# TeMTG: Text-Enhanced Multi-Hop Temporal Graph Modeling for Audio-Visual Video Parsing


Yaru Chen
University of Surrey
Surrey, United Kingdom
Aria_yc@126.com

Peiliang Zhang
Wuhan University of Technology
Wuhan, China
cheungbl@ieee.org

Fei Li
University of Wisconsin-Madison
Madison, United States
leefly072@126.com

Faegheh Sardari
University of Surrey
Surrey, United Kingdom
f.sardari@surrey.ac.uk

Ruohao Guo
Peking University
Beijing, China
ruohguo@stu.pku.edu.cn

Zhenbo Li
China Agricultural University
Beijing, China
lizb@cau.edu.cn

Wenwu Wang
University of Surrey
Surrey, United Kingdom
w.wang@surrey.ac.uk



## Abstract

Audio-Visual Video Parsing (AVVP) task aims to parse the event categories and occurrence times from audio and visual modalities in a given video. Existing methods usually focus on implicitly modeling audio and visual features through weak labels, without mining semantic relationships for different modalities and explicit modeling of event temporal dependencies. This makes it difficult for the model to accurately parse event information for each segment under weak supervision, especially when high similarity between segmental modal features leads to ambiguous event boundaries. Hence, we propose a multimodal optimization framework, TeMTG, that combines text enhancement and multi-hop temporal graph modeling. Specifically, we leverage pre-trained multimodal models to generate modality-specific text embeddings, and fuse them with audio-visual features to enhance the semantic representation of these features. In addition, we introduce a multi-hop temporal graph neural network, which explicitly models the local temporal relationships between segments, capturing the temporal continuity of both short-term and long-range events. Experimental results demonstrate that our proposed method achieves state-of-the-art (SOTA) performance in multiple key indicators in the LLP dataset.


## CCS Concepts

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision tasks**;


Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*ICMR'25, Chicago, USA*



ACM ISBN 979-8-4007-1877-9/25/06
https://doi.org/10.1145/3731715.3733495

## Keywords

Audio-Visual Video Parsing, Semantic Enhancement, Multi-hop Temporal Graph, Weakly Supervised Learning

**ACM Reference Format:**
Yaru Chen, Peiliang Zhang, Fei Li, Faegheh Sardari, Ruohao Guo, Zhenbo Li, and Wenwu Wang. 2025. TeMTG: Text-Enhanced Multi-Hop Temporal Graph Modeling for Audio-Visual Video Parsing. In *Proceedings of The 15th ACM International Conference on Multimedia Retrieval (ICMR'25).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3731715.3733495


## 1 Introduction

In an Audio-Visual Video Parsing (AVVP) task [1], our goal is not only to detect what events occur at what times, but also to determine which modality detects the event. AVVP techniques can be used in a variety of real-world applications, such as intelligent video surveillance, and content-based video indexing. Compared to other related tasks [2–5], a distinguishing characteristic of this task is the temporal asynchrony between events that occur in different modalities. As shown in Fig. 1 (a), we see cats for 10 seconds, while we hear it between 1-3 and 8-10 s, and we can still hear the sound of speech even if no one appears in the video. Therefore, events are often categorized into three types: audio events, visual events, and audio-visual events. An AVVP model is often trained using weakly labelled data where event labels are given only for the whole video, instead of its individual frames. This setting increases the difficulty for models to learn the temporal details and modality correlations.

