AVE-PM: An Audio-visual Event Localization Dataset for Portrait Mode Short Videos

Wuyang Liu
liuwuyang@whu.edu.cn
School of Cyber Science and Engineering, Wuhan University
Wuhan, Hubei, China

Yi Chai
chaiyi@whu.edu.cn
School of Cyber Science and Engineering, Wuhan University
Wuhan, Hubei, China

Yongpeng Yan
yanyongpeng@whu.edu.cn
School of Cyber Science and Engineering, Wuhan University
Wuhan, Hubei, China

Yihuan Huang
yihuanhuang@whu.edu.cn
School of Cyber Science and Engineering, Wuhan University
Wuhan, Hubei, China

Yanzhen Ren
renyz@whu.edu.cn
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education
School of Cyber Science and Engineering, Wuhan University
Wuhan, Hubei, China

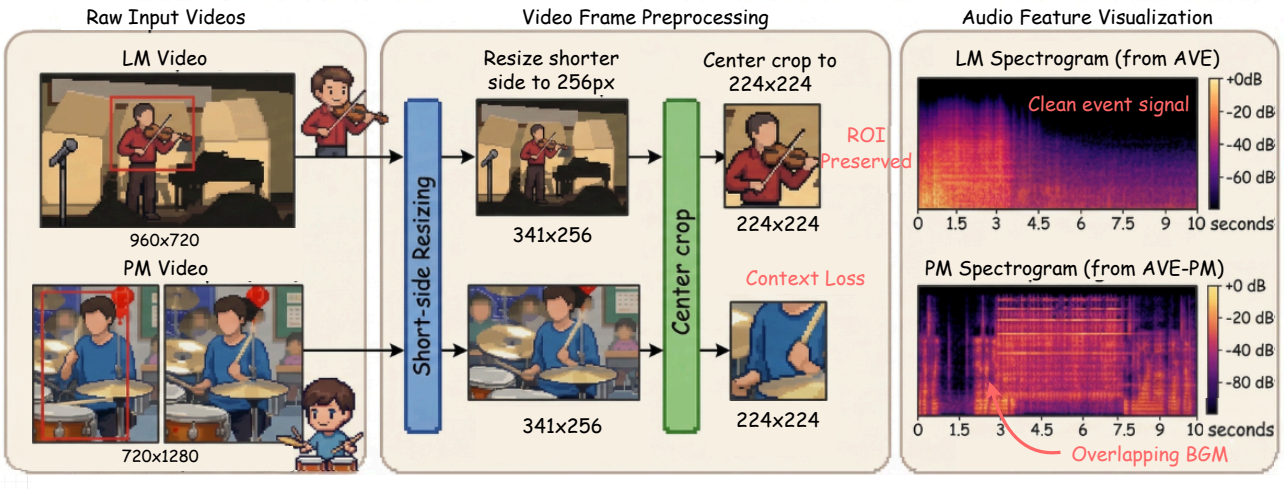


Figure 1: Illustration of the unique visual and auditory challenges in AVEL for Portrait Mode (PM) short videos.

Abstract

While existing datasets for audio-visual event localization (AVEL) predominantly comprise landscape-oriented long videos with simple audio context, short videos have become the primary format of online video content due to the proliferation of smartphones. Short videos are characterized by portrait-oriented framing and layered audio compositions (*e.g.*, overlapping sound effects, voiceovers, and music), which brings unique challenges unaddressed by conventional AVEL datasets. To this end, we introduce **AVE-PM**, the first AVEL dataset specifically designed for portrait mode short videos, comprising 25,335 clips that span 86 fine-grained categories with frame-level annotations and sample-level labels indicating background music (BGM) presence. Our cross-mode evaluation reveals that state-of-the-art AVEL models suffer an average 18.66% performance drop during cross-mode evaluations. Further analysis identifies two critical challenges: (*i*) Spatial bias in PM videos, where standard center cropping degrades performance due to distinct object distribution patterns. (*ii*) Audio complexity, as BGM interference in short videos compromises audio reliability. Our work establishes a foundational benchmark and provides empirically validated strategies for advancing AVEL research in the mobile-centric video era. Dataset and code are available at https://github.com/dzdydx/ave-pm.

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**; Speech recognition.

Keywords

Audio-visual Event Localization, Portrait Mode Video, Multimodal Learning

1 Introduction

As a pivotal task in multimodal scene understanding, audio-visual event localization (AVEL) has gained significant attention due to its wide-ranging applications [28, 30]. Since the publication of the AVE dataset [32], considerable progress has been made in this field [9, 26, 27, 33, 41, 44]. Recent introductions of diverse datasets including LLP [31], XD-Violence [37] and UnAV-100 [11] have further expanded the scope of investigation.

