This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MUG: Pseudo Labeling Augmented Audio-Visual Mamba Network for Audio-Visual Video Parsing

Langyu Wang Bingke Zhu* Yingying Chen Yiyuan Zhang Ming Tang Jinqiao Wang
[1] Foundation Model Research Center, Institute of Automation,
Chinese Academy of Sciences, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
wangly54321@163.com {bingke.zhu, yingying.chen, yiyuan.zhang, tangm, jqwang}@nlpr.ia.ac.cn

Abstract

*The weakly-supervised audio-visual video parsing (AVVP) aims to predict all modality-specific events and locate their temporal boundaries. Despite significant progress, due to the limitations of the weakly-supervised and the deficiencies of the model architecture, existing methods are lacking in simultaneously improving both the segment-level prediction and the event-level prediction. In this work, we propose an audio-visual **M**amba network with pseudo labeling a**UG**mentation (MUG) for emphasising the uniqueness of each segment and excluding the noise interference from the alternate modalities. Specifically, we annotate some of the pseudo-labels based on previous work. Using unimodal pseudo-labels, we perform cross-modal random combinations to generate new data, which can enhance the model's ability to parse various segment-level event combinations. For feature processing and interaction, we employ an audio-visual mamba network. The AV-Mamba enhances the ability to perceive different segments and excludes additional modal noise while sharing similar modal information. Our extensive experiments demonstrate that MUG improves state-of-the-art results on LLP dataset in all metrics (e.g., gains of 2.1% and 1.2% in terms of visual Segment-level and audio Segment-level metrics). Our code is available at https://github.com/WangLY136/MUG.*

1. Introduction

Multimodal learning is now a crucial field in machine learning. Many audio-visual tasks such as audio-visual event localization [33] and audio-visual question answering [38] assume that modalities are aligned and both visual and audio modalities contain learnable cues. However, in the real world, audio-visual events are often unaligned, *e.g.*, seeing a group of people playing basketball but hearing the honking of cars on the road. After observing this prevalent modal mismatch, Tian *et al.* [34] proposed the audio-visual video parsing task. The aim of audio-visual video parsing is to classify video events and localize them according to time and modality. Due to the laborious labeling process, the LLP dataset proposed by Tian *et al.* is only trained using

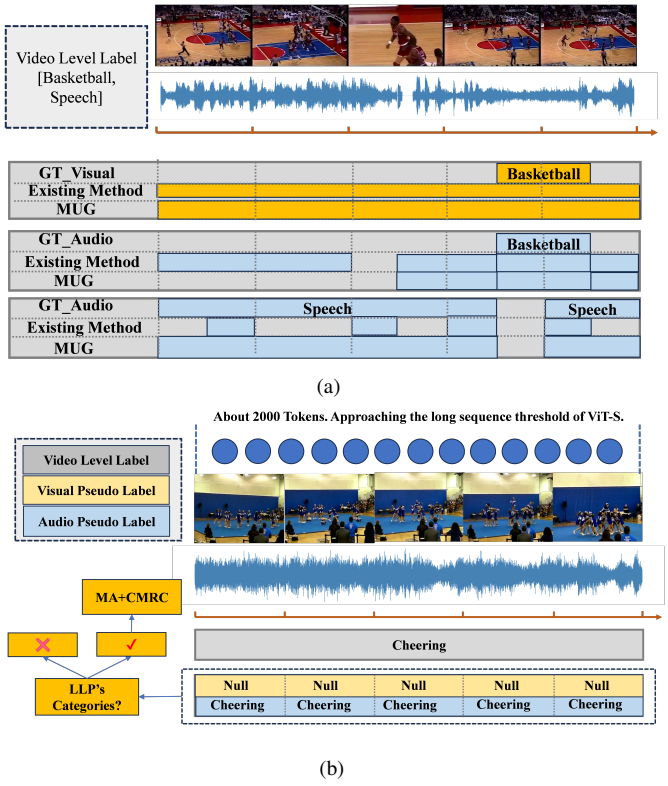


Fig.1. (a): With only video level label, AVVP requires parsing out visual events, audio events, and their temporal boundaries. (b): Constrained by the weakly-supervision learning and the noise in LLP dataset, previous works fail to learn a large number of event combinations at the segment-level, and a large quantity of empty pseudo-labels occur. Meanwhile, the input tokens are approaching the long sequence threshold of ViT-S [45].

*Corresponding author.

weakly-supervised approach, where only video-level labels are provided for the whole training process (Fig.1a).

Limited by weakly-supervised learning, the models can only learn the features of different events from video-level labels. Although previous works [9, 23] have been devoted to extracting pseudo-labels, the quality of the pseudo-labels is restricted due to the presence of noise in the labels of the LLP dataset. Existing frameworks fail to learn a large number of event combinations at the segment-level, which significantly affects the model's prediction of segment-level events. Meanwhile, the events occurring in each modality are independent of each other and may even be completely unrelated. An event in one modality may either complement the prediction of another modality or introduce noise. Some works [8, 29, 35, 44] attempt to develop more robust audio-visual encoders for embedding more effective audiovisual features, but they are insufficient in retaining the original modality features while sharing the features. Finally, the AVVP task requires the simultaneous prediction of events at both the Segment-level (single frame) and the Event-level (multiple frames). Therefore, it is necessary not only to conduct inference on single-frame images/audio but also to capture the strong causal relationships among multiple frames of images/audio. The commonly used Transformer (primarily due to its attention mechanism) models in the past have certain limitations when dealing with long sequences (a large number of tokens). The token length of the visual input sequence in AVVP is close to 2000, which is approaching the threshold of ViT-S [45]. Recently, some Mamba-based backbones [6, 13] have demonstrated great potential in long-sequence modeling and can utilize the causal order to model sequences. However, for single-frame image, there is no sequential dependency. Instead, the ability to model the overall space is required [17, 45]. Therefore, Mamba has deficiencies in the recognition of single-frame image.