A baseline method [1] was developed for the AVVP task by employing hybrid attention networks (HAN), where multi-modal multiple instance learning (MMIL) is used to aggregate the multi-modal temporal contexts, together with the identification and suppression of noisy labels for each modality. Subsequently, the researchers [6, 7] explored contrastive learning, distillation learning, and other techniques to connect semantically similar segments within and between modalities. With the emergence of large-scale pre-trained models, Lai et al. [8] utilized pre-trained CLAP [9] and CLIP [10] to extract features and generate segment-level pseudo-labels. Other researchers have explored the use of pseudo labels as

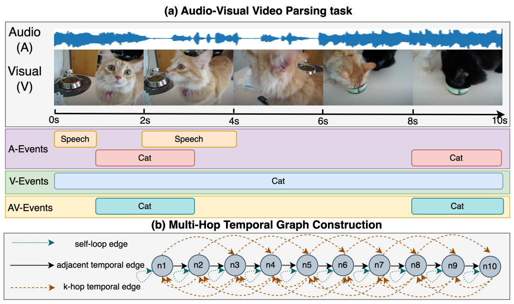


**Figure 1: (a) Illustration of the AVVP task. (b) Structure for Multi-Hop Temporal Graph (assume $K = 2$).**

references for the model to distinguish the semantics of each event that occurred in each segment [11, 12].

Previous studies have explored the connections between the categories of events [13, 14], but not the propagation and continuation of events in the temporal dimension. In addition, large-scale pre-trained models have been used to provide semantic information [15, 16], which is used as pseudo labels and classification auxiliary. However, they have not been integrated deeply into feature representation. As a result, semantic consistency within the features cannot be guaranteed, causing potential misalignment in audio-visual feature fusion. Recently, text embeddings have been used to improve multimodal representation learning [17]. This method focuses primarily on encoding event-related semantics while neglecting background information, which, however, may contain crucial contextual cues to distinguish events.

To tackle these challenges, we propose TeMTG, which combines text-enhanced semantic guidance with multi-hop temporal graph (MTG) modeling for weakly supervised AVVP. We first introduce a fusion mechanism that leverages large-scale pre-trained models to generate segment-level pseudo labels and corresponding text embeddings, which are refined by a modality-specific multi-layer perceptron (MLP) [18] to enhance semantic guidance and feature discriminability. Then, we construct a K-hop temporal graph to explicitly model segment-wise dependencies. By linking segments through K-hop edges, our model captures audio-visual correlations over time, improving temporal reasoning. Experimental results show that TeMTG effectively addresses AVVP limitations and achieves SOTA performance.

## 2 Proposed Methodology

In the AVVP task, a video clip $S$ can be divided into $T$ segments, represented as $S = \{A_t, V_t\}_{t=1}^{T}$, where $A_t$ and $V_t$ denote the audio and visual features of the $t$-th segment. The task requires segmenting events into three categories: audio events $y_t^a \in \{0,1\}^C$, visual events $y_t^v \in \{0,1\}^C$, and audio-visual events $y_t^{av} \in \{0,1\}^C$, where $C$ is the total number of event classes. An event is considered audio-visual if it appears in both modalities simultaneously, i.e., $y_t^{av} = y_t^a * y_t^v$. During training, only weak video-level labels $y \in \{0,1\}^C$ are provided. The goal of AVVP is, given a video clip divided into $T$ segments, to determine for each segment whether an event (among $C$ possible classes) occurs in the audio modality, the visual modality, or both.

### 2.1 Framework

We adopt CoLeaF [7] as our baseline model and propose a novel multimodal optimization framework that integrates text enhancement and multi-hop temporal graph modeling. As shown in Fig. 2, we first use the text encoder to generate text embeddings for the audio and visual stream of each video segment, respectively. These embeddings are then fused with the audio and visual features, respectively, through a feature fusion module. Next, the fused multimodal features are fed into the feature aggregation module which adopts self-attention and cross-attention to preserve unimodal feature learning and enhance cross-modal interactions. Then we construct a multi-hop temporal graph and propagate information using multi-head graph attention (GAT) [19] to model both short-term continuity and long-term dependencies among video segments. Finally, we employ MMIL pooling [1] to aggregate temporal features and generate the final video-level predictions, including audio predictions $P_a$, visual predictions $P_v$, and joint audio-visual predictions $P$. As CoLeaF used two branches for feature aggregation, we placed our proposed multi-hop temporal GAT after each branch.

