PreFM: Online Audio-Visual Event Parsing via Predictive Future Modeling

Xiao Yu[1,2] **Yan Fang**[1,2] **Yao Zhao**[1,2] **Yunchao Wei**[1,2,✉]
[1]Institute of Information Science, Beijing Jiaotong University
[2]Visual Intelligence + X International Joint Laboratory
[✉]Corresponding Author
xiaoyu@bjtu.edu.cn wychao1987@gmail.com

Abstract

Audio-visual event parsing plays a crucial role in understanding multimodal video content, but existing methods typically rely on offline processing of entire videos with huge model sizes, limiting their real-time applicability. We introduce Online Audio-Visual Event Parsing (On-AVEP), a novel paradigm for parsing audio, visual, and audio-visual events by sequentially analyzing incoming video streams. The On-AVEP task necessitates models with two key capabilities: (1) Accurate online inference, to effectively distinguish events with unclear and limited context in online settings, and (2) Real-time efficiency, to balance high performance with computational constraints. To cultivate these, we propose the **Pre**dictive **F**uture **M**odeling (PreFM) framework featured by (a) predictive multimodal future modeling to infer and integrate beneficial future audio-visual cues, thereby enhancing contextual understanding and (b) modality-agnostic robust representation along with focal temporal prioritization to improve precision and generalization. Extensive experiments on the UnAV-100 and LLP datasets show PreFM significantly outperforms state-of-the-art methods by a large margin with significantly fewer parameters, offering an insightful approach for real-time multimodal video understanding. Code is available at https://github.com/XiaoYu-1123/PreFM.

1 Introduction

Multimodal learning [5, 72, 38, 79] is a significant topic in the machine learning research area. Among various modalities, audio [62] and vision [60, 47] are the primary ways humans perceive the world, making audio-visual learning (AVL) [19, 42, 35, 15] essential. Among various progress [43, 45, 31, 34] related to AVL, audio-visual event parsing (AVEP), i.e., understanding events in videos, becomes increasingly important with the explosive growth of video content on streaming platforms.

AVEP involves processing both modality-aligned (audio-visual) and modality-misaligned (audio-only or visual-only) events in video content. Prevailing methods [13, 14, 78] operate offline, analyzing entire video sequences to utilize global context for accurate video events understanding. Though offering precise predictions, the necessity of whole-video processing, often coupled with large models and consequently high computational costs, makes these approaches unsuitable for real-time applications that require immediate detection and swift responses in dynamic environments such as autonomous driving [65, 70], wearable devices [26, 2], and human-robot interaction [50, 32].

To tackle these limitations, we introduce **On**line **A**udio-**V**isual **E**vent **P**arsing (On-AVEP), a new paradigm that parses audio, visual, and audio-visual events in streaming videos with an online processing manner. The core characteristic of On-AVEP is to perceive the environmental state and generate timely feedback using only historical and current multimodal information, while balancing model performance and efficiency particularly in resource-aware and dynamic environments.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

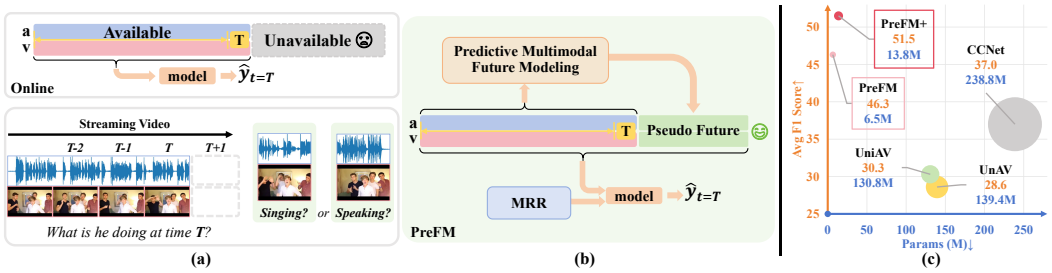

Figure 1: (a) Illustration of parsing events in online scenarios: if a man opens his mouth and produces a vocal sound at time T, it is unclear based solely on information from 0 to T whether this marks the beginning of a musical phrase (as part of "*singing*") or the start of a conversation (as in "*speaking*"). Precisely parsing events with these limited context is crucial for accurate online inference. (b) Simplified architecture of our PreFM framework, highlighting predictive future modeling and modality-agnostic robust representation (MRR). (c) Comparison of performance and efficiency against SOTA methods on the UnAV-100 [13] dataset.

Specifically, On-AVEP necessitates the model possess two key capabilities: (1) **Accurate Online Inference**: This requires that the model adapts to complex and dynamic scene variations and accurately predicts ongoing events by relying exclusively on past and current information without any future context. As illustrated in Figure 1(a), the model needs to distinguish similar events with unclear, limited context due to the lack of future information. (2) **Real-time Efficiency**: To meet the immediate response demands of On-AVEP applications, the model needs to achieve accurate event parsing with low computational cost, balancing performance and complexity well to satisfy the needs of online video processing.

To cultivate these essential capabilities, we introduce the **Pre**dictive **F**uture **M**odeling (PreFM) framework as illustrated in Figure 1(b). PreFM aims to predict future states through effective temporal-modality feature fusion and leverage knowledge distillation and temporal prioritization for training efficiency. To achieve (1) accurate online inference, PreFM employs **predictive multimodal future modeling**, using available data and fusing their features to infer beneficial future audio-visual cues. The cross-temporal and cross-modal feature interactions are utilized to effectively reduce noise within the pseudo-future context and enhance current representations. For (2) balancing real-time efficiency and overall parsing performance, PreFM integrates two designs during training: **modality-agnostic robust representation** distills rich, modality-agnostic knowledge from a large pre-trained teacher model for more generalized representation, and **focal temporal prioritization** encourages the model to focus on the most temporally critical information for online decisions, thereby boosting the model's inference accuracy while keeping high inference efficiency.

Extensive experiments on two challenging datasets, UnAV-100 [13] and LLP [54], demonstrate that the PreFM framework significantly outperforms existing state-of-the-art (SOTA) methods in both segment-level and event-level metrics. Moreover, PreFM exhibits substantial advantages in model efficiency, striking a superior balance between performance and model complexity, with the margin of $+\mathbf{9.3}$ in event-level average F1 score and merely $\mathbf{2.7}\%$ parameters as highlighted in Figure 1(c).

In summary, our main contributions are:

- **(I)** We introduce Online Audio-Visual Event Parsing (On-AVEP), a new paradigm for real-time multimodal understanding. To our knowledge, this is the first work to systematically address the challenge of parsing audio, visual, and audio-visual events from streaming video. We further establish that success in this paradigm requires two critical capabilities: (a) accurate online inference from limited context, and (b) real-time efficiency to balance performance with computational cost.

- **(II)** We propose the PreFM framework, a novel and efficient architecture for On-AVEP. PreFM's core innovations include: (a) Predictive Multimodal Future Modeling mechanism to overcome the critical problem of missing future context; and (b) a combination of Modality-agnostic Robust Representation and Focal Temporal Prioritization to enhance model robustness and efficiency during training, providing an insightful approach to multimodal real-time video understanding.

- **(III)** We establish new SOTA performance with unprecedented efficiency. Extensive experiments on two public datasets show that PreFM drastically outperforms previous methods (e.g., +9.3 Avg F1-score on UnAV-100), while using a fraction of the computational resources (e.g., only 2.7% of the parameters of the next best model), validating it as a powerful and practical solution.

2 Related Work

Online Video Understanding encompasses online action detection for identifying actions [56, 7, 44], action anticipation for predicting future [6, 73], and online temporal action localization [48, 51] for determining action boundaries. Frameworks like JOADAA [17], TPT [67] and MAT [58] jointly model detection and anticipation tasks, bridging the present and future. Recent research focuses on model reliability through uncertainty quantification [18] and adaptability through open-vocabulary detection [61]. Concurrent advancements explore leveraging large language models for complex online understanding tasks [33, 3, 69]. However, these methods rely solely on the visual modality and neglect the crucial auditory perception, motivating our research into online audio-visual event parsing aiming to integrate both sensory streams for a more robust and holistic real-time understanding.

Audio Visual Video Parsing (AVVP) aims to temporally classify videos within segments as audible or visible events. Early weakly-supervised methods [54, 66] use attention to infer temporal structure. Subsequent works [10–12, 4] further address modality imbalance and interaction. A significant recent trend involves leveraging external knowledge, using language prompts [9] or pre-trained models like CLIP [46]/CLAP [8] to denoise or generate finer pseudo-labels from weak supervision [29, 77, 30]. Building on this, methods such as CoLeaF [49], NREP [27], and MM-CSE [71] focuse on sophisticated feature disentanglement and interaction for improved performance.

Audio Visual Event Localization (AVEL) is first introduced to temporally locate events that are both visually and auditorily present within trimmed video clips [53]. Subsequent methods [63, 41, 28, 36, 23, 74] leverage cross-modal attention, background suppression, contrastive smaples and adapters to improve localization accuracy. AVE-PM [37] is developed to handle portrait-mode short videos, while OV-AVEL [75] extends the task into an open-vocabulary setting. For densely annotated, untrimmed videos featuring multiple overlapping events, UnAV [13] releases the UnAV-100 benchmark and inspires models like UniAV [14], LOCO [64], FASTEN [39] and CCNet [78], which employ multi-temporal fusion, local correspondence correction and cross-modal consistency for dense event localization. Recent efforts [68, 15, 52, 40] also aim to omni-understanding using powerful large language models. However, these approaches generally rely on full-video inputs and huge model sizes, making them unsuitable for real-time parsing. Our work distinguishes itself by unifying AVEL and AVVP into a comprehensive online audio-visual event parsing framework, designed for efficient real-time processing and capable of identifying events regardless of whether they are solely auditory, visual or audio-visual.