Contemporary AVEL datasets are predominantly constructed using landscape-oriented long videos sourced from platforms like YouTube [11, 31, 32] and movies [37]. However, the proliferation of smartphones and social media has established portrait-oriented short videos as the primary format of online video content [25]. In this paper, we present the **Audio-visual Event in Portrait Mode (AVE-PM)** dataset, the first portrait mode short video dataset dedicated to AVEL research. The dataset contains 25,335 10-second video clips that span over 86 fine-grained categories with human-annotated event onsets and offsets. Sample-level binary labels are provided to indicate the presence/absence of background music in each video. Detailed illustration of AVE-PM is presented in Fig. 2. All the videos are sourced from Douyin [1], ensuring authentic representation of unconstrained user-generated content comparable to the AVE dataset [32].

We conduct cross-mode evaluation to investigate whether existing AVEL models trained on landscape mode datasets can generalize to portrait mode videos, and vice versa. For a rigorous comparison, we selected 10 overlapping categories from AVE dataset [32] and AVE-PM, constructing two subsets: Selected-LM and Selected-PM. We conducted cross-mode evaluations with multiple state-of-the-art AVEL models. An average 18.66% performance drop demonstrates significant degradation in all the selected models, which reveals the domain gap between landscape and portrait mode videos.

More specifically, the change of filming device and user behavior in short video era not only implies a simple change in the aspect ratio, but also brings fundamental changes to content characteristics. As demonstrated in [13], portrait mode videos exhibit stronger subject focus (typically humans) with reduced background context and increased first-person perspective content. Moreover, users tend to create complex audio compositions featuring layered soundtracks (*e.g.*, overlapping sound effects, voiceovers and music). These distinctive characteristics present novel challenges for AVEL systems, as existing methods struggle to generalize to short videos when trained with long videos.

To further explore the fundamental differences between landscape mode long videos and portrait mode short videos, we identified two key aspects from the perspective of AVEL tasks: (*i*) the influence of spatial bias in the video domain, and (*ii*) the complexity of audio content.

We measure the influence of spatial bias in the video domain by visualizing YOLOv5 [15] bounding box distributions, which reveals that portrait mode videos exhibit more concentrated object information, primarily located in the central-lower portion of the frame, whereas landscape mode videos display objects that occupy a smaller proportion of the frame, being tightly focused around the very center, as shown in Fig. 4. Upon this observation, we explore multiple preprocessing recipes, finding that standard center cropping degrades performance on PM content due to critical object information concentrated in the lower-central frame regions.

Regarding audio composition, short videos frequently incorporate noisy background music (BGM) for entertainment purposes, which may interfere with event understanding. To this end, we systematically investigated audio signal processing methods for BGM removal to propose a common practice for audio preprocessing in short videos. Results show that self non-negative matrix factorization (NMF) emerges as the most robust approach, consistently enhancing all three models (AVELN: +2.50%, CPSP: +1.00%, CMBS: +1.62%) by leveraging intra-video non-event segments for adaptive BGM suppression.

In summary, our contributions are: (*i*) We curate AVE-PM, the first audio-visual event dataset designed to accurately localize events in portrait mode short videos. (*ii*) We conduct cross-mode evaluation to reveal that existing AVEL methods struggle to generalize between LM and PM videos (avg. 18.66% performance drop). (*iii*) In the visual branch, we demonstrate that PM videos exhibit distinct object distribution patterns, with standard center cropping degrading performance on PM videos. (*iv*) In the audio branch, we explore audio signal processing strategies to remove the background music in short videos. Results show that self NMF method boosts three AVEL models by up to +2.50% in accuracy.

Table 1: Comparison with related audio-visual datasets. LM: landscape mode. PM: portrait mode. EB: event boundaries.

Dataset	Type	Videos	Classes	Avg. Duration	EB
Audioset [10]	LM	2.1M	527	10s	✗
PM-400 [13]	PM	76K	400	27s	✗
3MASSIV [12]	PM	50K	34	20s	✗
AVE [32]	LM	4,143	28	10s	✓
LLP [31]	LM	11,849	25	10s	✓
XD-Violence [37]	LM	4,754	6	2.74m	✓
UnAV-100 [11]	LM	10,790	100	42.1s	✓
AVE-PM (Ours)	PM	25,335	86	10s	✓