### 2.2 Text-Enhanced Multimodal Feature Fusion

To enhance the semantic representation of the audio-visual features, we introduce text embeddings to provide explicit semantic guidance, effectively mitigating the limitations of weakly supervised learning. Specifically, inspired by [8], we first use CLAP and CLIP to generate pseudo labels at segment level $p_t^a, p_t^v \in \{0,1\}^C$. Then, we convert each pseudo label into a text description in the following format: "*This is the sound of* x *audio event*" or "*This is the image of* x *visual event*". If a segment contains multiple audible or visible events, we concatenate their corresponding descriptions with conjunctions (e.g. "*This is the sound of event A and event B*"). If a segment contains no events, the corresponding text is set as "*There is no sound in the segment*" or "*There is no event in the image*". Then, we feed these texts into the text branches of CLAP and CLIP to generate the corresponding text embeddings $e_t^a, e_t^v \in \mathbb{R}^{b \times T \times d}$, in which $b$ is the batch size, and $d$ is the feature dimension.

Afterwards, we design a modality-specific fusion strategy based on MLP [18] to effectively integrate semantic text information into audio-visual feature representations. We first get the audio and visual feature representations $f_t^a, f_t^v \in \mathbb{R}^{b \times T \times d}$ from their feature extractors, and then concatenate the features and their text embeddings along the feature dimension to obtain the fused input:

$$z_t^a = (f_t^a \parallel e_t^a) \in \mathbb{R}^{b \times T \times 2d} \tag{1}$$

where $\parallel$ is the concatenating operation. This feature is then mapped to the fused audio feature through a two-layer MLP:

$$f_t^{ea} = \sigma\left(W_2\left(ReLU\left(W_1 z_t^a + b_1\right)\right) + b_2\right) \tag{2}$$

where $W_1 \in \mathbb{R}^{2d \times m}$ and $W_2 \in \mathbb{R}^{d \times m}$, are the weight matrices of the two-layer MLP, $b_1$ and $b_2$ are bias terms, $m$ is the hidden layer dimension, and $\sigma(\cdot)$ denotes the LayerNorm operation. Similarly, the fusion process for the visual modality is defined as follows:

$$z_t^v = (f_t^v \parallel e_t^v) \in \mathbb{R}^{bs \times T \times 2d} \tag{3}$$

$$f_t^{ev} = \sigma\left(W_2\left(ReLU\left(W_1 z_t^a + b_1\right)\right) + b_2\right) \tag{4}$$

Finally, we apply a linear layer to project the fused audio and visual features back to the original feature dimension $d$, ensuring compatibility with the input of the downstream task and maintaining consistency with the original feature space after fusion.

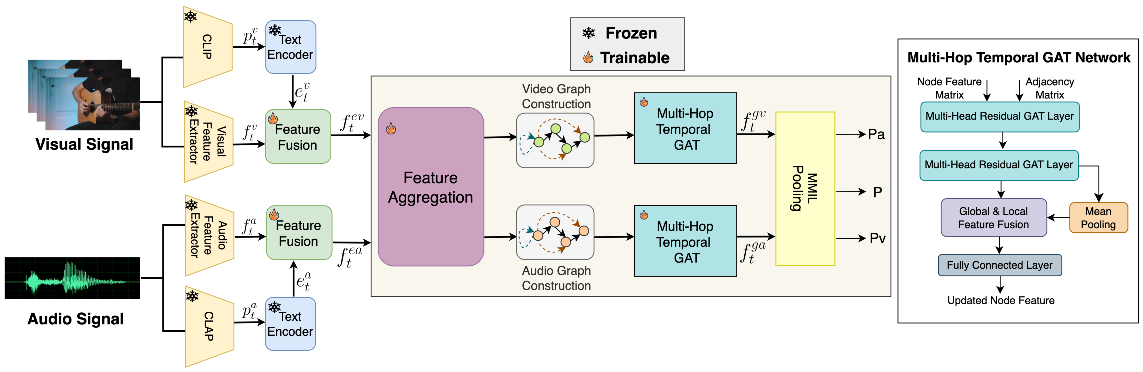


**Figure 2: TeMTG architecture: Audio and visual features are fused with text embeddings, aggregated, and then processed by multi-hop temporal graphs with multi-head residual GAT layers to model event dependencies.**