3 Methods

In this section, we first introduce the problem setup in Sec. 3.1 and present a brief overview of our method in Sec. 3.2, with core designs: Predictive Multimodal Future Modeling (Sec. 3.3), Modality-agnostic Robust Representation (Sec. 3.4), and Focal Temporal Prioritization (Sec. 3.5). Finally, we discuss the specifics of our approach during training and online inference in Sec. 3.6.

3.1 Preliminaries

On-AVEP involves predicting events within streaming videos by sequentially processing multimodal information. This task is primarily divided into two sub-tasks: online audio-visual event localization (On-AVEL) and online audio-visual video parsing (On-AVVP). In the former, given a sequence of audio-visual data pairs $\{V_t, A_t\}_{t=1}^{T}$ and the corresponding label $y_{t=T}$, where T denotes the current time step, the model is required to predict the multi-label event vector $\hat{y}_{t=T} \in \{0,1\}^{C_{av}}$, where C_{av} represents the total number of audio-visual event categories. While the latter task involves predicting $\hat{y}_{t=T} \in \{0,1\}^{C_a+C_v}$, where C_a and C_v represent the number of audio-only and visual-only events, respectively. In both sub-tasks, models typically take the pre-processed visual-audio feature vectors

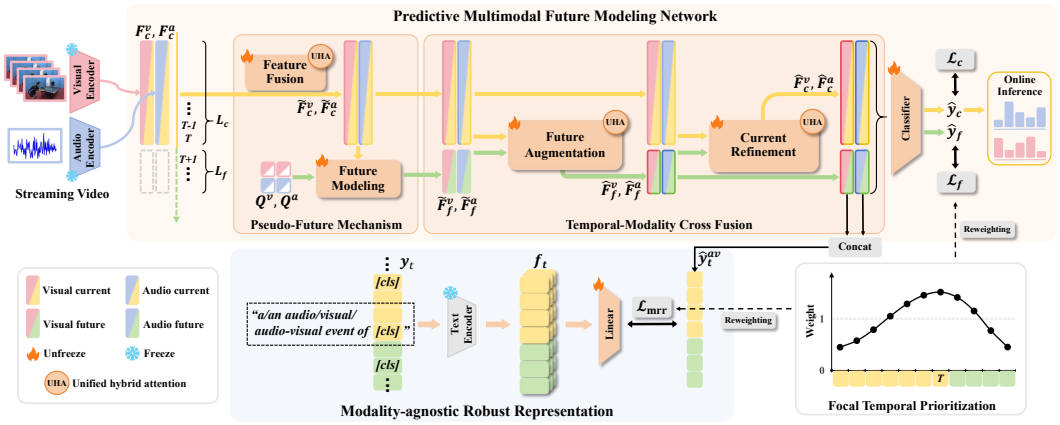

Figure 2: The pipeline of PreFM. It takes real-time audio-visual streams, using predictive modeling to generate multimodal future context, modality-agnostic robust representation to enhance performance by transferring knowledge, and focal temporal prioritization to emphasize the current time step T.

$\{f_t^a\}_{t=1}^T$ and $\{f_t^v\}_{t=1}^T \in \mathbb{R}^{T\times D}$ (D: feature dimension) within existing video datasets [54, 25, 53, 13] for subsequent operations.

3.2 Overview

As illustrated in Figure 2, during online inference, PreFM sequentially processes incoming audio and visual features F_c^a, F_c^v available up to the current time T. To address the challenge of missing future information, which is crucial for event disambiguation, the core **Predictive Multimodal Future Modeling** network (Sec. 3.3) dynamically generates pseudo-future multimodal sequences. This process starts with a *Pseudo-Future Mechanism* that fuses current-time multimodal features and subsequently models initial pseudo-future predictions $\tilde{F}_f^a, \tilde{F}_f^v$, then *Temporal-Modality Cross Fusion* where pseudo-future cues and current representations are mutually enhanced through comprehensive cross-temporal and cross-modal interactions. The resulting contextually augmented representations $\hat{F}_c^a, \hat{F}_c^v$ are then utilized for event parsing at time T.

To train an effective and efficient PreFM model, in addition to direct supervision on predictions for both the current window and the pseudo-future sequences, PreFM utilizes **Modality-agnostic Robust Representation** (MRR, Sec. 3.4). Through MRR, event labels y_t are transformed into target modality-agnostic features f_t using a pre-trained teacher model; PreFM's internal event representations $\hat{y}_t^{av}$ are then guided to align with these target features via a dedicated distillation loss term. Furthermore, **Focal Temporal Prioritization** (Sec. 3.5) is implemented by reweighting the contributions of different relative time steps, encouraging the model to make precise predictions at current time.

3.3 Predictive Multimodal Future Modeling Network

Inspired by advances in online action detection [58, 17, 67], our approach to On-AVEP centers on predictively modeling multimodal pseudo-future sequences using only currently available data. To help PreFM better utilize and consolidate all available modal and temporal cues, we propose a Universal Hybrid Attention (UHA) block to bridge different modalities across time. Given the target query sequence Q and the flexible list of k context sets $\{F_i\}_{i=1}^k$ where each F_i can represent various temporal segments of different modalities, UHA merges these features into Q as follows:

$$\text{UHA}(Q, \{F_i\}_{i=1}^k) = \text{FFN}(\text{LN}(Q + \sum_{i=1}^{k} \text{Attn}(Q, F_i, F_i))) \tag{1}$$

Where Attn is multi-head attention [57], LN is Layer Normalization, and FFN is a Feed-Forward Network. UHA serves as the foundational attention block for subsequent fusion operations.

Pseudo-Future Mechanism This mechanism first fuses current audio-visual information and then models an initial prediction of the future sequence. Given input features up to current time T, $\{(f_t^v, f_t^a)\}_{t=1}^T$, we define a current working window of length $\boldsymbol{L_c}$. This yields the initial current audio and visual features $F_c^a = \{f_t^a\}_{t=T-L_c+1}^T$ and $F_c^v = \{f_t^v\}_{t=T-L_c+1}^T$, both in $\mathbb{R}^{L_c \times D}$.

First, we perform an initial *feature fusion* between F_c^a and F_c^v. Each sequence is processed by our UHA block with both as context. The fused current features $\tilde{F}_c^a, \tilde{F}_c^v \in \mathbb{R}^{L_c \times D}$ are produced by:

$$\tilde{F}_c^m = \text{UHA}(F_c^m, \{F_c^a, F_c^v\}), m \in \{a, v\} \tag{2}$$

Next, *future modeling* generates initial pseudo-future sequences of length $\boldsymbol{L_f}$. We use learnable tokens $Q^a, Q^v \in \mathbb{R}^{L_f \times D}$ as queries. These attend to the corresponding fused current features:

$$\tilde{F}_f^m = \text{Attn}(Q^m, \tilde{F}_c^m, \tilde{F}_c^m), m \in \{a, v\} \tag{3}$$

This step yields the initial multimodal pseudo-future sequences $\tilde{F}_f^a, \tilde{F}_f^v \in \mathbb{R}^{L_f \times D}$.

Temporal-Modality Cross Fusion Having obtained the fused current features $(\tilde{F}_c^a, \tilde{F}_c^v)$ and initial pseudo-future sequences $(\tilde{F}_f^a, \tilde{F}_f^v)$, this stage performs further interactions to mutually refine them, reducing potential noise within the pseudo-future, while simultaneously enriching the current representations with foresight gleaned from the modeled future.

First, *future augmentation* refines the initial pseudo-future predictions with UHA block:

$$\hat{F}_f^m = \text{UHA}(\tilde{F}_f^m, \{\tilde{F}_f^a, \tilde{F}_f^v, \tilde{F}_c^m\}), m \in \{a, v\} \tag{4}$$

This yields augmented pseudo-future sequences $\hat{F}_f^a, \hat{F}_f^v \in \mathbb{R}^{L_f \times D}$. Notably, the context list within the UHA block enables a rich combination of self-attention, as well as cross-interactions across modalities and time. For instance, the augmented visual pseudo-future $\hat{F}_f^v$ is obtained by interacting with $\tilde{F}_f^v$ (for self-attention), $\tilde{F}_f^a$ (for cross-modal attention), and $\tilde{F}_c^v$ (for cross-temporal attention) as shown in Figure 3.

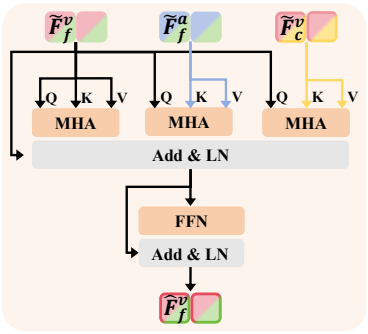

Figure 3: Temporal-modality cross fusion for the pseudo-future $\hat{F}_f^v$.

Next, *current refinement* integrates the augmented future back into the current representations:

$$\hat{F}_c^m = \text{UHA}(\tilde{F}_c^m, \{\tilde{F}_c^a, \tilde{F}_c^v, \hat{F}_f^m\}), m \in \{a, v\} \tag{5}$$

This results in the final contextually-aware current feature sequences $\hat{F}_c^a, \hat{F}_c^v \in \mathbb{R}^{L_c \times D}$.

