1 Experiment statistical significance

To assess the statistical significance of the experimental results, we repeated the cross-mode evaluation $n = 5$ times independently with different random seeds. For the accuracy under each experimental condition, we calculated its standard deviation (SD) and standard error of the mean (SEM). The standard deviation was calculated using the formula $SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$, and the standard error of the mean was calculated using the formula $SEM = \frac{SD}{\sqrt{n}}$, where x_i is the accuracy rate of the i-th experiment, $\bar{x}$ is the average accuracy rate of multiple experiments, and n is the number of experiments.

Table 1: The average accuracy (%) and SEM of selected methods on S-LM and S-PM cross-mode evaluation.

Methods	Train on LM		Train on PM	
	Test on LM	Test on PM	Test on PM	Test on LM
AVEL	72.89 ± 1.23	52.12 ± 1.09	73.55 ± 0.75	57.47 ± 0.80
CPSP	74.29 ± 0.21	52.40 ± 0.33	74.98 ± 0.89	63.84 ± 1.09
CMBS	73.50 ± 0.85	51.99 ± 1.12	76.62 ± 0.41	64.68 ± 0.99
LAVISH	85.47 ± 0.23	64.59 ± 0.83	86.39 ± 0.52	74.24 ± 0.42

2 Annotation Tool

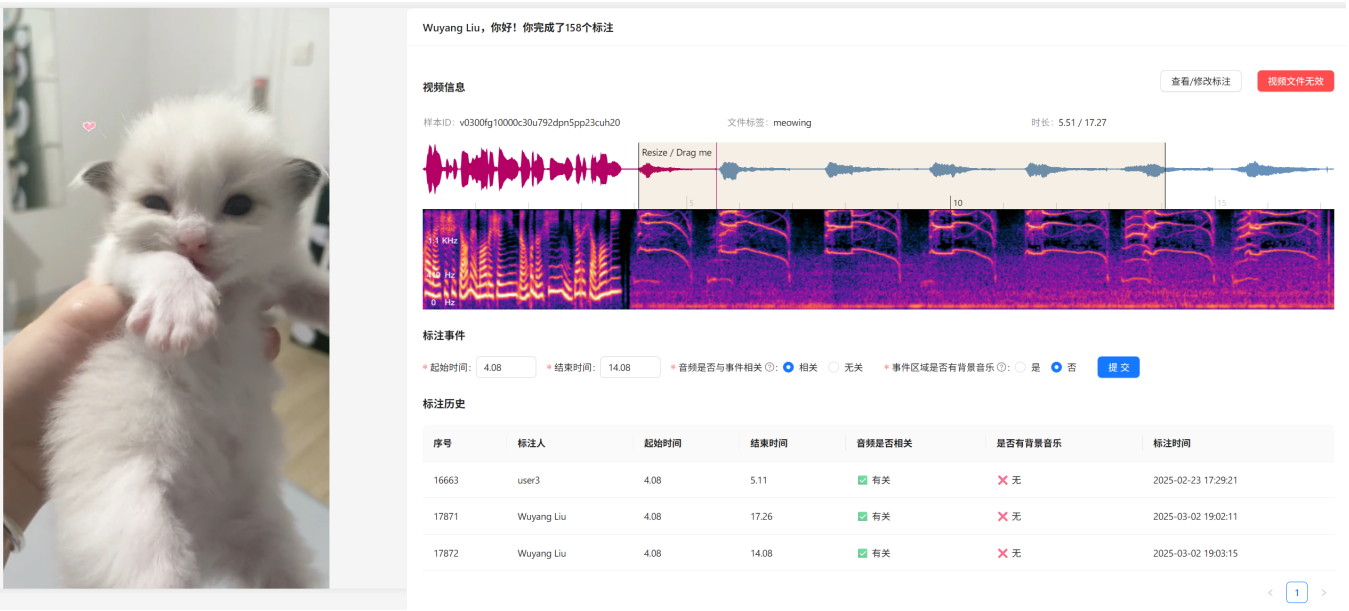

Figure 1: The web-based annotation tool for the AVE-PM dataset. The left panel displays the video player, while the right panel includes (from top to bottom): video metadata (name, duration, and category), an interactive waveform for selecting event boundaries, a form with checkboxes for BGM and audio-event relevance, and a reference table showing annotations from other annotators.

The annotation tool is a web-based interface designed for labeling audio-visual events in the AVE-PM dataset. The left side of the interface features a video player, allowing annotators to review the content. The right side is divided into four sections: (1) Video metadata (e.g., title, duration, category), (2) an interactive waveform of the audio track, where annotators can drag to select precise start and end times of events, (3) a form with fields for event timestamps and two binary checkboxes ("Contains BGM" and "Audio is irrelevant to the event"), and (4) a reference table displaying annotations from other annotators to ensure consistency. The tool's code is open-sourced at https://github.com/dzdydx/ave-annotate.

