing researchers to the whole dataset.

4 Privacy and Ethics

Data collectors have ethical obligations and standards to uphold when conducting data collection efforts. While specifics vary per site, three common obligations and guidelines have been followed:

1. Adherence to the legal terms and the consortium conditions of use.
2. Confidentiality and privacy of participants must be protected.
3. To prevent any breach of confidentiality, sensitive areas should be avoided.

The collecting partner holds consent forms and/or release forms for all videos. The data has been manually de-identified to remove personally identifiable information (PII) and reviewed pre-release.

5 Limitations

Our dataset aims to cover various aspects of daily life to reflect the real world, but we acknowledge that it still possesses certain biases.

First, despite our efforts to collect massive everyday videos from geographically and demographically diverse sources, the current 22 scenes and 10 cities are still far from complete coverage of everyday scenes. Although we have collected footage in countryside and field locations, most of the filming is located in urban or college town areas.

Another challenge is the potential bias and noise in our data collection process. The unscripted nature of the videos introduces a source of variability, as the collectors may select videos based on their personal preferences.

Finally, although the annotations were performed by multiple annotators and merged to minimize bias, there still objectively exist variations in the understanding of scenes. The interpretation of the scene may be biased towards their knowledge background and native word choices, which can affect the language-based narrations and action board in subtle ways.

6 Future Work

The *360+x* dataset is a collaborative project that aims to advance the fundamental AI research for panoramic multi-modal machine perception and scene understanding. We value and encourage global collaborations with researchers and participants from diverse and underrepresented regions, as they are essential for capturing the full spectrum of daily life activities.

Thus, our data collection and annotation methods are designed to be comprehensive and transparent, so that researchers from different regions and backgrounds can join us in expanding the diversity and quality of the dataset. We also plan to extend the current benchmarks to other video-audio scene understanding tasks, e.g. audio-visual diarization, scene querying, pre/post conditions and forecasting, to advance the state-of-the-art techniques in this field.

We also ensure that data collection is supervised with respect to privacy and confidentiality. We comply with legal terms and acquire consent from subjects for filming and using the data only for research purposes.

To ensure the long-term usability and relevance of the dataset, we will provide periodic updates and maintenance for the dataset. This includes checking and fixing any issues with data accessibility and integrity, as well as adding new data or features to keep the dataset up-to-date with the latest research trends and challenges.

7 Social Impact

Our contribution has the potential to positively impact video understanding through multi-modality learning. It provides the research community with a rich-modality view for scene understanding with rigorous privacy and ethics standards. Additionally, it offers a diversity and density of activities and reproducible benchmarks for technical advances.

We acknowledge that large-scale data collection with inadequate oversight could raise privacy and ethical concerns. Therefore, we intend to hinder potential negative applications by making *360+x* data available only for users who sign a license agreement with the statement enumerating the allowable uses of the data.

8 Benchmark Reproduction

We base our code on the official GitHub repositories of OGM-GE[1], video-pace [2], video-clip-order[3] and actionformer[4]. We thank the authors for publicly sharing their code, which facilitates and advances the research community. We also make our code publicly available here: https://github.com/x360dataset/x360dataset-dev.

Our code consists of two parts: “data preparation” and “experiments”. The “data preparation” part involves pre-processing the videos and extracting frames and audio files, generating directional-audio time delays and pre-computing features from pre-trained models. The “experiments” part includes a data loader for the *360+x* dataset, and experiment folders for classification, video self-supervised learning and temporal action localization, respectively.

To enable easy reproduction of our results, we also provide the extracted frames, audio and the pre-computed features on the project website. The model weights are also made public and can be easily downloaded from the code repository or project website.

The experiments were conducted on a single Tesla A100 GPU, but the maximum VRAM usage is only 4G, which is computationally efficient and facilitates reproducibility on other types of GPUs. The training time for the classification task is 3 hours, for the self-supervised tasks is 5 hours, and for the temporal segmentation task is 1 hour.

[1] https://github.com/GeWu-Lab/OGM-GE_CVPR2022

[2] https://github.com/laura-wang/video-pace

[3] https://github.com/xudejing/video-clip-order-prediction

[4] https://github.com/happyharrycn/actionformer_release