

Appendix - Auto-ACD: A Large-scale Dataset for Audio-Language Representation Learning

Anonymous Author(s)

1 DATASET ANALYSIS

In this section, we conduct a more thorough analysis of the proposed dataset, Auto-ACD. In Section 1.1, we meticulously examine the distribution of vocabulary within the dataset. In Section 1.2, we further elaborate on the reasons and methods for data filtering, and display the data that have been excluded. Additionally, in Section 1.3, we compare Auto-ACD with existing audio-language datasets of many aspects. In Section 1.4, we present additional examples from Auto-ACD, which contains diverse information and corpus.

1.1 Dataset Corpus

For better analysis, we visualize the corpus within our dataset. As depicted in Figure. 2, and the common audio tags, *man speak* and *music play* still predominate in frequency within our data. It is noteworthy that terms describing settings, such as *small room* and *music studio*, also emerge with considerable frequency, corroborating the presence of descriptive elements related to soundscapes within Auto-ACD. Moreover, numerous adjectives like *lively* and *passionately* which characterize the attributes of sounds, feature prominently, suggesting that Auto-ACD not only catalogues sound events but also intricately describes the qualities of these auditory phenomena. These are a plethora of audio events, such as *birds chirping*, *engine idling* and *water splashing*, further demonstrating the diverse audio events in Auto-ACD.

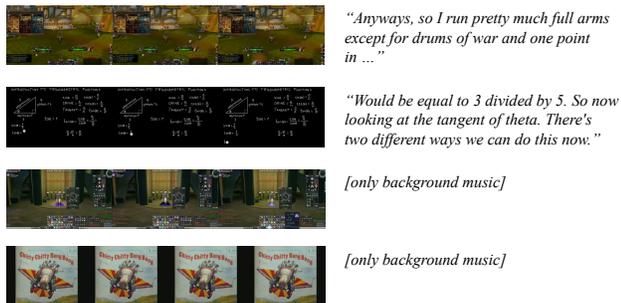


Figure 1: Samples deleted in filter processing. The text on the right side represents transcriptions of speech from the audio in the video, processed using WhisperX.

1.2 Dataset Filtering

Our data relies on strong audio-visual correspondence. However, many entries within AudioSet contain considerable noise, posing challenges to achieving such coherence, for instance, videos synthesized background music alongside serene speeches or videos depicting gameplay or software tutorials. Such videos typically

only encompass two types of audio events: speech and music. Consequently, the generated captions often contain sparse information and exhibit high error rates. Hence, we employ an analysis of audio-visual labels and synchronization to filter these samples. The specific details of this filtering process are described in Section 3.3 of the main text.

In Figure. 1, we present the video frame sequences and the outcomes of audio ASR (Automatic Speech Recognition) by WhisperX [1] for the excluded data. It is evident that the majority of discarded entries are primarily due to the audio and video are not unrelated or not synchronized.

1.3 Dataset Statistics

In total, we collect 1.5 million audio samples, each with a duration of 10 seconds, accompanied by one detailed caption. As indicated in Table 1, in comparison to other datasets, Auto-ACD not only surpasses them significantly in terms of volume, but also boasts a longer average sentence length. It stands as the sole automatically collected dataset that includes contextual information within its descriptions. Laion-Audio-630k may possess a higher vocabulary count, but the majority of its lexicon comprises user-uploaded device information and timestamps, which are irrelevant noise to the audio content.

Dataset	Quantity	Length	# Vocab.	Env.	Auto.
AudioCaps [3]	57K	8.8	5K	×	×
Clotho [2]	30K	11.3	4K	×	×
LAION-Audio-630K [5]	630K	7.3	311K	×	✓
WavCaps [4]	400K	7.8	29K	×	✓
Auto-ACD (ours)	1.5M	18.1	22K	✓	✓

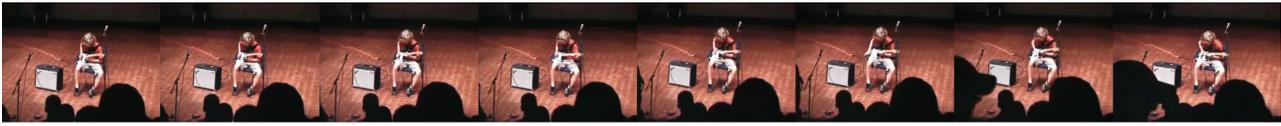
Table 1: Comparison with other audio caption datasets. “Length” and “# Vocab.” refer to average length and vocabulary. “Env.” and “Auto.” refer to environmental information and automatic pipeline, respectively.

1.4 Dataset Visualization

As shown in Table. 2, we show more generated captions for audios from VGGSound and AudioSet. Note that, we present the video sequences to demonstrate how visual information can assist the language description for audio. It can be observed that, the captions in Auto-ACD not only accurately depict sound events but also infer contextual information based on visual priors, that can also be inferred from audios, for example, (i) environmental details, for instance, “a lively performance arena”, “in a music studio” and “a peaceful zen garden”, (ii) sound attributes like “A civil defense siren blares loudly” and “music plays in the background”, (iii) sound variations, for example, “motorcycle engine revs up and down” and “a car speeds down a dirt track”.

Table 2: Data visualization in Auto-ACD. In each sample, the top line showcases the video frame sequence, the bottom line presents the corresponding audio caption. The sound events in the caption are highlighted in bold text, and environmental information is indicated in *italics* text.

No. Generated Caption

1.	
	A man sings while playing the guitar , accompanied by country music and the sound of drums , <i>in a music studio</i> .
2.	
	A civil defense siren blares loudly , indicating an emergency situation, <i>possibly in a city or urban environment</i> .
3.	
	The motorcycle engine revs up and down while driving through a residential neighborhood, accompanied by some speech and light engine sounds .
4.	
	A crowd of people cheer while music plays in the background, <i>creating a lively atmosphere in a concert</i> .
5.	
	The sound of a loud engine revving can be heard as a car speeds down a dirt track at night.
6.	
	The sound of a singing bowl resonates , accompanied by faint tones of a sine wave and a tuning fork <i>in a peaceful zen garden</i> .
7.	
	A group of people cheer and sing while an urban battle cry echoes in the background.
8.	
	Music plays as a crowd cheers and a band performs on stage with vibrant lights <i>in a lively performance arena</i> .