Annotators were instructed to: (1) Select a single event per sample, ensuring clarity and consistency, (2) mark event boundaries as accurately as possible by dragging on the waveform visualization, aligning with audible and visual cues, (3) determine if the event segment

contains background music (BGM) by checking the "Contains BGM" box when applicable, and (4) check "Audio is irrelevant to the event" if the audio within the selected segment does not correspond to the visual event. These steps aimed to standardize annotations while accounting for auditory complexities like BGM or irrelevant sounds.

3 Visualizing Audio Preprocessing

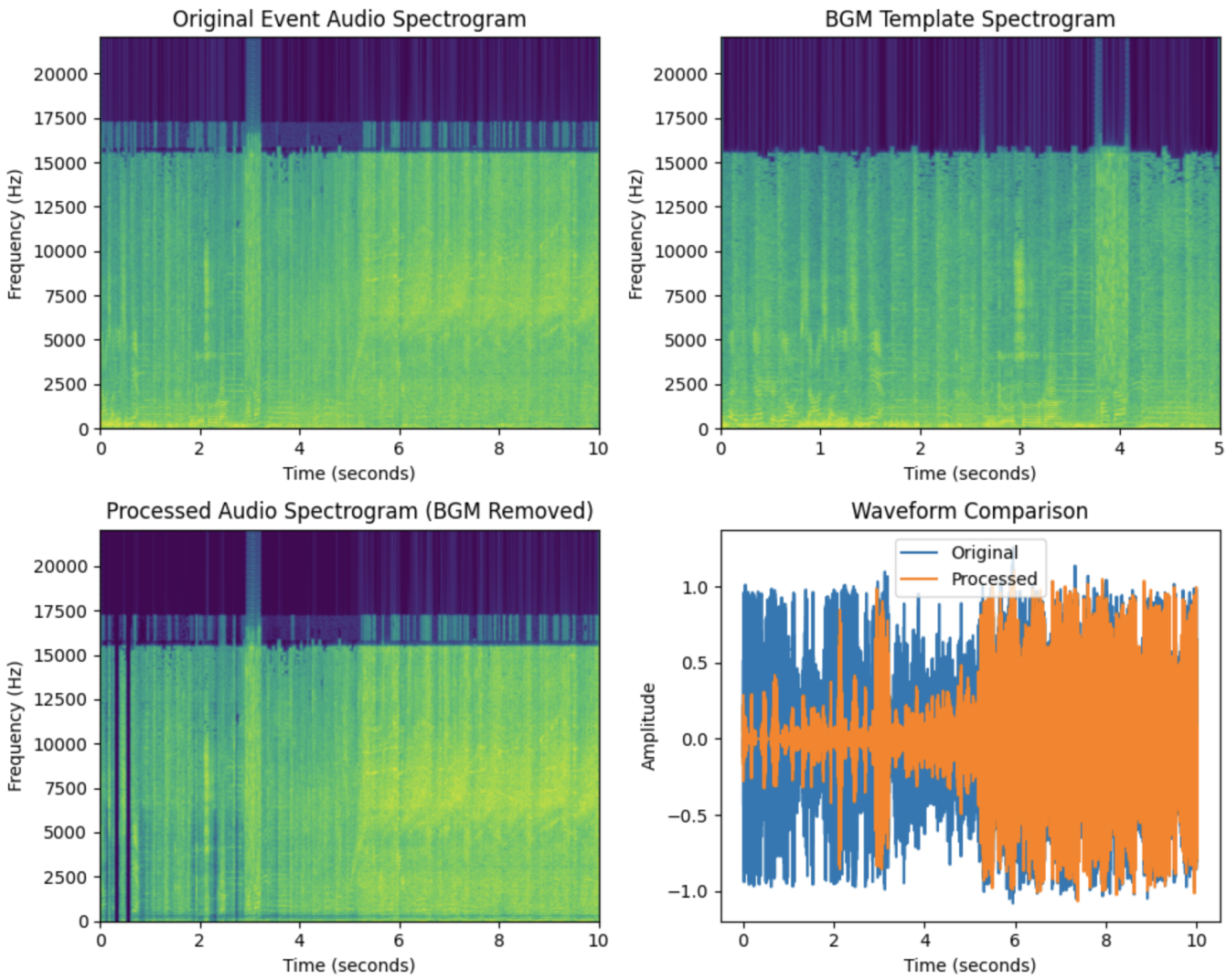

Figure 2: An illustration of the spectrum of the audio signal before and after removing the background music.

We adopt three audio processing methods to remove the background music in short videos: (1) Self NMF. Extract BGM templates from non-event segments within the same video, then remove BGM via NMF decomposition. (2) Template NMF. Utilize clean event templates from BGM-free samples to suppress BGM in contaminated audio via NMF decomposition. (3) Adpative LMS Filtering. Apply Least Mean Squares (LMS) adaptive filters derived from template libraries to attenuate BGM components.

Fig. 2 effectively illustrates the impact of audio preprocessing, where the Self NMF approach successfully isolates and attenuates BGM without compromising the primary event's acoustic signature. The clear contrast between subplots underscores the method's efficacy in enhancing the signal-to-noise ratio for downstream tasks.