In order to effectively enhance the model's perception ability of segment-level features, we propose a brand-new data augmentation strategy applicable to the AVVP task. We extract all the pseudo-labels from previous works [23] and manually annotate the obviously incorrect pseudo-labels (empty labels) among them. Subsequently, we extract the visual track of one video and the audio track of another video, and randomly combine them into a new video. The label of new video is the intersection of the original visual pseudo-label and the auditory pseudo-label. It is worth noting that the pseudo-labels that cannot be annotated and their corresponding videos will be discarded during the random combination. At the same time, we introduce the text modality into the AVVP task. We extract the semantic information of the pseudo-labels and adaptively fuse it with the visual/audio features to eliminate additional modal noise.

Inspired by the Mamba-Transformer architecture in previous works [11, 17, 25], based on the HAN [34] model, we propose a brand-new baseline that can simultaneously improve the model's performance at both the segment-level and the event-level. Specifically, we first utilize a Mamba-based attention to capture the key information in the sequence. Subsequently, we propose a cross-modal adaptive mamba fusion structure which can captures the cross-modal similar information while retaining the intermodal information through a shared matrix. In order to prevent the causal model from forgetting the early token information, we add an additional dynamic branch to alleviate this problem. To enhance the cross-modal similar information captured in the previous steps, we introduce a Mamba feature enhancement module. We incorporate the HAN model at the end of the network, and its simple Transformer (attention architecture) further strengthens the long-range spatial dependencies.

Extensive experimental results on LLP data demonstrate that MUG outperforms existing state-of-the-art models on several metrics. Compared with pure Transformer or CNN architectures, our method achieves more advanced results. Our contributions are summarized as follows:

- We propose a data augmentation approach to effectively improve the model's prediction ability for segment-level, and it can be applied to multiple downstream models;
- We investigate a Mamba-Transformer network, which simultaneously improves the detection accuracy of the model at both the segment-level and the event-level;
- Text features are introduced to exclude the noise of another modality and constrain the prediction of unimodal.

2. Related Works

Audio-Visual Video Parsing (AVVP). The aim of audio-visual video parsing is to identify visual and audio events in a video and locate their timestamps under weakly supervised conditions. Tian *et al.* [34] first introduced the AVVP task and proposed a framework based on hybrid attention networks and multimodal multi-instance learning. Based on this, numerous studies have focused on network architecture construction and the application of attention mechanisms. Yu *et al.* [44] proposed a multimodal pyramidal attention network for capturing and integrating multilevel features. Mo *et al.* [29] proposed a multimodal grouping network to learn dense and differentiated audio-visual encodings. Wu *et al.* [37] designed an algorithm to obtain modality-related labels by exchanging audio and visual tracks. Duan *et al.* [8] used a bi-directionally guided multi-dimensional attentional mechanism to improve performance on a variety of downstream tasks. Gao *et al.* [10] proposed a joint modal mutual learning process that adaptively and dynamically calibrated the evidence for a variety of audible, visible, and audible-visible events. In addition, there are some works that use label denoising strategies or generates pseudo-labels for finer-grained supervised

learning. Cheng *et al.* [3] dynamically identified and removed modality-specific noisy labels in a two-phase approach. Lai *et al.* [23] used frozen CLIP [30] and CLAP [39] to extract features and generated pseudo-labels to aid prediction. Fan *et al.* [9] proposed to perform dynamic re-weighting method to adjust the pseudo-labels. Zhou *et al.* [49] built on pseudo-labels to propose a novel decoding strategy to solve the problem of parsing potentially overlapping events. However, the above methods do not address the limitations inherent in the LLP dataset, which leads to insufficient accuracy of pseudo-labels and insufficient quality of the dataset. Previous works also faced the problem of introducing noise from another modality.

Data Augmentation. Data augmentation techniques are crucial in model training, enhancing generalization and robustness, especially when data is limited. There have been many very mature and widely used data augmentation methods for images, for example [1, 4, 5, 12, 19, 46]. However, the application of data augmentation in the video domain remains relatively sparse. Kim *et al.* [22] extended the data augmentation strategy for images to the temporal dimension of videos as a way to learn temporal features in videos. Yun *et al.* [47] extended CutMix in the field of image recognition to the video domain by proposing VideoMix. Zhang *et al.* [48] observed the effect of hue changes on video understanding and proposed a data enhancement method called motion related enhancement. Building on this foundation, we propose a pseudo-label-based cross-modal random combination method for AVVP task, which can effectively improve the generalization and robustness of the model.

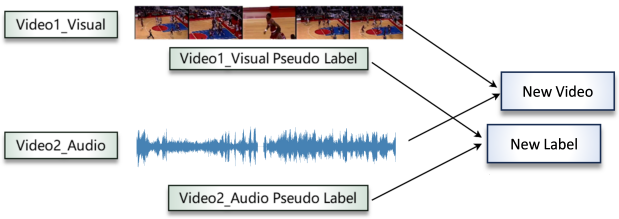


Fig.2. The process of the CMRC.