Finally, these augmented current and future features are projected by a shared classification head $h(\cdot)$ and a Sigmoid function $\mathcal{S}(\cdot)$ to obtain event predictions $\hat{y}_c \in \mathbb{R}^{L_c \times C}$ (for the current window) and $\hat{y}_f \in \mathbb{R}^{L_f \times C}$ (for the future window):

$$\hat{y}_c = \mathcal{S}(h(\texttt{Concat}(\hat{F}_c^a, \hat{F}_c^v))), \hat{y}_f = \mathcal{S}(h(\texttt{Concat}(\hat{F}_f^a, \hat{F}_f^v))) \tag{6}$$

Here, `Concat` denotes feature concatenation, and C is the event class count (either C_{av} or $C_a + C_v$). For online inference, the prediction in $\hat{y}_c$ corresponding to time T is used. While during training, $\hat{y}_c$ and $\hat{y}_f$ are supervised across the time steps of $[T - L_c + 1, T + L_f]$, using BCE loss with annotations $y_c \in \mathbb{R}^{L_c \times C}$ and $y_f \in \mathbb{R}^{L_f \times C}$:

$$\mathcal{L}_c = \text{BCE}(\hat{y}_c, y_c), \mathcal{L}_f = \text{BCE}(\hat{y}_f, y_f) \tag{7}$$

3.4 Modality-agnostic Robust Representation

Learning from the rich, modality-agnostic event representations established by powerful pre-trained teacher models [46, 8, 16, 59, 20] is efficient to obtain robust and generalizable representations

while maintaining efficiency. For each time step t, we convert the event labels y_t into "*a/an audio/visual/audio-visual event of* [cls]" as text prompt, which is then processed by the text encoder of frozen teacher model OnePeace [59] to obtain modality-agnostic event features f_t. Simultaneously, the student's representation $\hat{f}_t^{av} = \texttt{Concat}(\hat{f}_t^a, \hat{f}_t^v)$ can be easily extracted from our PreFM model, and distilled by the target representation f_t. Cosine similarity is used as distillation loss as follows:

$$\mathcal{L}_{mrr} = \frac{1}{L_c + L_f} \sum_{t=T-L_c+1}^{T+L_f} (1 - \frac{\hat{f}_t^{av} \cdot h'(f_t)}{\|\hat{f}_t^{av}\| \cdot \|h'(f_t)\|}) \tag{8}$$

Where $h'(\cdot)$ denotes the projector module implemented by a linear layer to align different dimensions.

3.5 Focal Temporal Prioritization

To emphasize the predictions close to the present moment, we introduce a focal temporal prioritization scheme to the loss calculation, highlighting the significance of prediction at time step T instead of a uniform weighting. Specifically, we define temporal priorities using a Gaussian function centered at the current time T: $g(t, \sigma) = \exp\left(-\frac{(t)^2}{2\sigma^2}\right)$, where t is the relative temporal distance from time T, and σ controls the width of the focus. We define the temporal weights $w_c(t) \in [T - L_c + 1, T]$ for the current window, and $w_f(t) \in [T + 1, T + L_f]$ for pseudo-future sequences,

$$w_c(t-T) = \frac{L_c \cdot g(t-T, L_c)}{\sum_{k=T-L_c+1}^{T} g(k-T, L_c)}, \quad w_f(t-T) = \frac{L_f \cdot g(t-T, L_f)}{\sum_{k=T+1}^{T+L_f} g(k-T, L_f)} \tag{9}$$

Let $\mathcal{L}_c(t)$, $\mathcal{L}_f(t)$ and $\mathcal{L}_{mrr}(t)$ be the per-timestep loss in Eq. 7 and Eq. 8. We use $w(t-T) = \texttt{Concat}\{w_c(t-T), w_f(t-T)\}$ to obtain the while weights sequence vector. The final loss is computed as:

$$\mathcal{L} = \sum_{t=T-L_c+1}^{T} w_c(t-T) \cdot \mathcal{L}_c(t) + \sum_{t=T+1}^{T+L_f} w_f(t-T) \cdot \mathcal{L}_f(t) + \lambda \sum_{t=T-L_c+1}^{T+L_f} w(t-T) \cdot \mathcal{L}_{mrr}(t) \tag{10}$$

The hyperparameter λ balances the robust representation term (typically 1.0).

3.6 Training and Online Inference

Random Segment Sampling for Training To adapt training for the online nature of On-AVEP and enhance data utilization, we design a random segment sampling strategy. During training, for a video of total length T_{all}, the target prediction times $T_k \in [1, T_{all}]$ are generated by $T_k = kL_c + \delta$. Here, k serves as an index for iterating across the video, and $\delta \in [0, L_c - 1]$ is a periodically selected random integer offset. The L_c-length feature sequences $\{(f_t^v, f_t^a)\}_{t=T_k-L_c+1}^{T_k}$ act as model inputs, and zero-padding is applied at the beginning if $T_k < L_c - 1$. This strategy provides diverse training segments with a fixed history length L_c, suitable for the online setting.

Online Inference During inference, the model works in a truly online manner, processing the input video stream with a sliding window of length L_c and stride 1. At each step T_{infer}, the model takes features from $[T_{infer} - L_c + 1, T_{infer}]$, generates the multimodal pseudo-future context, and gets the final event predictions for the current time step T_{infer}.

4 Experiments

4.1 Experimental Setups

Dataset **UnAV-100** [13] is a large-scale dataset designed for dense audio-visual event localization in untrimmed videos. It contains 10,790 videos of varying lengths covering 100 event categories, with over 30,000 annotated audio-visual event instances. **LLP** [54] provides 11,849 trimmed 10-second clips across 25 categories for audio-only and visual-only event parsing. For online scenarios, we concatenate LLP clips into longer video sequences. Specifically, half of these sequences are formed

Table 1: Comparison with SOTA methods on On-AVEL task. Feature extractors: *I.V.* denotes I3D [1]+VGGish [24], *O.* denotes OnePeace [59] and *C.C.* denotes CLIP [46]+CLAP [8]. "PreFM+" replaces the feature extractor from CLIP/CLAP to OnePeace and the hidden dimension is expanded from 256 to 512. Methods marked with "*" take the entire video as input, fully utilizing the complete context.

Methods	Extractors	Segment-Level		Event-Level						Params↓	FLOPs↓	Inference		
		F1	mAP	0.1	0.3	0.5	0.7	0.9	Avg			Memory↓	FPS↑	Latency ↓
UnAV [13]	*I.V.*	47.5	58.3	50.9	37.1	28.7	18.2	9.4	28.6	139.4M	52.4G	764.7MB	10.6	94.3ms
UniAV [14]	*O.*	47.8	66.9	50.3	38.9	29.9	21.1	12.3	30.3	130.8M	22.7G	1020.5MB	15.6	64.1ms
CCNet [78]	*O.*	54.8	62.3	58.3	46.3	37.5	27.3	15.8	37.0	238.8M	72.1G	1179.4MB	7.5	133.3ms
PreFM (Ours)	*C.C.*	59.1	70.1	61.5	53.6	46.9	39.6	29.2	46.3	**6.5M**	**0.4G**	**56.4MB**	**51.9**	**19.3ms**
PreFM+ (Ours)	*O.*	**62.4**	**70.6**	**66.3**	**58.2**	**52.2**	**44.5**	**35.4**	**51.5**	13.8M	0.5G	144.2MB	42.0	23.8ms
UnAV* [13]	*I.V.*	56.1	67.8	59.3	56.0	52.7	46.7	35.1	50.6	139.4M	52.4G	764.7MB	10.6	94.3ms
UniAV* [14]	*O.*	59.2	70.0	62.8	59.0	55.1	48.7	35.0	52.9	130.8M	22.7G	1020.5MB	15.6	64.1ms
CCNet* [78]	*O.*	65.0	70.6	69.0	65.1	61.0	53.1	40.1	58.3	238.8M	72.1G	1179.4MB	7.5	133.3ms

Table 2: Comparison with SOTA methods on On-AVVP task. Feature extractors: *R.C.C.* denotes R3D [55]+CLIP [46]+CLAP [8] and *R.R.V.* denotes R3D [55]+ResNet152 [21]+Vsh: VGGish [24]. "PreFM+" increases the hidden dimension from 128 to 256. Methods marked with "*" take the entire 10-second video clips as input, fully utilizing the complete context.