2 Related work

2.1 Audio-visual event datasets

Large-scale audio-visual datasets like Kinetics-Sound [1], AudioSet [10] and VGGSound [3] contribute to advancing audio-visual learning and recognition tasks in machine perception. However, these datasets only contain clip-level annotations with event boundaries. Audio-visual event localization (AVEL) is more intricate because it requires both classification and localization of audio-visual events. AVE dataset [32] is the first AVEL dataset, which is a subset of AudioSet [10] with event temporal boundaries annotated. LLP dataset [31] introduced audio-visual event parsing where video samples contain multiple events. UnAV-100 dataset [11] proposed dense localization of multiple audio-visual events in untrimmed videos. All these datasets are all sourced from landscape-oriented videos, while

[1]Douyin is a popular social media application built for smartphones and primarily features portrait mode short-form videos. https://www.douyin.com/

recent research has focused on developing datasets and methods for audio-visual recognition in diverse video formats, especially short videos in portrait mode. 3MASSIV [12] is a multilingual and multimodal dataset of short social media videos which includes a great proportion of portrait mode videos. However, it focuses on visual concepts rather than specific actions, with only 34 coarse concepts in total. PortraitMode-400 (PM-400) [13], the first dataset consisting of portrait mode short videos for action recognition, has addressed challenges unique to this format. Detailed comparison with related audio-visual datasets is shown in Tab. 1.

2.2 Audio-visual event localization

Recent advances in audio-visual event localization focus on enhancing cross-modal alignment and temporal modeling. Attention mechanisms have been widely adopted, including bidirectional global-local attention [39] and cross-modal co-attention [22, 43]. To address modality interactions, AVSDN [20] applies a sequence-to-sequence cross-modal architecture, while relation-aware networks [41] and semantic modulation frameworks [33] explicitly model audio-visual correlations. Special architectures like MM-Pyramid [46] and MPN [45] leverage multi-scale features, while [38] improve event continuity modeling with span-based approaches. Weakly-supervised methods address label scarcity via contrastive learning strategies [47, 48] and novel loss functions [44]. To mitigate noise interference, [40] adopts background suppression techniques. Optimization methods like OGM-GE [24] alleviate modality imbalance. Recent innovations also explore efficient adaptation of pre-trained vision transformers [21] and latent summarization for temporal inconsistency [8, 14]. However, since all the available datasets are mainly constructed with clips from landscape-oriented long videos, the ability to generalize on portrait mode videos has not yet been discussed.

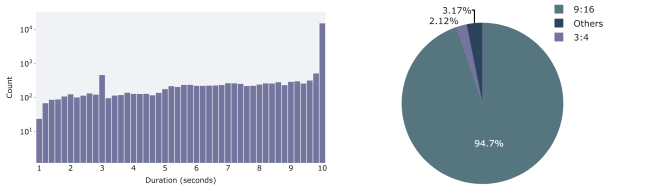
3 The AVE-PM dataset

3.1 Taxonomy

Following the practice of AVE dataset [32], the most commonly used dataset in AVEL, we select specific categories from PortraitMode-400 [13], the first dataset dedicated to portrait mode video recognition. The hierarchical tree structure taxonomy of PortraitMode-400 is then inherited in AVE-PM. Although most of the videos in PortraitMode-400 contain an audio track, not all the categories precisely match the common definition of audio-visual events (e.g., *Makeup* and *performing acupuncture*). Therefore, we built the ontology graph of PortraitMode-400 and compared it with the ontology of AudioSet [10] to obtain 200 candidate categories in PortraitMode-400 that possibly contain audio-visual events. Then, we randomly sampled 20 videos from each category and provided them to expert annotators as a test run, where we filtered out 100 candidate categories for further annotation. Finally, we regrouped these categories into 8 high-level domains that cover most of the occasions in daily life, spanning from human activities to natural sounds.

3.2 Dataset construction

Data collection. According to the video ids provided in [13], we collected raw videos from Douyin platform, a popular social media application built for smartphones and primarily features portrait



(a) Distribution of event duration **(b) Distribution of aspect ratios**

Figure 2: Illustrations of statistics on AVE-PM. (a) Distribution of event duration. (b) Distribution of aspect ratios in AVE-PM, where 94.7% videos are in portrait mode with 9:16 format (width:height).

mode short videos. We performed an audio quality analysis and observed that a large proportion of videos contain background music. While previous datasets such as AVE [32] and UnAV-100 [11] have excluded such videos to ensure clean audio tracks, we argue that this approach may not fully align with real-world scenarios, as background music is prevalent in short videos. Removing these videos alters the data distribution, thus limiting the potential for further applications. Therefore, we choose to provide a `haveBGM` flag for each annotated video so that the quality of the dataset is guaranteed while utilizing these noisy videos remains an option.