State Space Models (SSM). The state space model [15] has become one of the most important backbones in deep learning. It originates from classical control theory and provides linear scalability of sequence length for modeling remote dependency. To enhance the practical feasibility, Gu *et al.* [14] proposed the S4 module. Subsequently, Smith *et al.* [32] proposed a SSM model supporting multiple inputs and multiple outputs, and Hasani *et al.* [16] proposed a SSM model for liquid structures. Based on this, Gu *et al.* [13] proposed the mamba architecture. It merges the previous SSM structure with the MLP module in Transformer into a single module, thus obtaining an architectural design with a selective state space. Inspired by ViT [7] and Swin Transformer [27], Zhu *et al.* [50] and Liu *et al.* [26] proposed Vision Mamba and VMamba, respectively, which showed impressive potential on numerous vision tasks. Meanwhile, Mamba structures have begun to emerge on multimodal tasks. xie *et al.* [21] proposed a mamba multimodal fusion structure for medical images. Li *et al.* [24] utilised a coupled state space model to enhance information fusion between different modalities, which greatly improves the inference speed. All these works lay the foundation for subsequent multimodal mamba research.

3. Our Approach

We make improvements from both the data and model perspectives. On the data side, we propose a data augmentation method specifically for the AVVP task. We first annotate some obviously erroneous pseudo-labels. Building on this, we randomly combine the visual and audio modalities of any two videos to generate new data. This method effectively enhance the quality of the dataset, allowing the model to learn the features of each segment more thoroughly. On the model side, we propose an audio-visual mamba network. We capture both temporal and local information simultaneously using Mamba-based attention and share some parameters of the SSM. This approach not only enhances the model's perception of each segment but also shares cross-modal similar information while preserving unimodal information. For effective projection and better model optimization, we incorporate the segment-wise pseudo labels generated in recent work [46] to provide fine-grained supervision. We also introduce the text modality to exclude irrelevant modality noise. The MUG framework is shown in Fig.3. We describe the problem formulation in section 3.1. Then we illustrate the data augmentation and mamba framework in section 3.2 and section 3.3.

3.1. Problem Definition

The AVVP task aims to identify the event of every segment into audio event, visual event and audio-visual event, together with their classes. For the benchmark dataset of Look, Listen, and Parse (LLP), a T-second video is split into T non-overlapping segments, expressed as S = $\{A_t, V_t\}_{t=1}^{T}$, where A and V represent audio and visual segment in time t respectively. In each segment, $y_t^a \in \mathbb{R}^C$, $y_t^v \in \mathbb{R}^C$, $y_t^{av} \in \mathbb{R}^C$ represent to the audio event labels, visual event labels and audio-visual event labels, C is the number of event types. However, we only have weak labels in training split, but have detailed event labels with modalities and temporal boundaries for evaluation.

3.2. Data augmentation in AVVP

Manual annotation (MA). Lai *et al.*[23] used frozen CLIP and CLAP to compute the visual segment pseudo-labels $\hat{y}_v^m$ and audio segment pseudo-labels $\hat{y}_a^m$, respectively. In process of generating pseudo-labels, Lai *et al.* use the pre-

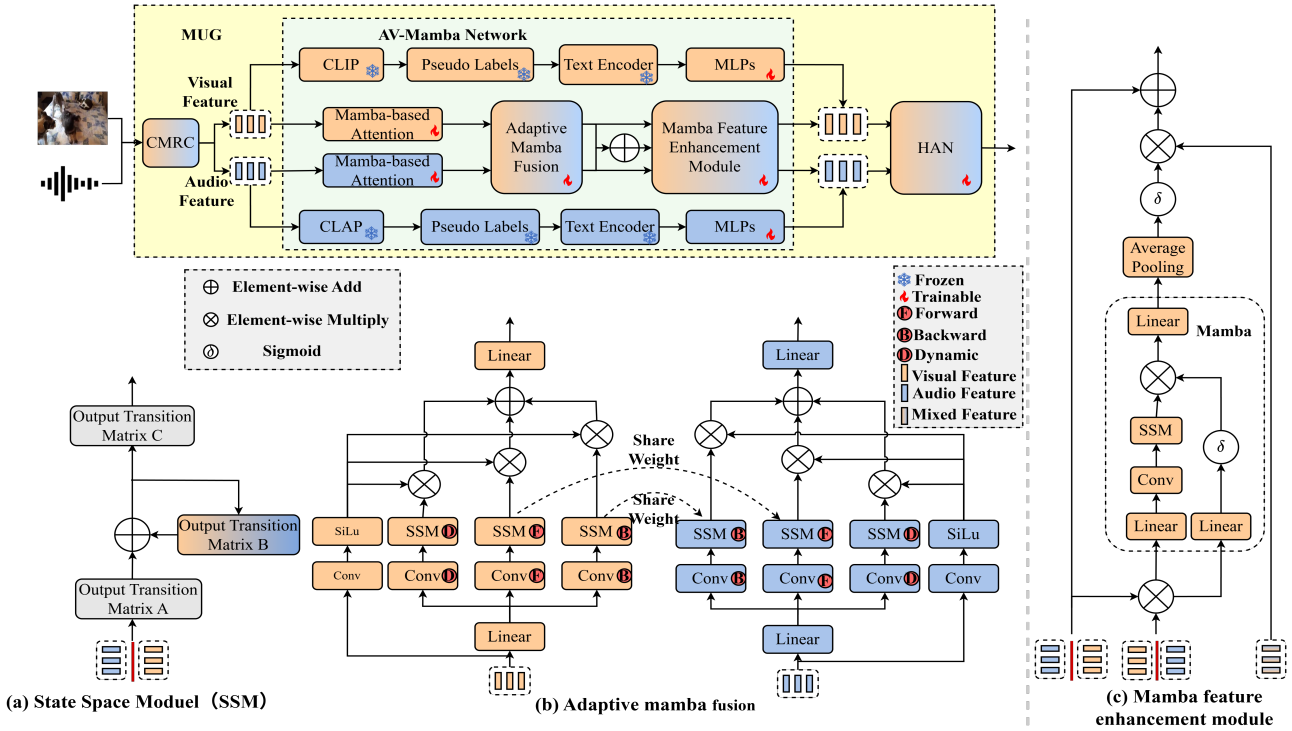


Fig.3. The framework of **MUG**. MUG consists of two parts: the data augmentation CMRC and the AV-Mamba Network. Pseudo-labels are extracted by VALOR [23], which can be used to provide fine-grained supervision and extract semantic features by CLIP/CLAP .