Methods	Extractors	Segment-Level						Event-Level						Params↓	FLOPs↓	Inference		
		$F1_a$	$F1_v$	$F1_{av}$	mAP_a	mAP_v	mAP_{av}	0.5_a	0.5_v	0.5_{av}	Avg_a	Avg_v	Avg_{av}			Memory↓	FPS↑	Latency ↓
VALOR [29]	*R.C.C.*	49.7	52.4	45.4	72.9	68.4	56.7	36.5	46.1	34.6	35.2	42.8	33.0	4.9M	0.45G	**20.1MB**	62.2	16.1ms
CoLeaF [49]	*R.R.V.*	50.7	44.5	41.0	62.8	45.8	37.3	37.9	36.4	29.6	37.3	35.5	29.7	5.7M	0.25G	114.1MB	60.4	16.6ms
LEAP [76]	*R.R.V.*	50.6	49.3	45.8	73.3	64.3	54.6	40.1	42.5	35.9	38.4	39.6	34.3	52.0M	1.09G	204.7MB	19.3	51.8ms
NREP [27]	*R.C.C.*	53.7	51.4	45.5	66.5	52.7	42.3	38.9	45.6	34.2	38.3	42.3	33.5	9.6M	1.69G	90.2MB	26.4	37.9ms
MM-CSE [71]	*R.C.C.*	53.3	56.5	48.9	74.6	70.0	57.5	39.4	50.8	38.4	37.7	46.9	36.2	6.2M	0.91G	33.0MB	36.1	27.7ms
PreFM (Ours)	*R.C.C.*	60.0	59.3	53.3	80.0	73.7	61.3	47.1	50.9	42.0	46.3	50.6	41.2	**3.3M**	**0.22G**	20.7MB	**94.4**	**10.6ms**
PreFM+ (Ours)	*R.C.C.*	**61.0**	**60.0**	**54.6**	**80.2**	**73.8**	**61.4**	**48.5**	**51.7**	**43.1**	**47.6**	**51.0**	**42.2**	12.1M	0.48G	55.9MB	53.5	18.7ms
VALOR* [29]	*R.C.C.*	65.6	61.8	56.5	81.4	73.7	61.4	55.1	54.9	46.7	54.0	54.2	46.0	4.9M	0.45G	20.1MB	62.2	16.1ms
CoLeaF* [49]	*R.R.V.*	60.5	58.0	52.4	71.7	60.7	49.3	48.3	53.0	42.1	48.7	51.8	42.5	5.7M	0.25G	114.1MB	60.4	16.6ms
LEAP* [76]	*R.R.V.*	61.6	61.5	56.5	80.6	71.3	60.2	52.3	56.4	47.7	51.2	55.0	46.7	52.0M	1.09G	204.7MB	19.3	51.8ms
NREP* [27]	*R.C.C.*	67.3	63.7	57.9	77.4	66.2	53.9	55.9	57.5	47.8	54.9	56.7	47.1	9.6M	1.69G	90.2MB	26.4	37.9ms
MM-CSE* [71]	*R.C.C.*	67.0	64.0	57.6	82.3	74.8	61.7	56.9	56.8	47.3	54.7	56.0	46.1	6.2M	0.91G	33.0MB	36.1	27.7ms

by randomly concatenating clips to simulate the rapid scene variations often encountered in online streaming content; the other half are formed by concatenating clips from the same event category to represent longer, continuous event occurrences. Following recent works [29, 77, 9, 49, 27, 71], segment-wise pseudo labels from CLIP [46, 22] and CLAP [8] are used for supervision.

Metric For model performance, we follow prior work [58, 54], using F1-score and mean Average Precision (mAP) as segment-level metrics. For event-level evaluation, consecutive positive segments are treated as a complete event instance. We calculate event-level F1-scores by setting tiou = [0.1:0.1:0.9] [13] and average F1-score (Avg F1-score) for overall performance. For the On-AVVP task, we adhere to the established protocol from VALOR [29], evaluating audio-only (A), visual-only (V), and combined audio-visual (AV, denoted with subscript "av") events. Regarding model efficiency, we assess the number of trainable parameters, FLOPs per inference, peak inference memory and FPS.

Implementation details For both tasks, we set 60 training epochs, with the first 10 epochs dedicated to warm-up. A batch size of 128 is used, and AdamW serves as the optimizer with a weight decay of $1e^{-4}$. We set the value L_c of 10 and L_f of 5 as the default setting. CLIP [46] and CLAP [8] are used to extract visual and audio features with a temporal stride set to 1 second, respectively. All experiments are conducted on a single RTX 3090. For the learning rate and the hidden dimension within the attention block, we use $1e^{-3}$ and 256 for On-AVEL, $5e^{-4}$ and 128 for On-AVVP.

4.2 Comparison with Existing Work

Our method is benchmarked against recent SOTA methods UnAV [13], UniAV [14] and CCNet [78] on UnAV-100 [13] for the On-AVEL task, while VALOR [29], CoLeaf [49], LEAP [76], NREP [27], and MM-CSE [71] on LLP [54] for the On-AVVP task. We provide two versions of our method: the basic version "PreFM", and the improved version "PreFM+" with larger hidden size.

Performance Comparison As shown in Table 1, PreFM clearly achieves new SOTA results for *On-AVEL* task, surpassing the second-best method with significant improvement of **+7.8** in mAP and **+9.3**

in Avg F1-score. Furthermore, our enhanced version, PreFM+, extends these gains to **+8.3** in mAP and **+14.5** in Avg F1-score with only a moderate increase in parameters, highlighting the excellent scalability of the PreFM architecture for applications requiring higher precision. Similarly for *On-AVVP* task shown in Table 2, PreFM demonstrates consistent advantages, achieving improvements of **+3.8** in mAP_{av} and **+5.0** in Avg F1-score_{av} and PreFM+ further elevating performance to **+3.9** in mAP_{av} and **+6.0** in Avg F1-score_{av} over the second-best methods. Notably, we also present the original offline results of these baseline methods (marked with "*") to show their performance under full-context conditions. Even when compared to these results, our online PreFM achieves comparable performance despite predicting with limited context.

These substantial performance gains across both tasks are largely attributed to our core predictive multimodal future modeling (PMFM) design. By dynamically generating and integrating pseudo-future contextual cues from streaming data, PMFM empowers our method to effectively parse environmental states and accurately capture temporal boundaries.

Efficiency Analysis Regarding the *On-AVEL* task (Table 1), PreFM's efficiency is remarkable. PreFM utilizes merely **2.7**% parameters (*6.5M vs 238.8M*) compared to the next best performing method, and it requires only **0.6**% FLOPs (*0.4G vs 72.1G*) and **4.8**% peak memory (*56.4MB vs 1179.4MB*) for a single inference, while running at an impressive 51.9 FPS with merely a latency of **19.3ms**. The compelling efficiency advantage is also evident in the *On-AVVP* task (Table 2). Such ability to deliver SOTA performance with drastically reduced overhead highlights that PreFM is designed with a strong emphasis on practical deployability, rendering it a highly suitable and efficient solution for resource-constrained real-time applications.

4.3 Ablation Studies

Main Component To systematically evaluate the contribution of each proposed component, we conduct comprehensive ablation studies on the On-AVEL task, with results presented in Table 3(a). The simple prediction strategy (row 1) uses only data at time T and performs badly. Our baseline (row 2), which just extends accessible data to context L_c but no more improvements, achieves an Avg F1-score of 40.8%. Introducing the pseudo future mechanism (PF, row 3) significantly boosts performance to 42.4 (+1.6 vs baseline), underscoring the importance of future context modeling over relying solely on past or current information. Further incorporating modality-agnostic robust representation ($\mathcal{L}_{mrr}$, row 5) or random segment sampling (RS, row 6) individually builds upon this, yielding Avg F1-scores of 44.2 (+3.4 vs baseline) and 44.0 (+3.2 vs baseline) respectively, demonstrating their distinct benefits. The focal temporal prioritization ($w(t)$) consistently improves results when applied (e.g., row 4 vs 3, and row 7 vs 5), confirming its effectiveness in focusing the model on critical information at the current moment. Finally, our full PreFM model (row 8), integrating all components, achieves a final Avg F1-score of 46.3, marking a substantial +5.5 improvement over the baseline and validating the collective effectiveness of our design.

Impact of future-oriented losses We investigate the impact of direct future supervision $\mathcal{L}_f$ and future part of robust representation loss $\mathcal{L}_{mrr,f}$ used in pseudo-future (PF) mechanism. Results are shown in Table 3(b). A comparison of the first two rows shows that merely incorporating the extra parameters in the future module without applying any future supervision yields negligible performance gains. Conversely, the results in the subsequent three rows indicate that designing losses to explicitly guide the model in anticipating and modeling the future, whether through direct supervision or robust representation distillation, enhances model performance. These findings clearly demonstrate that the performance benefits derived from our pseudo-future mechanism are primarily attributable to the effective learning guided by these targeted future-oriented losses, rather than merely an increase in model capacity.

Impact of temporal-modality cross fusion Generating reliable audio-visual pseudo-future is challenging due to inherent predictive noise. Table 3(c) compares our Temporal-Modality Cross Fusion (TMCF) with ablated variants that utilize only self-attention (Self), audio-visual modality fusion (M only), or temporal-only fusion (T only), focusing on their accuracy in predicting the future (the first three relative time steps) and overall event parsing performance. The inferior performance of these simplified variants underscores that uni-dimensional interactions are insufficient for producing robust future sequences, leaving its reliability and noise levels suboptimal. In contrast, our full TMCF,

Table 3: (a) Overall ablation study. PF: the pseudo future mechanism, $w(t)$: focal temporal prioritization, $\mathcal{L}_{mrr}$: modality-agnostic robust representation, RS: random segment sampling. (b) Ablation studies for future-oriented losses. (c) Ablation studies for temporal-modality cross fusion. Params: trainable parameters, S-L: Segment-Level, E-L: Event-Level.

	PF	$w(t)$	$\mathcal{L}_{mrr}$	RS	S-L mAP	E-L Avg
(1)	Simple Predictions				66.6	29.9
(2)	✗	✗	✗	✗	69.1	40.8$_{+0.0}$
(3)	✓	✗	✗	✗	69.7	42.4$_{+1.6}$
(4)	✓	✓	✗	✗	69.7	43.0$_{+2.2}$
(5)	✓	✗	✓	✗	69.8	44.2$_{+3.4}$
(6)	✓	✗	✗	✓	70.5	44.0$_{+3.2}$
(7)	✓	✓	✓	✗	69.4	45.4$_{+4.6}$
(8)	✓	✓	✓	✓	70.1	**46.3**$_{+5.5}$

(a)

$\mathcal{L}_f$	$\mathcal{L}_{mrr,f}$	PF	Params	S-L mAP	E-L Avg
✗	✗	✗	2.6M	69.8	44.7$_{+0.0}$
✗	✗	✓	6.5M	69.6	44.8$_{+0.1}$
✗	✓	✓	6.5M	69.9	45.2$_{+0.5}$
✓	✗	✓	6.5M	69.6	45.5$_{+0.8}$
✓	✓	✓	6.5M	70.1	**46.3**$_{+1.6}$

(b)

	S-L F1 T+1	S-L F1 T+2	S-L F1 T+3	E-L Avg
Self	55.4	54.6	53.7	**44.6**
T only	56.7	56.0	55.5	**45.3**
M only	56.6	55.7	55.1	**45.0**
TMCF	**57.5**	**56.5**	**55.4**	**46.3**

(c)

Table 4: (a) Ablation studies for different length of past and future. (c) Ablation studies for different pre-trained teacher models.