Data annotation. We developed a custom video annotation tool for clearer visualization and annotated raw videos by crowdsourcing. Presented with the category of target audio-visual event, annotators are asked to mark the onset and offset of the event on the waveform graph of provided video, as well as confirm the presence of background music within the region. To facilitate accurate temporal boundary identification, which can be challenging based solely on visual cues, annotators are provided with both the waveform and spectrogram of the audio track. To ensure annotation quality, approximately 20% of videos have at least two annotations from two different annotators. A third annotation is required if two annotations differs too much (*i.e.*, a discrepancy of 0.5 seconds or more in either the onset or offset) from each other.

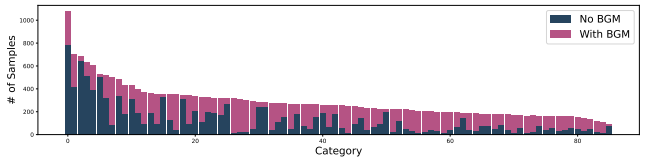


Figure 3: Distribution of the BGM ratios of each category. The bars represent the count of samples with and without BGM for each category. The average BGM ratio of AVE-PM dataset is 59.11%.

Post processing. Since the durations of raw videos vary from 8 seconds to 1 minute, we cut the raw videos into multiple 10-second clips, following the practice of AVE dataset [32]. We then discard the clips in which the event lasts less than 1 second. We also filtered out the categories where valid clips are less than 100, resulting in discarding 14 categories out of 100 annotated categories. Through post

processing, we managed to guarantee that each category contains at least 114 clips.

3.3 Statistical analysis

The AVE-PM dataset comprises 24,450 10-second video clips, each containing a single audio-visual event with temporally annotated onset and offset timestamps. The annotated audio-visual events have an average duration of 8.55 seconds. Additionally, each sample is accompanied by a binary `haveBGM` label indicating the presence of background music. We split the dataset into training, validation and testing sets with a ratio of 7:1:2. The samples from each category are distributed into each subset according to this ratio, thereby guaranteeing consistency in the data distribution across the subsets. The illustrations of statistics are presented in Fig. 2.

In Fig. 3, we present the proportion of samples containing background music for each category. The entire dataset contains 59.11% of samples with background music. The tendency of users to add background music varies significantly depending on the content of the videos. For instance, in the categories of *meowing* and *playing fingerstyle guitar*, the proportion of samples with background music is less than 10% of the total samples in each category. In contrast, in the categories of *freeskating* and *skateboarding*, this proportion exceeds 80%. Although the target events remain audible, the presence of background music still poses a challenge for accurate event localization.

To ensure alignment with real-world data distributions prevalent in short video-based social media platforms, we preserve samples with background music rather than discarding these videos. We maintain that investigating this phenomenon is imperative, particularly given the prevalence of vertical short videos as the dominant online video format. Understanding event semantics in such content holds crucial practical significance, yet existing AVEL benchmarks predominantly utilize landscape mode long videos. This design choice not only enhances ecological validity but also enables systematic exploration of background music interference mitigation - a critical capability for deploying AVEL systems in contemporary digital environments.

4 Cross-mode audio-visual event localization

Landscape mode and portrait mode videos exhibit inherent differences in spatial priors and audio composition, driven by user behavior and device constraints. We hypothesize that models trained on one orientation may struggle to generalize to the other. To validate this, we conduct a comprehensive cross-mode evaluation.

4.1 Experiment setup

To ensure a rigorous comparison between landscape mode and portrait mode audio-visual event localization, we select 10 overlapping categories from the 28 classes of the AVE dataset to construct a subset. We utilize all samples from the corresponding categories of the AVE dataset to build the AVE subset landscape mode (S-LM), which comprises 1,536 samples, accounting for 37% of the total 4,143 samples in the AVE dataset. Subsequently, we select an equal number of samples per category from the AVE-PM dataset to construct the AVE-PM subset portrait mode (S-PM). By ensuring identical taxonomy and equal data distribution per category across both subsets, we establish a fair testing condition to validate the differences in audio-visual event localization between landscape and portrait videos, where the primary distinction between the subsets lies in the data content itself.

We selected four distinct methods for comparison to encompass a diverse range of network architectures. **AVELN** [32] is a dual multimodal residual network designed for the joint modeling of auditory and visual cues. It is also the first model designed for the AVEL task. **CPSP** [47] employs a contrastive positive sample propagation method to enhance feature representation learning. **CMBS** [40] is a cross-modal background suppression network aimed at reducing noise and improving localization performance. These models all adopt separate visual and audio encoders, utilizing pre-trained VGG and VGGish networks to extract video and audio features. In a different direction, **LAVISH** [21] explores the use of a pretrained Swin transformer [23], introducing a latent audio-visual hybrid adapter that achieves competitive performance with fewer tunable parameters.