trained CLIP and CLAP to compute the cosine similarity of the features and the 25 categories of labels, and pass a threshold to obtain one-hot encoded pseudo-labels. However, the LLP dataset itself is not annotated sufficient accuracy and the 25 categories cannot encompass all events. The limitation of the LLP dataset is one of the reasons that lead to the insufficient accuracy of pseudo-labels. We convert the generated one-hot pseudo-labels to CSV file, and we can find that there are many null labels (*i.e.*, no events occurred) in the visual modality, which is abnormal since a video may lack sound but rarely lacks image. For such pseudo-labels, we compare them with the LLP dataset and manually annotate them. Pseudo-labels that belong to the 25 categories are accurately annotated, while those outside this range are left unannotated and excluded during the cross modality random combination process.

Cross modality random combination (CMRC). The inability to use data enhancement in previous works is due to the absence of unimodal label information. With the availability of high-quality pseudo-labels, data enhancement for the AVVP task became feasible. For the LLP dataset, we first count the distribution of all video-level labels that appear more than 50 times in the entire training set. Based on this distribution, we selectively combine the visual features of one video with the audio features of another video to generate new video data. The label of the new video is the union of the visual modality pseudo-labels and the audio modality pseudo-labels. Fig.2 represents the process of the CMRC. The pseudo-labels that cannot be annotated and their corresponding videos (*i.e.*, pseudo-labels and videos that do not fall within the 25 categories) are eliminated when generating new data. In order to avoid overfitting and introduce excessive additional data noise, we generate five batches of data according to the actual distribution, with quantities of 1585, 3242, 4610, 6080 and 12096 respectively. We separately test data augmentation for different batches to find the critical point of this method. When the amount of generated data is small, there is still room for model optimization. When the generated data is abundant, the noise in the dataset will increase, and meanwhile, the risk of overfitting rises. CMRC is reasonable because in the real world, visual and audio modalities are not always correlated, because sound signals can originate from various directions.

3.3. AV-Mamba Network

Mamba-based attention (MBA). Audio and visual features are extracted using pre-trained VGGish and ResNet-152, denoted as $\{f_t^a\}_{t=1}^T$, $\{f_t^v\}_{t=1}^T$. F^a and F^v stand for the feature set in the same video, which are defined as $F^a = \{f_1^a, ..., f_T^a\} \in \mathbb{R}^{T\times d}$ and $F^v = \{f_1^v, ..., f_T^v\} \in \mathbb{R}^{T\times d}$, d is the feature dimension. To aggregate features from different segments and enhance the feature expressiveness of each segment, we propose a segment-based attention module implemented via Mamba. This module is similar to the convolutional block attention module [36] but it is implemented by Mamba, thereby better handling the causal relationship.

For each segment, two feature vectors are generated by global maximum pooling and global average pooling. These feature vectors are fed into a shared fully connected layer, resulting in the final attention weight vector W_t^m:

$$W_t^m(f_t^m) = \delta(Mamba(AvgPool(f_t^m)) + Mamba(MaxPool(f_t^m))), \quad (1)$$

where f_t^m denotes audio or visual features within a video, whose dimensions are (B, T, D), representing batchsize, segments, and 128/2048 (audio/visual), respectively. δ indicates sigmoid function, $m \in \{a, v\}$. $Mamba$ denotes Mamba block.

Similar to aforementioned method, two feature vectors are produced through max pooling and average pooling along the temporal (segment) dimension. The features derived from max pooling and average pooling are then concatenated along the temporal dimension to obtain a feature vector S_t^m, $m \in \{a, v\}$:

$$S_t^m(f_t^m) = \delta(Mamba(AvgPool(f_t^m)); (MaxPool(f_t^m))). \quad (2)$$

The outputs W_t^m and S_t^m are element-wise multiplied to obtain the final attention-enhanced feature $\hat{f}_t^m$, $m \in \{a, v\}$:

$$\hat{f}_t^m = S_t^m(W_t^m(f_t^m) \otimes f_t^m) \cdot (W_t^m(f_t^m) \cdot f_t^a). \quad (3)$$

Adaptive mamba fusion (AMF). In previous work [28, 40, 42], there have been explorations in multimodal Mamba fusion with surprising results. Building upon these findings, we propose an adaptive mamba fusion network for selective interaction of visual and audio features. Specifically, after Mamba-based attention, visual features and audio features are fed into AMF. Each of the two input features in the AMF is processed through four distinct branches, which capture unique features. $f_m^{Forward}$ is forward process, performing forward scans. $f_m^{Backward}$ is backward process, performing backward scans. The forward and backward scanning can be expressed as:

$$f_m^{Forward} = SSM([\hat{f}_1^m, \hat{f}_2^m, ..., \hat{f}_N^m], Conv1d), \quad (4)$$

$$f_m^{Backward} = SSM([\hat{f}_N^m, \hat{f}_{N-1}^m, ..., \hat{f}_1^m], Conv1d), \quad (5)$$

where $Conv1d$ represents 1D convolution, SSM represents the state space model, $m \in \{a, v\}$. However, only performing forward and backward scans can potentially ignore the first and last layer tokens. The Mamba has some risk of forgetting the initial token entered, so relying only on forward and backward is not adequate way to model all layers of features. For the input features, the importance of each segment is different and the events that occur are different. We have incorporated the dynamic scanning scheme from recent work [28], allowing the scanning process to start and end at any layer. This approach enhances the model's comprehension of different segments and thus improves the ability to parse segment-level events.