	L_c	L_f	Seg-Level F1	Seg-Level mAP	Event-Level 0.5	Event-Level Avg
(1)	10	1	58.8	70.1	46.8	45.9
(2)	10	5	59.1	70.1	46.9	**46.3**
(3)	10	10	57.3	69.9	46.4	45.4
(4)	5	5	57.2	69.9	44.1	43.5
(5)	20	5	48.9	65.7	38.9	38.5

(a)

Models	Dimensions	Seg-Level F1	Seg-Level mAP	Event-Level 0.5	Event-Level Avg
AudioClip [20]	1024	58.7	**70.2**	46.6	46.2
ImageBind [16]	1024	58.3	70.0	**47.2**	45.9
ONE-PEACE [59]	1536	**59.1**	70.1	46.9	**46.3**

(b)

by collaboratively leveraging both cross-modal and cross-temporal interactions from available content, generates more accurate and dependable pseudo-future sequences. This results in a higher-quality predictive context that more effectively mitigates noise and aids robust real-time event parsing.

Impact of context lengths L_c and L_f We investigate the impact of varying lengths for the working area L_c and the pseudo-future sequence L_f, with results presented in Table 4(a). The optimal performance, achieving an Avg F1-score of 46.3, is obtained with our default configuration of $L_c = 10$ and $L_f = 5$ (row 2). Analysis of L_f (rows 1, 2, 3, with $L_c = 10$ fixed) indicates that while a very short future window ($L_f = 1$) provides insufficient predictive insight, an overly long one ($L_f = 10$) can introduce distracting noise, both degrading performance. Similarly, examining L_c (rows 2, 4, 5, with $L_f = 5$ fixed) reveals that too little historical context ($L_c = 5$) offers inadequate support, whereas excessive history ($L_c = 20$) may include outdated or irrelevant information. These findings confirm the importance of appropriately sized context windows, with $L_c = 10$ and $L_f = 5$ providing the most effective balance for the immediate event parsing task.

Different teacher models in MRR We evaluate the influence of different pre-trained teacher models on our modality-agnostic robust representation (MRR) module. Specifically, we compare OnePeace [59], ImageBind [16], and AudioClip [20] across multiple metrics. As the results demonstrate in Table 4(b), no single model consistently outperforms the others on every measure. However, OnePeace delivers better segment-level F1-scores and average event-level performance, which lead us to adopt it as our default teacher.

Temporal impact of the pseudo-future Figure 4(a) illustrates how our pseudo future (PF) mechanism affects prediction accuracy across time steps relative to the current moment T. From the orange line, we observe that the model's peak performance occurs significantly earlier (around relative time $T - 6$), with accuracy declining as it approaches T, indicating a strong reliance on full context. In contrast, the purple line shows that incorporating the PF not only achieves generally higher accuracy but also shifts its performance peak much closer to the actual target time T (around $T - 2$). These observations underscore a fundamental principle in event parsing: accurate event identification intrinsically depends on a comprehensive contextual window. Thus, the reliance on future context presents a significant hurdle for online systems. Our PF mechanism effectively

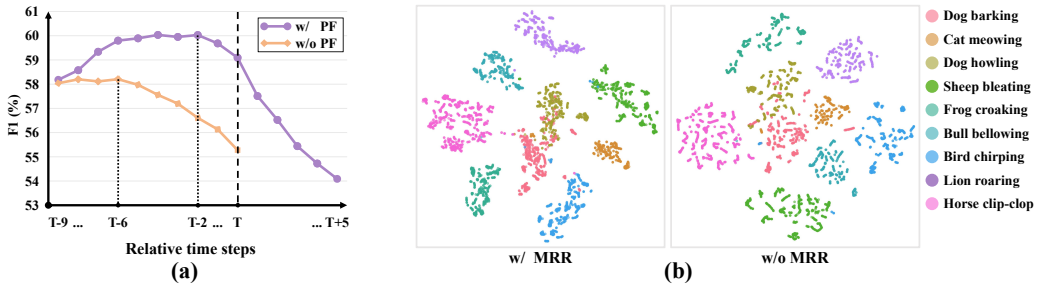

Figure 4: (a) The performance across different relative time steps. (b) t-SNE visualization of the pre-classifier features. We use nine animal events from UnAV-100 [13] for better illustration.

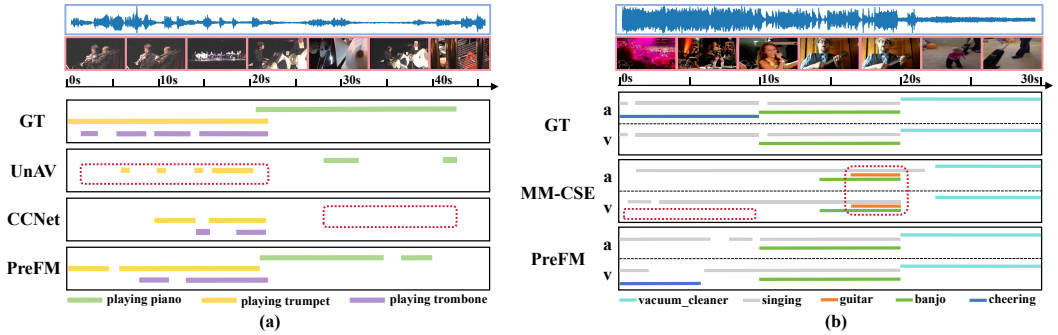

Figure 5: (a) The visualization on the On-AVEL task. (b) The visualization on the On-AVVP task. GT: ground truth. The red dotted box indicates the area of mispredictions.

anticipates event trends, models and utilizes audio-visual future information to enhance prediction accuracy near the present moment, thereby mitigating immediate contextual limitations.

Impact of MRR on latent feature representation Figure 4(b) qualitatively evaluates our modality-agnostic robust representation (MRR) via t-SNE visualization of latent features from nine predefined animal events from UnAV-100 [13]. With MRR, event classes form more compact and well-separated clusters, unlike the more chaotic clusters from the model without MRR. This suggests that while MRR may shift the latent space, it guides the model towards more discriminative representations, enhancing event separability and overall performance.

4.4 Qualitative Analysis

Figure 5 presents a qualitative comparison of our PreFM with SOTA methods UnAV [13], CCNet [78] on On-AVEL task and MM-CSE [71] on On-AVVP tasks. Prior methods often exhibit limitations such as missed detections (e.g., "trombone" by UnAV, "piano" by CCNet, "cheering" by MM-CSE), fragmented predictions (e.g., "trumpet" by UnAV) depicted by red dotted box. In stark contrast, PreFM's predictions exhibit strong temporal continuity and precise event boundary localization, without the interruptions or errors in other methods. These visualizations intuitively showcase PreFM's enhanced recognition accuracy and the coherent, continuous nature of its event parsing.

5 Conclusions

In this work, we introduce online audio-visual event parsing to enable real-time multimodal event understanding in streaming videos. We identify accurate online inference and real-time efficiency as two crucial capabilities in this setting, and propose the PreFM framework, featuring a novel predictive multimodal future modeling to infer future context and modality-agnostic robust representation together with focal temporal prioritization for model's generalization. Extensive experiments on the UnAV-100 and LLP datasets validate that PreFM significantly outperforms prior methods, achieving state-of-the-art performance while offering a superior balance between accuracy and computational efficiency, thus presenting a viable solution for practical real-time multimodal applications.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (No. 92470203, U23A20314), the Beijing Natural Science Foundation (No. L242022), and the Fundamental Research Funds for the Central Universities (No. 2024XKRC082).

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+ x: A panoptic multi-modal scene understanding dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19373–19382, 2024.

[3] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18407–18418, 2024.

[4] 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.

[5] Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, Julio Saez-Rodriguez, Simona Cristea, Hani Goodarzi, Mohammad Lotfollahi, Fabian J Theis, and Bo Wang. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, 2025.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[7] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 269–284. Springer, 2016.

[8] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[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*, 36:40610–40622, 2023.

[10] Jie Fu, Junyu Gao, Bing-Kun Bao, and Changsheng Xu. Multimodal imbalance-aware gradient modulation for weakly-supervised audio-visual video parsing. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4843–4856, 2023.

[11] 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.

[12] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Learning probabilistic presence-absence evidence for weakly-supervised audio-visual event perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[13] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 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*, pages 22942–22951, 2023.

[14] Tiantian Geng, Teng Wang, Yanfu Zhang, Jinming Duan, Weili Guan, and Feng Zheng. Uniav: Unified audio-visual perception for multi-task video localization. *arXiv e-prints*, pages arXiv–2404, 2024.

[15] Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. *arXiv preprint arXiv:2411.19772*, 2024.

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[17] Mohammed Guermal, Abid Ali, Rui Dai, and Francois Bremond. Joadaa: joint online action detection and action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6889–6898, 2024.