All the experiments are conducted on a single NVIDIA RTX 4090 GPU. We use Adam [16] optimizer, a multistep scheduler and a batch size of 32. Each training job on full AVE-PM dataset takes about 4 hours, while on S-LM and S-PM it takes about 1 hour. The selected methods use slightly different learning rates and data preprocessing configurations. For consistency, we retain these settings only in the cross-mode evaluation experiments Tab. 2, while keeping preprocessing uniform across all other experiments.

Table 2: Cross-mode evaluation accuracy (%) of selected methods on S-LM and S-PM.

Method	Train	Test	Acc.	Acc. drop	Avg. Acc.
AVELN [32]	LM	LM	**70.42**	-20.55	60.14
		PM	49.87		
	PM	LM	59.65	-11.64	65.47
		PM	**71.29**		
CPSP [47]	LM	LM	**73.83**	-21.32	63.17
		PM	52.51		
	PM	LM	65.53	-7.88	69.47
		PM	**73.41**		
CMBS [40]	LM	LM	**73.36**	-11.01	67.87
		PM	62.36		
	PM	LM	50.34	**-25.41**	63.04
		PM	**75.74**		
LAVISH [21]	LM	LM	**85.85**	**-21.77**	**74.97**
		PM	64.08		
	PM	LM	74.41	-12.64	**80.72**
		PM	**87.04**		

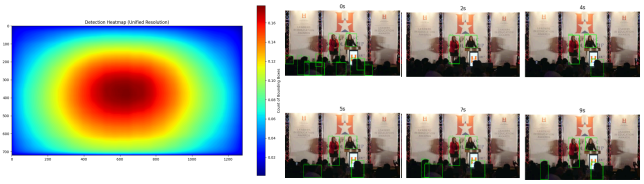
4.2 Cross-mode evaluation results

To demonstrate the domain differences between landscape mode (LM) and portrait mode (PM) videos in the context of audio-visual event localization, we conducted a cross-mode evaluation on the

S-LM and S-PM subsets. We trained the selected models on different subsets and evaluated their performance on the test sets of both subsets, as shown in Tab. 2.

From the experimental results, the first observation we can make is that all models exhibit their best performance when trained and tested on the same subset. This indicates that audio-visual event localization in portrait mode videos is not a trivial problem that can be simply addressed by training existing models on current AVE datasets and directly applying them to portrait mode videos. This suggests that training with portrait mode videos is necessary for existing methods to be applied to diverse scenarios like localizing audio-visual events in short videos.

Another observation is that all models show varying degrees of performance degradation in cross-mode evaluation. Among the models trained on the S-LM subset and tested on the S-PM subset, LAVISH experiences the largest accuracy drop, with a decrease in accuracy of 21.77%. Conversely, among the models trained on the S-PM subset and tested on the S-LM subset, CMBS shows the largest accuracy drop, with a decrease in accuracy of 25.41%. This implies that there are significant domain prior differences between portrait mode and landscape mode videos, and existing methods are not effectively designed to generalize between these two modes. Therefore, further research is needed to address the unique characteristics of portrait mode data.



(a) Object spatial distributions in S-LM subset. (left) aggregated bounding box heatmap across the entire S-LM subset; (right) representative frames with detected bounding boxes from a sample LM video.

(b) Object spatial distributions in S-PM subset. (left) aggregated bounding box heatmap across the entire S-PM subset; (right) representative frames with detected bounding boxes from a sample PM video.

Figure 4: Comparative analysis of object spatial distributions in landscape and portrait videos. (a) Landscape-mode videos show centralized object distribution with minimal spread. (b) Portrait-mode videos exhibit vertically elongated distribution with concentration in the lower-central region. Each subfigure presents both the aggregated heatmap (left) and annotated sample frames (right).

5 Analysis on spatial priors

We aim to further investigate the underlying causes of the observed performance degradation, hypothesizing that distinct spatial priors exist between landscape and portrait videos. To validate this hypothesis, we seek to localize the visual objects associated with the target event in each frame for further analysis.

Specifically, we first resize all video frames to a uniform resolution—1280×720 for the S-LM dataset and 720×1280 for the S-PM dataset. For each frame, we use YOLOv5 [15] to detect the regions containing event-related objects and set the pixels within the bounding boxes to 1, while the remaining areas are set to 0. By aggregating and averaging the pixel-wise values across all frames in both subsets, we generate heatmaps illustrating the spatial distribution of objects in each subset.