Furthermore, in the forward/backward SSM, two modalities share a portion of parameters due to the consistency in their scanning directions. As shown in Fig.3, forward SSM and backward SSM include 6 parameters: input transition matrices A state transition matrices B,and output transition matrices C. The state transition matrices B has the most significant impact on the system, as they govern the evolution of the current hidden state . Consequently, We share parameters of state transition matrices B between two modalities, while keeping input transition matrices a and output transition matrices C independent [42]. This strategy not only reduces the number of parameters and the potential risk of overfitting, but also preserves unimodal independence while capturing cross-modal similarity information. Subsequently, each feature of both modalities is subjected to a gating strategy in order to fuse the scanning results and obtain the feature f_m^{AMF} by element-by-element summation, $m \in \{a, v\}$. Finally, as shown in Fig.3, the features undergo a simple add to obtain mixed feature f_{mix}^{AMF}. f_m^{AMF} and f_{mix}^{AMF} are fed as three input to the Mamba feature enhancement module. We employ multiple learnable parameters to control the degree of interaction between different modalities, preventing excessive interference between modalities. By modeling sequences in causal order, Mamba strengthens the connections between adjacent frames and enhance the perception ability of multi-frame events.

Mamba feature enhancement module(MFE). The features that have undergone AMF will be fed into the mamba feature enhancement module, as illustrated in Fig.3. This module accepts three types of input: visual features, audio features, and mixed features. Initially, the feature maps from the two modalities undergo average pooling to reduce their dimensionality. Next, the features from the corresponding time steps of the two modalities are subjected to an element-wise multiplication operation. These features are then transformed into channel enhancement vectors through the activation of a Sigmoid function, which performs channel-wise enhancement on the original feature maps at each time step. By leveraging the element-wise product of features from different modalities, the mapping of these similar features is strengthened. This process effectively amplifies the shared features and fine details across the two modalities while suppressing the dissimilar features, thereby mitigating the interference caused by discrepancies between modal features. Finally, we obtain the enhanced features f_m^{MFE}, $m \in \{a, v\}$.

Pseudo-label semantic interaction module (PLSIM). The information in text can be used as a cue to effectively improve the performance of the models [35, 41, 43, 49]. The above work has proven that text can serve as a priori information to culling out the other modal noise and constrain uni-modal event prediction. Contrary to the pseudo-labels employed in [23] , we eliminate logical operations with

video-level labels when generating pseudo-labels for testing purposes. New pseudo-labels can be expressed as $\hat{y}_t^m$.

Since both pseudo-labels $\hat{y}_t^m$ and real labels y are one-hot coded, the event categories corresponding to pseudo-labels f_{event}^a and f_{event}^v can be easily extracted. Next, we convert the event categories in the pseudo-labels into concepts that can be understood by CLIP/CLAP. The title for each event is formulated by prepending the prefix 'A photo of' or 'this is a sound of' to the natural language description of the event. These captions are processed by a frozen CLIP/CLAP text encoder to obtain pseudo-label semantic features F_{CLAP}^a and F_{CLIP}^v for linguistic consistency:

$$F_{CLAP}^a = CLAP(f_{event}^a), F_{CLIP}^v = CLIP(f_{event}^v). \quad (6)$$

In addition, we use multiple MLPs Δ_m^n to map the semantic information of the text, which can be written as:

$$\gamma_{a1} = \Delta_m^1(F_{CLAP}^a), \gamma_{a2} = \Delta_m^2(F_{CLAP}^a), \quad (7)$$

$$\rho_{v1} = \Delta_m^3(F_{CLIP}^v), \rho_{v2} = \Delta_m^4(F_{CLIP}^v), \quad (8)$$

where Δ_m^1, Δ_m^2, Δ_m^3, Δ_m^4 are different MLPs operations to generate the semantic parameters respectively. We use the extracted audio/visual features to fuse with the semantic information, which can be represented as:

$$F_{audio} = f_a^{MFE} \odot \gamma_{a1} + \gamma_{a2} + f_a^{MFE}, \quad (9)$$

$$F_{visual} = f_v^{MFE} \odot \rho_{v1} + \rho_{v2} + f_v^{MFE}, \quad (10)$$

where $\odot$ denotes Hadamard product. γ_{a1} and ρ_{v1} denote scale scaling, γ_{a2} and ρ_{v2} denote bias control. F_a and F_v are audio and visual features fused with semantic features. It is worth noting that when generating new data using CMRC, PLSIM will also to carry out expansion to achieve the fusion of the corresponding features. In the experiments, we find that when using batches augmented with 12,000 samples, the improvement of the PLSIM module decreased significantly and even produced negative effects. This may be due to the fact that excessive data augmentation introduces excessive label noise, thereby affecting the performance of the PLSIM module.

4. Experiments

4.1. Experimental setup

LLP Dataset. The LLP dataset [34] is used to evaluate our method. This dataset has 11849 videos with 25 categories taken from YouTube and consist of a wide variety of scene content including daily activities, music performances, vehicle sounds etc. The dataset has 10000 videos with weak labels as the training set, 1200 videos and 649 videos as the testing set and the validation set with fully annotated labels.

Evaluation Metrics. Following previous works, we use F1-scores on audio, visual and audio-visual events as evaluation metrics. These are computed both at segment and event level. We also include the aggregate metrics "Type@AV" and "Event@AV", again compute at the segment and event level. See [34] for a full explanation of metrics.

Implementation Details. We conduct the training and evaluation processes on a NVIDIA RTX A6000 GPU with 48GB memory. Following the data preprocessing in previous works, we decode a 10-second video at 8 fps into 10 segments. Audio input tokens are extracted through pre-trained VGGish [20], and visual tokens are obtained through the pre-trained models ResNet-152 [18] and R(2+1)D. Our model is trained using Adamw with batch-size of 64 and a learning rate of $3e^{-4}$ for 20 epochs.