[18] Hongji Guo, Hanjing Wang, and Qiang Ji. Bayesian evidential deep learning for online action detection. In *European Conference on Computer Vision*, pages 283–301. Springer, 2024.

[19] Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. Crossmae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26721–26731, 2024.

[20] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[21] 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.

[22] Xingjian He, Sihan Chen, Fan Ma, Zhicheng Huang, Xiaojie Jin, Zikang Liu, Dongmei Fu, Yi Yang, Jing Liu, and Jiashi Feng. Vlab: Enhancing video language pre-training by feature adapting and blending. *IEEE Transactions on Multimedia*, 2024.

[23] Xiang He, Xiangxi Liu, Yang Li, Dongcheng Zhao, Guobin Shen, Qingqun Kong, Xin Yang, and Yi Zeng. Cace-net: Co-guidance attention and contrastive enhancement for effective audio-visual event localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 985–993, 2024.

[24] 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.

[25] Wenxuan Hou, Guangyao Li, Yapeng Tian, and Di Hu. Toward long form audio-visual video understanding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–26, 2024.

[26] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22910–22921, 2023.

[27] Xun Jiang, Xing Xu, Liqing Zhu, Zhe Sun, Andrzej Cichocki, and Heng Tao Shen. Resisting noise in pseudo labels: Audible video event parsing with evidential learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[28] Yuanyuan Jiang, Jianqin Yin, and Yonghao Dang. Leveraging the video-level semantic consistency of event for audio-visual event localization. *IEEE transactions on multimedia*, 26:4617–4627, 2023.

[29] 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.

[30] Yung-Hsuan Lai, Janek Ebbers, Yu-Chiang Frank Wang, François Germain, Michael Jeffrey Jones, and Moitreya Chatterjee. Uwav: Uncertainty-weighted weakly-supervised audio-visual video parsing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13561–13570, 2025.

[31] Guangyao Li, Henghui Du, and Di Hu. Boosting audio visual question answering via key semantic-aware cues. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5997–6005, 2024.

[32] Peizhen Li, Longbing Cao, Xiao-Ming Wu, Xiaohan Yu, and Runze Yang. Ugotme: An embodied system for affective human-robot interaction. *arXiv preprint arXiv:2410.18373*, 2024.

[33] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? *arXiv preprint arXiv:2501.05510*, 2025.

[34] Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3306–3314, 2024.

[35] Yan-Bo Lin and Gedas Bertasius. Siamese vision transformers are scalable audio-visual learners. In *European Conference on Computer Vision*, pages 303–321. Springer, 2024.

[36] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023.

[37] Wuyang Liu, Yi Chai, Yongpeng Yan, and Yanzhen Ren. Audio-visual event localization on portrait mode short videos. *arXiv preprint arXiv:2504.06884*, 2025.

[38] Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Continual multimodal contrastive learning. *arXiv preprint arXiv:2503.14963*, 2025.

[39] Yin Liu, Qin Wu, Mingyong Zeng, Yahan Liu, and Yuying Pan. Fasten: Video event localization based on audio-visual feature alignment and multi-scale temporal enhancement. *IEEE Signal Processing Letters*, 2025.

[40] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*, 2023.

[41] Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 5158–5167, 2023.

[42] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *International Conference on Machine Learning*, pages 25006–25017. PMLR, 2023.

[43] Shentong Mo, Weiguo Pian, and Yapeng Tian. Class-incremental grouping network for continual audio-visual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7788–7798, 2023.

[44] Zhanzhong Pang, Fadime Sener, and Angela Yao. Context-enhanced memory-refined transformer for online action detection. *arXiv preprint arXiv:2503.18359*, 2025.

[45] Weiguo Pian, Shentong Mo, Yunhui Guo, and Yapeng Tian. Audio-visual class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7799–7811, 2023.

[46] 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.

[47] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29029–29039, 2025.

[48] Sakib Reza, Yuexi Zhang, Mohsen Moghaddam, and Octavia Camps. Hat: History-augmented anchor transformer for online temporal action localization. In *European Conference on Computer Vision*, pages 205–222. Springer, 2024.

[49] 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, 2024.

[50] Alvin Shek, Bo Ying Su, Rui Chen, and Changliu Liu. Learning from physical human feedback: an object-centric one-shot adaptation method. In *2023 IEEE international Conference on Robotics and automation (ICRA)*, pages 9910–9916. IEEE, 2023.

[51] Youngkil Song, Dongkeun Kim, Minsu Cho, and Suha Kwak. Online temporal action localization with memory-augmented transformer. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[52] Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7293–7301, 2025.

[53] 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.

[54] 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.

[55] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[56] Elahe Vahdani and Yingli Tian. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4302–4320, 2022.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[58] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13824–13835, 2023.

[59] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

[60] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[61] Yi Wang, Jilan Xu, Yinan He, Zifan Song, Limin Wang, Yu Qiao, Cairong Zhao, et al. Does video-text pretraining help open-vocabulary online action detection? *Advances in Neural Information Processing Systems*, 37:47908–47930, 2024.

[62] Yihan Wu, Yifan Peng, Yichen Lu, Xuankai Chang, Ruihua Song, and Shinji Watanabe. Robust audiovisual speech recognition models with mixture-of-experts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 43–48. IEEE, 2024.

[63] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998, 2022.

[64] Ling Xing, Hongyu Qu, Rui Yan, Xiangbo Shu, and Jinhui Tang. Locality-aware cross-modal correspondence learning for dense audio-visual events localization. *arXiv preprint arXiv:2409.07967*, 2024.

[65] Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M Wolff, and Xin Huang. Vlm-ad: End-to-end autonomous driving through vision-language model supervision. *arXiv preprint arXiv:2412.14446*, 2024.

[66] 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.

[67] Haomiao Yuan, Yi Chen, Zheyan Ji, Zhichao Zheng, Yanhui Gu, and Junsheng Zhou. Throughout procedural transformer for online action detection and anticipation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[68] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[69] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

[70] Jimuyang Zhang, Zanming Huang, Arijit Ray, and Eshed Ohn-Bar. Feedback-guided autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15000–15011, 2024.

[71] Pengcheng Zhao, Jinxing Zhou, Yang Zhao, Dan Guo, and Yanxiang Chen. Multimodal class-aware semantic enhancement network for audio-visual video parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10448–10456, 2025.

[72] Xianbing Zhao, Soujanya Poria, Xuejiao Li, Yixin Chen, and Buzhou Tang. Toward robust multimodal learning using multimodal foundational models. *arXiv preprint arXiv:2401.13697*, 2024.

[73] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, pages 485–502. Springer, 2022.

[74] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2022.

[75] Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards open-vocabulary audio-visual event localization. *arXiv preprint arXiv:2411.11278*, 2024.

[76] Jinxing Zhou, Dan Guo, Yuxin Mao, Yiran Zhong, Xiaojun Chang, and Meng Wang. Label-anticipated event disentanglement for audio-visual video parsing. In *European Conference on Computer Vision*, pages 35–51. Springer, 2024.

[77] Jinxing Zhou, Dan Guo, Yiran Zhong, and Meng Wang. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision*, 132(11):5308–5329, 2024.

[78] Ziheng Zhou, Jinxing Zhou, Wei Qian, Shengeng Tang, Xiaojun Chang, and Dan Guo. Dense audio-visual event localization under cross-modal consistency and multi-temporal granularity collaboration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10905–10913, 2025.

[79] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions are outlined in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this paper are outlined in Sec A.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not involve theoretical results, assumptions or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed information on the experimental setup in Sec 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We submit our source code in the supplementary material and provide detailed information on the experimental setup in Sec 4.1.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the source code in the supplementary material, detail the experimental settings in Sec. 4.1, outline the training and inference details of our method in Sec. 3.6, and further supplemented additional online inference details in Sec. A.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following previous works [29, 76, 71, 13, 14, 78], we do not report extra error bars or statistical information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments in our method can be completed on a 3090 GPU. We also provide the model parameters and computational complexity of the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts of this paper are outlined in Sec A.1.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our paper uses the open-source code assets from VALOR [29], and the public dataset UnAV-100 [13] and LLP [54].

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our method does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Limitations and Broader Impact

Limitations While PreFM demonstrates promising results in online audio-visual event parsing, we identify a couple of key avenues for future exploration and enhancement. Firstly, the current PreFM design is primarily tailored for event detection and localization. Further research can extend its capabilities to more complex, semantically rich tasks such as video question answering or detailed captioning, and enhance its capacity for long-range temporal reasoning, potentially through integration with large language models. Secondly, while PreFM's predictive modeling of pseudo-future context is a key component for enhancing online inference, the inherent nature of future prediction means that the generated cues may not always perfectly foresee subsequent events. Although our temporal-modality cross fusion (TMCF) (detailed in Sec. 3.3) is designed to refine these predictions and mitigate potential noise by leveraging cross-modal and cross-temporal interactions—with its positive impact analyzed in Sec. 4.3— the noise may somewhat degrade performance. While TMCF offers an initial solution, further research can be developed to enhance the reliability and effectiveness of the future.

Broader Impact Our work on online audio-visual event parsing, using methods like PreFM, can greatly improve real-time AI systems. However, we must think carefully about serious ethical issues when using audio and video data. Important issues include protecting people's privacy from unwanted watching or access, reducing unfair biases that the AI might learn from its data should be considered.