The results in Fig. 4 reveal that portrait-mode videos exhibit more concentrated object distributions, primarily clustered in the central-lower region of the frame, as shown in Fig. 4b. In contrast, landscape-mode videos display objects occupying a smaller proportion of the frame, tightly centered around the middle, as shown in Fig. 4a. This disparity in spatial distribution contributes to the observed performance differences, as all selected methods employ center cropping as a preprocessing step for video frames. This approach aligns well with the spatial characteristics of landscape-mode videos, where target objects are predominantly concentrated near the center of the frame—thus, center cropping primarily discards peripheral background information. However, this method proves suboptimal for portrait-mode videos, as center cropping inadvertently removes critical object information located in the lower-central region of the frame. Consequently, the performance degradation in portrait-mode videos can be attributed to the mismatch between the center-cropping strategy and the actual spatial distribution of objects in such videos.

5.1 Video preprocessing recipes for PM videos

The identified spatial distribution mismatch suggests that conventional preprocessing pipelines, particularly those developed for landscape-oriented content, may be fundamentally ill-suited for portrait-mode videos. Most of the audio-visual event localization methods utilize VGG network [29] pretrained on ImageNet [4] for visual feature extraction, like AVELN [32], CPSP [47], CMBS [40] and other methods [8, 14, 26, 34, 39, 42, 44, 46]. Therefore, the center cropping size of these methods is set to 224×224 to match the input size of VGG. Recently proposed vision transformer [5, 23] based methods like LAVISH [21] and other methods [2, 19] utilize patch-embedded visual frames in the shape of 192×192 as direct input instead of using VGG features. Before conducting center cropping on input frames, all aforementioned methods either adopt shorter-side resizing to keep the original aspect ratio or simply resize the visual frame to a square shape of 224×224 or 192×192.

In this subsection, we investigate three video preprocessing strategies for audio-visual event localization: (a) **Shorter-side resizing** [29]. It involves resizing the shorter side of the frame to a fixed length or a random value within a range [35] while scaling the longer side proportionally. (b) **Inception-style augmentation** [6, 7, 18]. It involves random sampling, resizing and cropping, which generates a more diverse group of inputs. (c) **Longer-side**

Table 3: Comparison of accuracy (%) for different preprocessing strategies applied to S-PM videos. "Original" denotes the original preprocessing pipeline from the selected methods. "Inception" refers to Inception-style resizing, which incorporates random sampling, cropping, and resizing. The optimal result of each method is bold while the sub-optimal result is gray.

	Orig.	Shorter.		Longer.	Incep.
		center	random		
AVELN [32]	71.29	74.02	72.89	73.24	**74.98**
CPSP [47]	73.41	74.60	**77.59**	74.93	77.04
CMBS [40]	75.74	75.72	76.91	73.89	**77.62**
LAVISH [21]	87.04	86.30	86.56	**87.23**	86.01

resizing. To specifically address portrait-mode video challenges, we propose this alternative approach where the longer dimension is resized to a fixed length, with the shorter side proportionally scaled and then symmetrically padded with zeros to form a square input.

As shown in Tab. 3, the experimental results reveal distinct preprocessing preferences across methods. VGG-based methods achieve their best performance with Inception-style resizing (74.98% for AVELN and 77.62% for CMBS) or shorter-side resizing with random cropping (77.59% for CPSP), indicating that random operations enhance robustness to aspect ratio distortions. In contrast, the Vision Transformer-based LAVISH performs best with aspect ratio-preserving longer-side resizing (87.23%), outperforming its original square resizing (87.04%) and revealing ViTs' sensitivity to geometric integrity. Notably, shorter-side strategies degrade LAVISH's accuracy by 0.48-0.74%, likely due to disruptive padding patterns. These findings also suggest that both direct center cropping (Original) and shorter-side resizing with center cropping fail to achieve optimal performance across all methods. This observation aligns with our spatial distribution analysis in Fig. 4, where portrait-mode videos exhibit critical object information in the lower-central frame regions—areas systematically excluded by conventional center cropping strategies, suggesting that mainstream preprocessing pipelines may not be directly applicable to portrait-mode videos due to spatial prior incompatibilities.

5.2 Analysis on audio composition

Short video creators frequently incorporate artificial sound effects, voiceovers, and background music (BGM) during production—a practice that often obscures event-related acoustic information. Existing AVEL methods assume clean audio tracks with only target events present, which is an unrealistic presumption and causes severe performance degradation when tested on short videos, as shown in Tab. 2. Meanwhile, existing methods for short video understanding either neglect this critical factor [13] or adopt coarse-grained categorization for BGM-containing videos [12].