4.2. Comparisons with Prior Work

Quantitative. We compare our method with several popular baselines, such as HAN, MM-Pyramid, VALOR, DG-SCT in the same dataset (Table1). From the experimental results, we can see that MUG has improved in all metrics. Compared with the previous SOTA model CoLeaF, it achieves improvements in uni-modal performance, *e.g.*, 2.1% at the Visual Segment-level (66.6% vs. 64.4%), 1.2% at the Audio Segment-level (65.4% vs. 64.2%). Meanwhile, the multi-modal performance is also improved, *e.g.*, 0.6% at the AV Segment-level (59.9% vs. 59.3%) and 1.1% at the AV event level (55.3% vs. 54.2%). Meanwhile, in terms of the Event@AV metric, MUG has an improvement of 2.2% at the Segment-level (64.7% vs. 62.5%) and 2.1% the Event-level (57.7% vs. 55.6%). Compared with our baseline method VALOR, the proposed method can significantly improve the performances on all metrics.

Qualitative. As shown in Fig.4, we qualitatively compare our method with some previous works. Here we use MUG, HAN, JoMoLD [3] for comparison. In Fig.4 above, the first video contains both Speech and Violin and occurs in both modalities. In audio modality, our method localizes the time of Speech that occurs in the last second, while the other two methods fail to do so. In visual modality, although all three methods are accurate in detecting the event Violin, only our method accurately localizes the event Speech. Fig.4 (below) presents another video that contains three events, Singing, Guitar and Clapping. In audio modality, our method accurately locates the Singing and Clapping events, with an error of only one second on Guitar. In visual modality, our method accurately locates Singing and Guitar, with only one second error in Clapping. Overall, our method achieves superior parsing results. This indicates that MUG can perceive the events of each segment more accurately and exclude irrelevant interference.

4.3. Ablation experiments

Cross modality random combination (CMRC). Previous work was constrained by weakly supervised learning, which posed challenges for data augmentation. After pseudo-label

Method	Venue	Segment-level					Event-level				
		A	V	AV	Type@AV	Event@AV	A	V	AV	Type@AV	Event@AV
HAN[34]	ECCV'20	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MM-Pyr[44]	MM'22	60.9	54.4	50.0	55.1	57.6	52.7	51.8	44.4	49.9	50.0
MGN[29]	NeurIPS'22	60.8	55.4	50.4	55.5	57.2	51.1	52.4	44.4	49.3	49.1
JoMoLD[3]	ECCV'22	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
CMPAE[10]	CVPR'23	64.2	66.4	59.2	63.3	62.8	56.6	63.7	51.8	57.4	55.7
DGSCT[8]	NeurIPS'23	59.0	59.4	52.8	57.1	57.0	49.2	56.1	46.1	50.5	49.1
VALOR[23]	NeurIPS'23	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
CM-PIE[2]	ICASSP'24	61.7	55.2	50.1	55.7	56.8	53.7	51.3	43.6	49.5	51.3
LEAP[49]	ECCV'24	62.7	65.6	59.3	62.5	61.8	56.4	63.1	54.1	57.8	55.0
CoLeaF[31]	ECCV'24	64.2	64.4	59.3	62.6	62.5	57.6	63.2	54.2	57.9	55.6
MUG	-	**65.4**	**66.5**	**59.9**	**63.9**	**64.7**	**59.5**	**63.9**	**55.3**	**59.6**	**57.7**

Table1. Comparison with the state-of-the-art methods on the LLP dataset in terms of F-scores.

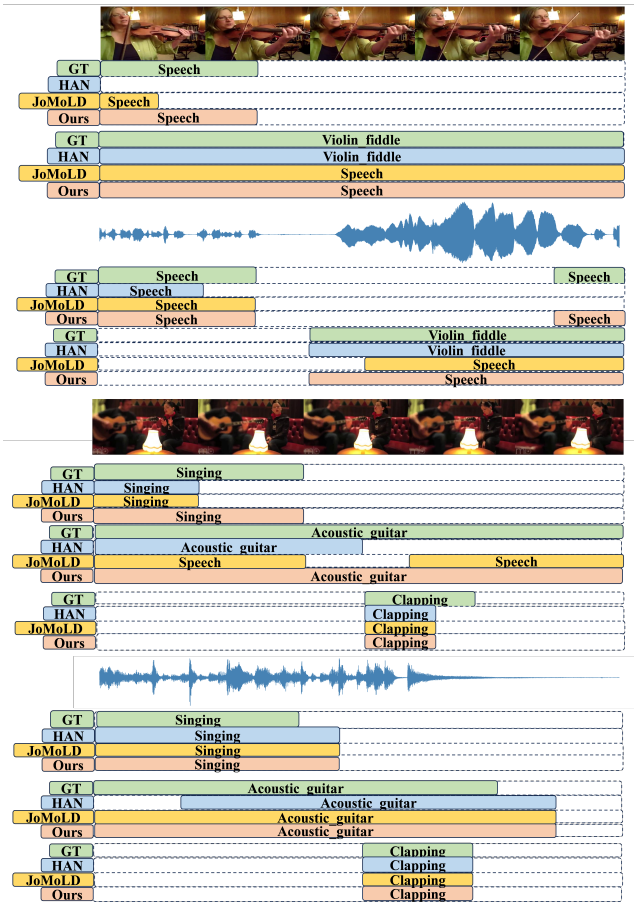


Fig.4. Examples of Qualitative results.

extraction and manual annotation, we are able to leverage the visual features of one video to randomly combine with the audio features of another video to generate completely new datasets and pseudo-labels. Note that in the random combination process, we excluded all null labels and combined them according to the distribution of the dataset. The ablation experiment of CMRC is shown in Table2. To facilitate the presentation of the results, we generate five batches of datasets, the results of which are shown in Table3. The results indicate that training the model with the combined data can effectively enhance accuracy (batch4). There is a significant enhancement in all metrics of visual modality, which is attributed to the fact that obviously wrong pseudo-labels (labels are null) are eliminated during random combination. Due to the limitations of the dataset itself, generating too much data may lead to overfitting, which prevents the model metrics from increasing or even decreasing. Fitting random combination provides a high-quality dataset, enabling the model to thoroughly learn the features of each segment and deepen its ability to perceive each segment. We apply this data augmentation method to a variety of different baselines to demonstrate the effectiveness of CMRC, as shown in Table4.