A.2 Additional Results and Analysis

Quantitative Analysis of the Pseudo-Future's Quality We evaluate them from two perspectives, their semantic similarity to ground-truth event features, and their effectiveness in predicting future events: *Top-k Similarity Accuracy*: We measure if a generated feature vector at a relative future time step (T+1-T+5) is the Top-1 or Top-5 closest match to its corresponding ground-truth class feature embedding, among all 100 classes in UnAV-100. *Future Prediction F1-Score*: We also report the standard segment-level F1-score for predictions made for future time steps. The results are presented in the Table 5.

This results provides two key insights: First, the pseudo-future features are remarkably realistic. The Top-5 Similarity Accuracy of over 94% demonstrates that the correct event feature is almost always ranked among the top candidates. Second, these high-quality features enable strong future prediction performance, as evidenced by the solid F1-scores.

Table 5: The quantitative results of the pseudo-future's quality.

Metric	T+1	T+2	T+3	T+4	T+5
Top-1 Similarity	45.4	44.6	44.0	43.3	42.6
Top-5 Similarity	95.5	95.3	95.1	94.8	94.3
F1	57.5	56.5	55.4	54.7	54.1

Ablation study on hyperparams The ablation study on the loss weighting hyperparameter λ is shown in Table 6. The results indicate that performance diminishes at the tested extreme values ($\lambda = 0.1$ and $\lambda = 10$), while the model exhibits stable and strong performance across a moderate range (from $\lambda = 0.5$ to $\lambda = 2$). Therefore, we adopt $\lambda = 1$ as the default setting in our experiments for simplicity.

Table 6: Ablation studies for hyperparams, loss weight λ.

λ	Seg-Level		Event-Level	
	F1	mAP	0.5	Avg
0.1	58.3	70.8	46.9	45.8
0.5	58.9	70.2	47.2	46.2
1	59.1	70.1	46.9	46.3
1.5	58.4	69.9	47.2	46.2
2	59.2	70.0	47.5	46.6
5	55.9	68.9	44.3	43.7
10	13.2	50.8	9.4	9.5

Different feature extractors Table 7 presents the ablation study on different feature extractors, evaluating their impact on the performance and efficiency of On-AVEP task. The results clearly indicate that employing more powerful foundation models as feature extractors generally leads to significant improvements in parsing performance. Specifically, while the I3D [1]+VGGish [24] combination is relatively lightweight, its performance is comparatively limited. In contrast, AudioClip [20] and CLIP [46]+CLAP [8] offer a favorable balance between performance and computational efficiency. Although OnePeace [59] achieves the best parsing results, its substantial computational requirements may hinder its practical deployment in real-world scenarios. Notably, the computational complexity of our proposed PreFM module remain relatively stable and low across all tested feature extractors. This underscores that the feature extraction stage constitutes the primary performance bottleneck and source of computational load, directly impacting the system's online processing capabilities.

Table 7: Ablation studies for different feature extractors. a: audio extractor part, v: visual extractor part.

Methods	Seg-Level		Event-Level		Dimensions		FLOPS↓		
	F1	mAP	0.5	Avg	a	v	a	v	PreFM
I3D [1]+VGGish [24]	30.7	48.7	23.7	24.0	128	2048	0.9G	3.5G	0.3G
AudioClip [20]	48.0	63.6	37.7	37.0	1024	1024	2.7G	5.4G	0.1G
CLIP [46]+CLAP [8]	59.1	70.1	46.9	46.3	768	768	23.1G	77.8G	0.5G
ONE-PEACE [59]	62.4	70.6	52.2	51.5	1536	1536	78.8G	389.8G	0.4G

Failure cases The quantitative findings and analysis about failure cases are shown below:

Confusion Between Similar Events We analyze the events with the lowest performance and their most common confusions in Table 8. This results reveal that PreFM struggles to distinguish between events that are semantically or acoustically similar. We hypothesize this is because our current framework does not explicitly incorporate a contrastive learning design to better separate the representations of events originating from similar audio-visual sources.

Table 8: The quantitative analysis of confusion between similar events.

Event	Precision	Recall	F1	Most confused with
People slurping	0.55	0.17	0.25	People eating, man speaking
People shouting	0.42	0.19	0.27	Baby laughter, engine knocking

Performance in Dense Scenes We analyze the impact of event density (the number of event classes within a video) on event-level performance in Table 9. These results show that PreFM's performance degrade in complex videos containing a large number of distinct event classes. This suggests that while our future modeling is effective, its benefits are less pronounced in scenarios with very rapid scene changes and drastic context shifts.

Table 9: The quantitative analysis of PreFM in dense scenes.

Num events	Event-Level Avg
1-3	0.54
4-6	0.23
>6	0.13

A.3 Additional Details

The detailed process of modality-agnostic representation refinement (MRR) For the MRR process, we select OnePeace [59] as the pre-trained teacher model. The generation of target teacher features f_t at each time step t involves the following steps: First, ground-truth event labels y_t are converted into textual prompts using the template "*a/an audio/visual/audio-visual event of* `[cls]`". These prompts are then processed by the OnePeace text encoder to yield the modality-agnostic event features. If multiple events are active at time t, the final f_t is computed by averaging the features corresponding to all active event classes. Concurrently, the student model's representation $\hat{f}_t^{av}$ is prepared. We extract the audio features $\hat{f}_t^a$ and visual features $\hat{f}_t^v$ from our model at time t, specifically from the layer before the final classification head $h(\cdot)$. These extracted features are subsequently concatenated in feature dimension to form the student's representation: $\hat{f}_t^{av} = \texttt{Concat}(\hat{f}_t^a, \hat{f}_t^v)$.

Specific values for focal temporal prioritization As detailed in Eq. 9 and Eq. 10, our focal temporal prioritization are designed to emphasize predictions closer to the current time T while maintaining the overall loss scale. This scale preservation ensures that the sum of weights for the context window, $\sum_{t=T-L_c+1}^{T} w_c(t)$, equals L_c, and for the future window, $\sum_{t=T+1}^{T+L_f} w_f(t)$, equals L_f. Table 10 presents the specific numerical values of these weights for each relative time step, calculated with our default settings of $L_c = 10$ and $L_f = 5$.

Table 10: Specific values of the focal temporal prioritization for current time steps ($t \in [T-9, T]$) and future prediction time steps ($t \in [T+1, T+5]$).

Time step	Current										Future				
	T-9	T-8	T-7	T-6	T-5	T-4	T-3	T-2	T-1	T	T+1	T+2	T+3	T+4	T+5
Weight value	0.76	0.83	0.89	0.95	1.01	1.06	1.09	1.12	1.14	1.14	1.12	1.10	1.03	0.94	0.81

Difference between random segment sampling and normal sampling The details of our random segment sampling strategy are described in Sec. 3.6. In contrast, normal sampling strategy just involves dividing a video of total length T_{all} into a sequence of k non-overlapping chunks, each of length L_c. For such a normal approach, the target prediction time T_k for each k-th chunk is deterministically set to its final time step, specifically defined as $T_k = kL_c - 1$. This means that predictions are consistently targeted only at the very end of these fixed chunks, unlike the more varied and diverse target prediction times generated by our random segment sampling method.

Dataset modification for online setting For the **UnAV-100** dataset [13], while its original annotations specify continuous time segments for events (e.g.,$[cls, T_{start}, T_{end}]$), we convert these into frame-level discrete labels for our online task. Specifically, for any given time T and a particular event within a video stream, a label of 1 indicates the event is currently occurring, while 0 indicates it is not.

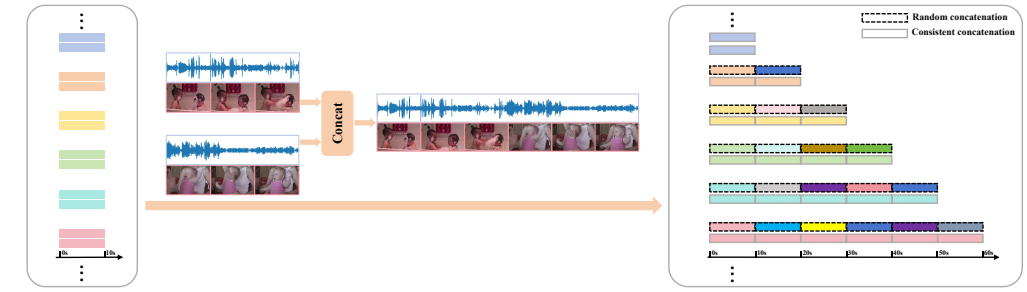

Figure 6: Visualization of the LLP dataset adaptation for the online setting. Various colored rectangles represent video clips of different categories. Video clips are processed by random concatenation and consistent concatenation.

The **LLP** dataset [54] initially provides 11,849 10-second video clips. To adapt it for our online evaluation setting, we concatenate 11,849 new, untrimmed video streams as shown in Figure 6. Each new stream is created by using the original 10-second clips as its base and concatenating additional clips. These resulting streams are specifically constructed to achieve one of six distinct target total durations: *10 seconds* (representing the original clip itself), *20 seconds, 30 seconds, 40 seconds, 50 seconds, or 60 seconds*. Approximately an equal number of streams are generated for each of these six target durations.

This concatenation process employs two distinct strategies: half of the 11,849 streams are formed by *random concatenation*, randomly combining clips from different categories. This aims to simulate the rapid and complex within-second scene variations commonly observed in current streaming content. The other half are constructed by *consistent concatenation*, identifying the first event category present in the base clip and then concatenating multiple additional clips that also contain this specific event, thereby simulating longer videos with a consistent, ongoing event context. This approach allows us to assess the model's adaptability to complex dynamic scenes and its capability for consistent understanding and discrimination within extended event contexts.