To address this challenge, we leverage the binary BGM annotation (haveBGM) of AVE-PM and investigate audio preprocessing strategies through two established signal processing techniques: Non-negative Matrix Factorization [17] (NMF) and adaptive filtering [36]. Our methodology capitalizes on two key observations: (*i*) BGM typically spans the entire video duration, while target events occur sporadically, and (*ii*) 40.89% of samples contain no BGM, providing clean event templates. We then implement three distinct approaches: (*i*) **Self NMF**. Extract BGM templates from non-event segments within the same video, then remove BGM via NMF decomposition. (*ii*) **Template NMF**. Utilize clean event templates from BGM-free samples to suppress BGM in contaminated audio via NMF decomposition. (*iii*) **Template LMS filtering**: Apply Least Mean Squares (LMS) adaptive filters derived from template libraries to attenuate BGM components.

Table 4: Comparison of accuracy (%) of different BGM removal techniques across different models.

Method	Original	Self NMF	Template NMF	LMS Filtering
AVELN[32]	71.29	73.79	**75.43**	72.86
CPSP[47]	73.41	74.41	74.60	**75.61**
CMBS[40]	75.74	**77.36**	75.40	74.50

The results in Tab. 4 demonstrate varying impacts of BGM removal techniques across different models. First, the effectiveness of preprocessing strategies differs significantly depending on the baseline method. For instance, Template NMF achieves the highest accuracy improvement for AVELN (75.43%, +4.14% over no removal) but slightly degrades CMBS performance (75.40%, -0.34%), while Adaptive Filtering benefits CPSP the most (75.61%, +2.20%) yet underperforms for CMBS (74.50%, -1.24%). Second, among the three preprocessing strategies, Self NMF is the only approach that consistently enhances performance across all models, improving AVELN (73.79%, +2.50%), CPSP (74.41%, +1.00%), and CMBS (77.36%, +1.62%). This suggests that leveraging intra-video non-event segments for BGM suppression offers robust adaptability to diverse model architectures.

The findings highlight the critical role of BGM removal in short video understanding. Self NMF's reliance on internal video segments likely mitigates such issues by aligning BGM extraction with the target audio's acoustic properties. For practical deployment, Self NMF emerges as a universally applicable strategy, though computational costs for real-time decomposition warrant further optimization.

6 Conclusion

In this paper, we introduce the Audio-visual Event in Portrait Mode (AVE-PM) dataset, the first dataset dedicated to audio-visual event localization in portrait mode short videos. Through comprehensive experiments, we demonstrated that existing AVEL models struggle to generalize across video modes, revealing a significant domain gap. We also identify the key differences between landscape mode and portrait mode videos, such as spatial bias and audio complexity, highlighting the need for specialized approaches. We make initial attempts to investigate optimal preprocessing techniques for both video and audio modalities. We hope AVE-PM provides a foundation for future research on portrait mode videos.

References

[1] Relja Arandjelovic and Andrew Zisserman. 2017. Look, Listen and Learn. In *Proceedings of the IEEE International Conference on Computer Vision*. 609–617.

[2] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. 2024. TIM: A Time Interval Machine for Audio-Visual Action Recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 18153–18163.

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A Large-Scale Audio-Visual Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 721–725.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009-06). 248–255.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929

[6] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6824–6835.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

[8] Fan Feng, Yue Ming, Nannan Hu, Hui Yu, and Yuanan Liu. 2024. CSS-Net: A Consistent Segment Selection Network for Audio-Visual Event Localization. *IEEE Transactions on Multimedia* 26 (2024), 701–713.

[9] Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. 2023. Learning Event-Specific Localization Preferences for Audio-Visual Event Localization. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 3446–3454.

[10] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780.

[11] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. Dense-Localizing Audio-Visual Events in Untrimmed Videos: A Large-Scale Benchmark and Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22942–22951.

[12] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 2022. 3MASSIV: Multilingual, Multimodal and Multi-Aspect Dataset of Social Media Short Videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21032–21043.

[13] Mingfei Han, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, and Heng Wang. 2024. Video Recognition in Portrait Mode. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 21831–21841.

[14] Yixuan He, Xing Xu, Xin Liu, Weihua Ou, and Huimin Lu. 2021. Multimodal Transformer Networks with Latent Interaction for Audio-Visual Event Localization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[15] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, Zeng Yifu, Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. 2022. *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation*. doi:10.5281/zenodo.7347926

[16] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs]

[17] Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for Non-Negative Matrix Factorization. In *Proceedings of the 14th International Conference on Neural Information Processing Systems (NIPS'00)*. MIT Press, Cambridge, MA, USA, 535–541.

[18] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2023. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 12581–12600.