Mamba-based attention (MBA). After extracting the features, we implement a simple segment-based attention mechanism by Mamba. This mechanism operates on the attention of ten frames or ten audio segments of each video, which effectively captures the salient features in a video and strengthens the feature representation of different segments. From the experimental results (Table2), it can be seen that all metrics show a decline when MBA is not used.

Adaptive mamba fusion (AMF). We introduce additional branch based on Vision Mamba [42, 50] for dynamically adjusting the scanning order. As shown in Table2, we can see that our approach achieves improvement. In the AVVP task, the events occurring in the two modalities are random, potentially being complementary or unrelated. Therefore,the AMF module shares cross-modal information while ensuring the independence of unimodal information. Besides, the addition of dynamic ordering branches mitigates the forgetting problem that mamba models pose in causal learning, which ignores information in early segments.

Mamba feature enhancement module (MFE). We design MFE to capture inter-modal similarities and enrich feature learning. The ablation experiment presented in Table2 demonstrates the effectiveness of our proposed module. The metrics of the model are slightly improved after using MFE. The module amplifies the similar features

Method	Segment-level					Event-level				
	A	V	AV	Type@AV	Event@AV	A	V	AV	Type@AV	Event@AV
MUG	**65.4**	**66.5**	**59.9**	**63.9**	**64.7**	**59.5**	**63.9**	**55.3**	**59.6**	**57.7**
wo/CMRC	62.7	65.2	58.6	62.2	62.2	56.5	61.8	53.8	57.4	54.6
wo/MBA	64.5	66.1	59.4	63.3	63.9	58.3	63.0	53.9	58.4	56.5
wo/AMF	64.1	66.3	59.7	63.4	63.4	58.1	63.0	54.8	58.6	56.1
wo/MFE	63.8	64.6	58.8	62.4	62.8	57.3	61.6	53.5	57.5	55.1
wo/PLSIM	64.8	66.5	59.6	63.7	64.0	58.3	63.3	53.4	58.4	56.3

Table2. Ablation experiments of MUG. wo/ denotes without.

Method	Segment-level				
	A	V	AV	Type@AV	Event@AV
Batch1	63.8	65.4	58.9	62.7	62.9
Batch2	64.4	65.8	59.5	63.2	63.3
Batch3	64.4	66.1	59.2	63.2	63.8
Batch4	65.4	**66.5**	**59.9**	**63.9**	64.7
Batch5	**66.3**	65.8	59.6	63.9	**64.9**
Method	Segment-level				
	A	V	AV	Type@AV	Event@AV
Batch1	57.3	62.5	53.6	57.8	55.5
Batch2	58.3	62.4	53.9	58.2	56.0
Batch3	58.2	62.8	53.7	58.2	56.3
Batch4	59.5	**63.9**	**55.3**	**59.6**	**57.7**
Batch5	**59.8**	62.7	54.1	58.9	57.1

Table3. Ablation results of randomly combining different batches of data. Batch1-Batch5 represent combinations of 0.25, 0.5, 0.75, 1 and 2 times the data as described in 3.2, respectively.

Method	Segment-level				
	A	V	AV	Type@AV	Event@AV
HAN	60.1	52.9	48.9	54.0	55.4
CMRC+HAN	**60.6**	**54.4**	**49.7**	**54.9**	**56.5**
MGN	60.8	55.4	50.4	55.5	57.2
CMRC+MGN	**61.3**	**57.5**	**53.0**	**57.3**	**58.4**
JoMoLD	61.3	63.8	57.2	60.8	59.9
CMRC+JoMoLD	**62.3**	**64.8**	**57.9**	**61.7**	**61.0**
Method	Segment-level				
	A	V	AV	Type@AV	Event@AV
HAN	51.3	48.9	43.0	47.7	48.0
CMRC+HAN	**51.8**	**50.0**	**43.0**	**48.2**	**48.9**
MGN	51.1	52.4	44.4	49.3	49.1
CMRC+MGN	**51.2**	**54.8**	**47.3**	**51.1**	**49.9**
JoMoLD	53.9	59.9	49.6	54.5	52.5
CMRC+JoMoLD	**54.3**	**62.0**	**50.8**	**55.7**	**53.5**

Table4. Results of CMRC on different baselines.

of the two modalities at the same segment, thus allowing the two modalities to achieve a complementary effect in the prediction. We consider that when events occurring in both modalities are similar, one modality can assist in predicting events in the other modality (*e.g.* hearing an engine and seeing a car). In the AVVP task, the similar features of two modalities have a greater probability of representing similar events, while the complementary features have a greater probability of representing different events. When the events occurring in the two modalities are not similar, noise tends to be introduced during modality interaction.