Regarding other datasets: The case of LFAV The LFAV dataset [25], comprising 5175 untrimmed videos with diverse audio, visual, and audio-visual events, is designed for long-form audio-visual video parsing and thus appears initially relevant for the On-AVEP task. However, we identify two critical limitations that preclude its effective use with our PreFM framework.

Firstly, complete access to the original video data is restricted. Of the officially stated 3721 training, 486 validation, and 968 test samples, our attempts allow us to retrieve only 3512, 461, and 910 samples respectively (totaling 4883 out of 5175 raw videos). The LFAV benchmark provides pre-extracted features using VGGish [24], ResNet18 [21], and R3D [55]. This reliance on fixed features prevents us from employing different feature extractors or leveraging pre-trained models (such as OnePeace [59] for our modality-agnostic robust representation) directly on the raw video data, which is a key aspect of our method.

Secondly, LFAV is curated under a weak supervision paradigm, offering only video-level annotations for its training set. The absence of readily available segment-level ground truth makes LFAV unsuitable for training critical components of our PreFM model. Specifically, mechanisms like our predictive future modeling and focal temporal prioritization require finer-grained temporal supervision than what LFAV's training annotations provide, rendering it incompatible with the training requirements for our online streaming prediction approach.

More online inference details All methods are evaluated using their officially provided checkpoints; for those without an available checkpoint, we reproduce the results using their official code. For any prediction at time T in online testing, only data from 0 to T is available.

For **On-AVEL** tasks, SOTA methods [13, 14, 78] pad the entire video beyond $T + 1$ with zeros as input because these methods originally utilize the complete video, and we use this padding to ensure a uniform input length under online settings. Our method, in contrast, does not use all available

historical data up to T; instead, it processes the segment $[T - L_c + 1, T]$ as input to derive the prediction at time T.

For **On-AVVP** tasks, SOTA approaches [29, 49, 76, 27, 71] employ the segment $[T - 9, T]$ as input, since these methods are designed for 10-second video clips. Similarly, our method utilizes the segment $[T - L_c + 1, T]$ as input for making predictions at time T.

Re-evaluation of efficiency metrics for fair comparison Table 11 (for the On-AVEL task) and Table 12 (for the On-AVVP task) present side-by-side comparisons of efficiency metrics. These include figures from our standardized re-evaluation (denoted as "Our Eval.") and those originally published in the respective papers (denoted as "Reported").

For our evaluations, we adhere to a strict and consistent protocol. The number of trainable parameters for all models is calculated as the sum of elements in all parameters requiring gradients (using `sum(p.numel() for p in model.parameters() if p.requires_grad)`). To measure FLOPs, we consistently employ the `thop` library for all methods, assessing a single forward pass (via `flops, _ = profile(model, inputs=(input,))`). All our efficiency tests are conducted under identical environmental conditions to ensure reproducibility and a fair basis for comparison.

Discrepancies may be observed between our "Our Eval." figures and the "Reported" values from the original publications. Such differences can arise from variations in measurement methodologies, specific versions of libraries used, or the underlying hardware and software environments. We present both sets of values to offer a transparent perspective, respecting the data from original publications while providing a benchmark that is directly comparable across methods under our unified testing framework.

Table 11: Comparison of re-evaluated ("Our Eval.") and originally reported ("Reported") efficiency metrics on the On-AVEL task. ("-" indicates the metric was not provided in the original paper.

Methods	Params		FLOPs	
	Our Eval.	Reported	Our Eval.	Reported
UnAV [13] (CVPR2023)	139.4M	-	52.4G	-
UniAV [14] (Arxiv2404)	130.8M	130M	22.7G	-
CCNet [78] (AAAI2025)	238.8M	-	72.1G	-
PreFM	12.3M	none	0.4G	none
PreFM+	36.9M	none	0.5G	none

Table 12: Comparison of re-evaluated ("Our Eval.") and originally reported ("Reported") efficiency metrics on the On-AVVP task. "-" indicates the metric was not provided in the original paper.

Methods	Params		FLOPs	
	Our Eval.	Reported	Our Eval.	Reported
VALOR [29] (NeurIPS2023)	4.9M	5.1M	0.45G	0.45G
Coleaf [49] (ECCV2024)	5.7M	-	0.25G	48.2G
LEAP [76] (ECCV2024)	52.0M	52.0M	1.09G	0.79G
NREP [27] (TNNLS2024)	9.6M	9.6M	1.69G	0.37G
MM-CSE [71] (AAAI2025)	6.2M	4.5M	0.91G	0.80G
PreFM	3.3M	none	0.22G	none
PreFM+	12.1M	none	0.48G	none

A.4 More Qualitative Analysis

On-AVEL Figure 7 provides further qualitative validation of our method's on the On-AVEL task through four distinct examples, comparing our results ("Ours") against the state-of-the-art CCNet [78] model ("SOTA"). These visualizations collectively demonstrate that our approach consistently yields event localizations that are more aligned with the ground truth annotations compared to CCNet.

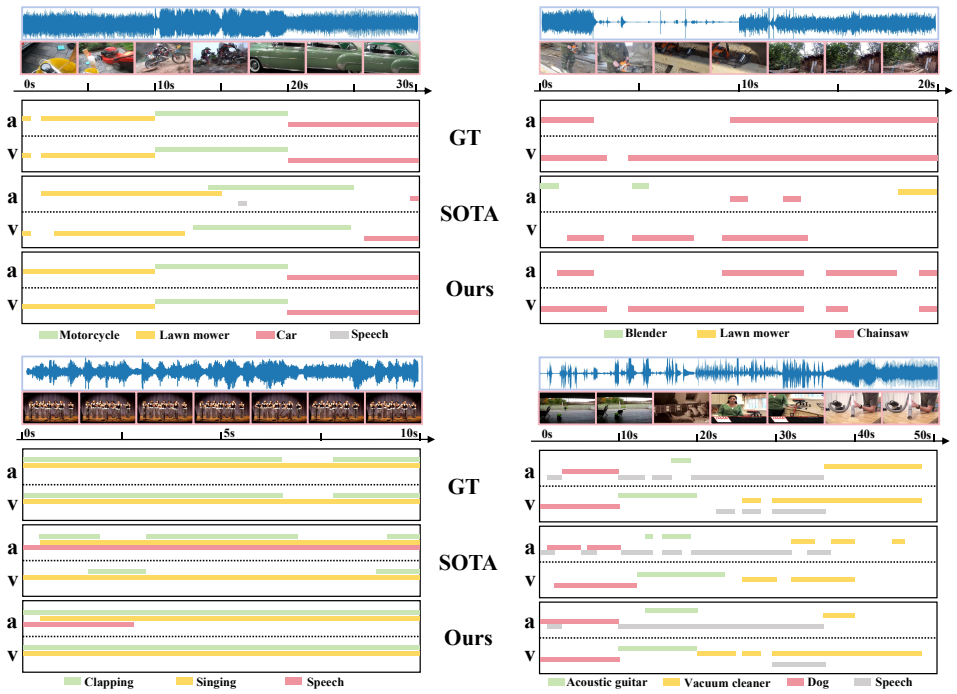

Figure 8: More visualization results on the On-AVVP task.

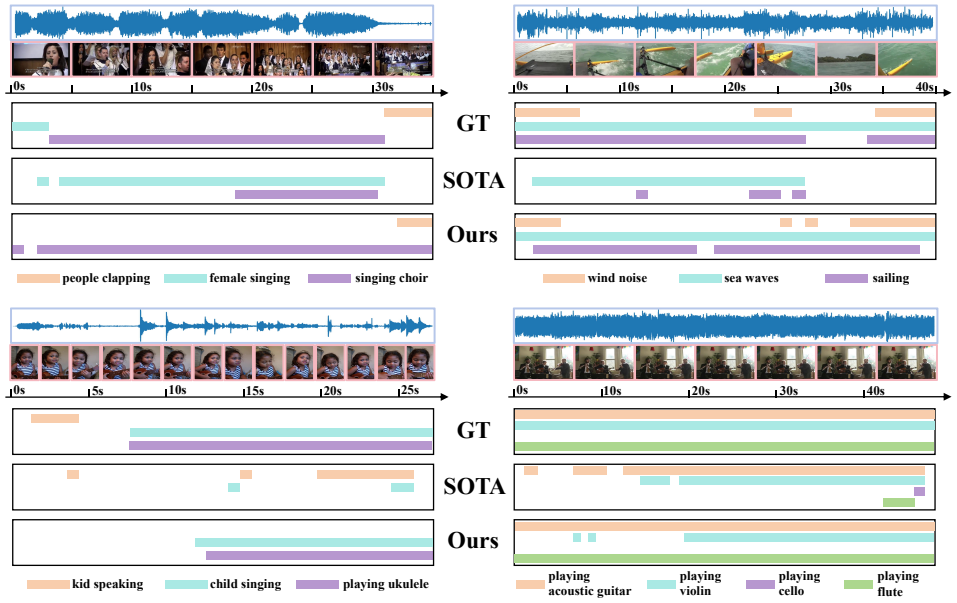

Figure 7: More visualization results on the On-AVEL task.

On-AVVP Similarly, we provide further qualitative results of our method's on the On-AVVP task through four distinct examples, comparing our results ("Ours") against the state-of-the-art MM-CSE [71] model ("SOTA"). The visualization results are shown in Figure 8. These visualizations highlight our method's superior performance in precisely parsing events and reducing errors compared to the SOTA model.

This robust handling of both unimodal and multimodal event characteristics signifies a key advantage of our approach for the online audio-visual event parsing task.