[19] Yan-Bo Lin and Gedas Bertasius. 2024. Siamese Vision Transformers Are Scalable Audio-Visual Learners. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XIV* (Berlin, Heidelberg, 2024-12-05). Springer-Verlag, 303–321.

[20] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. 2019. Dual-Modality Seq2Seq Network for Audio-Visual Event Localization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[21] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2023. Vision Transformers Are Parameter-Efficient Audio-Visual Learners. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2299–2309.

[22] Yan-Bo Lin and Yu-Chiang Frank Wang. 2020. Audiovisual Transformer with Instance Attention for Audio-Visual Event Localization. In *Computer Vision – ACCV 2020: 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 – December 4, 2020, Revised Selected Papers, Part VI*. Springer-Verlag, 274–290.

[23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 11999–12009.

[24] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8228–8237.

[25] Lang Qin, Ming Zheng, David C Schwebel, Li Li, Peixia Cheng, Zhenzhen Rao, Ruisha Peng, Peishan Ning, and Guoqing Hu. 2023. Content Quality of Web-Based Short-Form Videos for Fire and Burn Prevention in China: Content Analysis. *Journal of Medical Internet Research* 25 (2023), e47343.

[26] Janani Ramaswamy. 2020. What Makes the Sound?: A Dual-Modality Interacting Network for Audio-Visual Event Localization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[27] Varshanth Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. 2022. Dual Perspective Network for Audio-Visual Event Localization. In *Lecture Notes in Computer Science*. Springer Nature Switzerland, 689–704.

[28] Aliaksandra Shutsko. 2020. User-Generated Short Video Content in Social Media. A Case Study of TikTok. In *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing* (Cham, 2020), Gabriele Meiselwitz (Ed.). Springer International Publishing, 108–125.

[29] Karen Simonyan and Andrew Zisserman. 2023. Very Deep Convolutional Networks for Large-Scale Image Recognition.

[30] Li Sun, Haoqi Zhang, Songyang Zhang, and Jiebo Luo. 2020. Content-Based Analysis of the Cultural Differences between TikTok and Douyin. In *2020 IEEE International Conference on Big Data (Big Data)*. 4779–4786.

[31] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*. Springer-Verlag, 436–454.

[32] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-Visual Event Localization in Unconstrained Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.

[33] Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. 2023. Semantic and Relation Modulation for Audio-Visual Event Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 7711–7725.

[34] Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. 2024. Context-Aware Proposal–Boundary Network With Structural Consistency Fo r Audio-visual Event Localization. *IEEE Transactions on Neural Networks and Learning Systems* 35, 11 (2024), 15872–15882.

[35] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. 2021. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1895–1904.

[36] B. Widrow. 1979. A Review of Adaptive Antennas. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. 273–278.

[37] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not Only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Vol. 12375. Springer International Publishing, 322–339.

[38] Yiling Wu, Xinfeng Zhang, Yaowei Wang, and Qingming Huang. 2022. Span-Based Audio-Visual Localization. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, 1252–1260.

[39] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. 2019. Dual Attention Matching for Audio-Visual Event Localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

[40] Yan Xia and Zhou Zhao. 2022. Cross-Modal Background Suppression for Audio-Visual Event Localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR)*. IEEE, 19957–19966.

[41] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. 2020. Cross-Modal Relation-Aware Networks for Audio-Visual Event Localizatio n. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 3893–3901.

[42] Hanyu Xuan, Lei Luo, Zhenyu Zhang, Jian Yang, and Yan Yan. 2021. Discriminative Cross-Modality Attention Network for Temporal Inconsist Ent Audio-Visual Event Localization. *IEEE Transactions on Image Processing* 30 (2021), 7878–7888.

[43] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. 2020. Cross-Modal Attention Network for Temporal Inconsistent Audio-Visual E Vent Localization. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (2020), 279–286.

[44] Cheng Xue, Xionghu Zhong, Minjie Cai, Hao Chen, and Wenwu Wang. 2023. Audio-Visual Event Localization by Learning Spatial and Semantic Co-At Tention. *IEEE Transactions on Multimedia* 25 (2023), 418–429.

[45] Jiashuo Yu, Ying Cheng, and Rui Feng. 2021. MPN: Multimodal Parallel Network for Audio-Visual Event Localization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE.

[46] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. 2022. MM-Pyramid: Multimodal Pyramid Attentional Network for Audio-Visual Event Localization and Video Parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, 6241–6249.

[47] Jinxing Zhou, Dan Guo, and Meng Wang. 2023. Contrastive Positive Sample Propagation Along the Audio-Visual Event Line. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 7239–7257.

[48] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. 2021. Positive Sample Propagation along the Audio-Visual Event Line. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8432–8440.