Method	Segment-level					Parameters
	A	V	AV	Type@AV	Event@AV	
CNN	62.4	65.3	58.0	61.9	62.2	**6.5M**
Transformer	62.7	66.2	59.2	62.7	62.2	19.3M
MUG	**65.4**	**66.5**	**59.9**	**63.9**	**64.7**	7.6M
Method	Event-level					Parameters
	A	V	AV	Type@AV	Event@AV	
CNN	56.4	62.7	52.8	57.3	55.2	**6.5M**
Transformer	56.4	62.6	53.0	57.3	54.8	19.3M
MUG	**59.5**	**63.9**	**55.3**	**59.6**	**57.7**	7.6M

Table5. Comparison with Transformer and CNN.

MFE and AMF confirm the necessity to balance contribution in prediction by cross-modal interaction.

Pseudo-label semantic interaction module. We introduce text modality to enhance the model's understanding of the scene. We encode segment-level pseudo-labels as text features that semantically interact with the corresponding visual/audio features. We use the text modality as a constraint to mitigate the noise interference caused by the other modality. As shown in Table2, it can be seen that all metrics of the model are improved after PLSIM. AMF, MFE and PLSIM exclude additional noise while retaining similar information across different modalities.

Comparison with Transformer and CNN. To more comprehensively demonstrate the capabilities of the proposed Mamba-Transformer model, we replace the Mamba component in AV-mamba with either Transformer (multi-head attention) or ResNet (Conv). In order to match AV-mamba, we use a single layer of multi-head attention or a single layer of ResNet in the visual and audio tracks respectively. As shown in Table5, MUG achieves better results.

5. Conclusion

In this paper, we propose a pseudo labeling augmented audio-visual mamba network, which effectively enhances the model's capacity to learn from each segment. Data augmentation not only improves the quality of the dataset, but also allows model to acquire more fine-grained segment information. Additionally, a framework based on Mamba is proposed to enhance the perception ability both on single frame and multiple frames. Our approach proves its performance in a lot of experiments. Future work will focus on evaluating the effectiveness of MUG using larger datasets.

6. Acknowlegement

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0160601), Beijing Municipal Science and Technology Project (Z231100007423004), and National Natural Science Foundation of China (No. 62276260, 62472423, 62206283).

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 3

[2] Yaru Chen, Ruohao Guo, Xubo Liu, Peipei Wu, Guangyao Li, Zhenbo Li, and Wenwu Wang. Cm-pie: Cross-modal perception for interactive-enhanced audio-visual video parsing. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8421–8425. IEEE, 2024. 7

[3] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 431–448. Springer, 2022. 3, 6, 7

[4] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 3

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 3

[6] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 2

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 7

[9] Yingying Fan, Yu Wu, Bo Du, and Yutian Lin. Revisit weakly-supervised audio-visual video parsing from the language perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[10] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18827–18836, 2023. 2, 7

[11] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Yifan Wang, Pingping Zhang, Lijun Wang, and Huchuan Lu. Avs-mamba: Exploring temporal and multi-modal mamba for audio-visual segmentation. *arXiv preprint arXiv:2501.07810*, 2025. 2

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3

[14] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

[15] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021. 3

[16] Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. *arXiv preprint arXiv:2209.12951*, 2022. 3

[17] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[19] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 3

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[21] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. 3

[22] Taeoh Kim, Hyeongmin Lee, MyeongAh Cho, Ho Seong Lee, Dong Heon Cho, and Sangyoun Lee. Learning temporally invariant and localizable features via data augmentation for video recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 386–403. Springer, 2020. 3

[23] Yung-Hsuan Lai, Yen-Chun Chen, and Frank Wang. Modality-independent teachers meet weakly-supervised audio-visual event parser. *Advances in Neural Information Processing systems*, 36:73633–73651, 2023. 2, 3, 4, 5, 7

[24] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multi-modal fusion with coupled state space model. *arXiv preprint arXiv:2405.18014*, 2024. 3

[25] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024. 2

[26] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[28] Andong Lu, Wanyu Wang, Chenglong Li, Jin Tang, and Bin Luo. Rgbt tracking via all-layer multimodal interactions with progressive fusion mamba. *arXiv preprint arXiv:2408.08827*, 2024. 5

[29] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 35: 34722–34733, 2022. 2, 7

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[31] Faegheh Sardari, Armin Mustafa, Philip JB Jackson, and Adrian Hilton. Coleaf: A contrastive-collaborative learning framework for weakly supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 1–17. Springer, 2025. 7

[32] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 3

[33] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 1

[34] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 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 16*, pages 436–454. Springer, 2020. 1, 2, 6, 7

[35] Langyu Wang, Bingke Zhu, Yingying Chen, and Jinqiao Wang. Link: Adaptive modality interaction for audio-visual video parsing. pages 1–5, 2025. 2, 5

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[37] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 2

[38] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019. 1

[39] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[40] Xinyu Xie, Yawen Cui, Chio-In Ieong, Tao Tan, Xiaozhi Zhang, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *arXiv preprint arXiv:2404.09498*, 2024. 5

[41] Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24902–24912, 2024. 5

[42] Jiaxin Ye, Junping Zhang, and Hongming Shan. Depmamba: Progressive fusion mamba for multimodal depression detection. *arXiv preprint arXiv:2409.15936*, 2024. 5, 7

[43] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024. 5

[44] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6241–6249, 2022. 2, 7

[45] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? *arXiv preprint arXiv:2405.07992*, 2024. 1, 2

[46] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[47] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020. 3

[48] Yitian Zhang, Yue Bai, Huan Wang, Yizhou Wang, and Yun Fu. Don't judge by the look: A motion coherent augmentation for video recognition. *arXiv preprint arXiv:2403.09506*, 2024. 3

[49] Jinxing Zhou, Dan Guo, Yuxin Mao, Yiran Zhong, Xiaojun Chang, and Meng Wang. Label-anticipated event disentanglement for audio-visual video parsing. *arXiv preprint arXiv:2407.08126*, 2024. 3, 5, 7

[50] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3, 7