## 2.3 Multi-Hop Temporal Graph Modeling

To model both short- and long-term dependencies, we propose an MTG, which links each segment not only to its neighbors but also to others within a temporal range $K$. This bidirectional, modality-specific design enables flexible and effective temporal modeling by allowing each node to aggregate contextual information from both past and future segments, enhancing temporal relation reasoning.

Specifically, for the input features $f_t^{ea}$ and $f_t^{ev}$, we construct the temporal graphs $G^a = (N^a, E^a)$ and $G^v = (N^v, E^v)$ for each video at the segment level, where the nodes $N^a = \left\{n_1^a, n_2^a, \cdots, n_T^a\right\}, n_t^a \in \mathbb{R}^d$ and $N^v = \left\{n_1^v, n_2^v, \cdots, n_T^v\right\}, n_t^v \in \mathbb{R}^d$ represent the audio and visual features of each segment, and the edges $E^a$ and $E^v$ show the temporal relationships between the segments. Hence, for each audio or visual node $n_t$, we first define their K-Hop bidirectional temporal connection edges as follows:

$$\begin{aligned} E^a = & \left\{ (n_t, n_{t-k}) \mid 1 \le t \le T,\ 1 \le k \le K,\ t-k \ge 1 \right\} \\ & \cup \left\{ (n_t, n_{t+k}) \mid 1 \le t \le T,\ 1 \le k \le K,\ t+k \le T \right\}. \end{aligned} \tag{5}$$

where $k$ is the hop distance. Additionally, each node has a self-loop to preserve its original information. The final adjacency matrix for audio and visual temporal graph $A^a$ and $A^v$ are:

$$A_{ij}^a, A_{ij}^v = \begin{cases} 1, & if 0 \le j-i \le K \\ 0, & otherwise \end{cases} \tag{6}$$

where $i$ and $j$ are the index of the nodes.

Then we employ the multi-head residual GAT to perform feature aggregation on the constructed temporal graph, allowing nodes to dynamically weight different time steps based on contextual information. Specifically, given the multi-hop temporal graph $G$ and its adjacency matrix $A$, the attention weight between two connected nodes $n_i$ and $n_j$ is computed as follows:

$$\alpha_{ij} = \frac{exp\left(LeakyReLU\left(a^{(h)T}\left[W^{(h)}n_i \parallel W^{(h)}n_j\right]\right)\right)}{\sum_{k\in\phi_i} exp\left(LeakyReLU\left(a^{(h)T}\left[W^{(h)}n_i \parallel W^{(h)}n_k\right]\right)\right)} \tag{7}$$

where $\cdot^T$ represents transposition, $h$ is the multi-head attention, $W \in \mathbb{R}^{d\times d}$ is a learnable weight matrix, $a \in \mathbb{R}^{2d\times 1}$ is the attention vector, and $\phi_i$ is the set of neighbors of the node $i$. The final update of the node features is as follows:

$$n_i' = \varepsilon\left(\frac{1}{H}\sum_{h=1}^{H}\sum_{j\in\phi_i}\alpha_{ij}^{(h)} \cdot W^{(h)}n_j\right) \tag{8}$$

where $\varepsilon(\cdot)$ is a nonlinear activation function, and $H$ is number of heads for multi-head attention. To enhance model stability, we incorporate residual connections, batch normalization, and dropout in the design of K-hop GAT layer, further optimizing the effectiveness of temporal information propagation.

After GAT propagation, we use global mean pooling to extract the temporal representation of the entire video, enhancing cross-node information integration. Subsequently, we use an MLP to further fuse the global video features with each feature node, enabling each node to model local temporal relationships while also perceiving the global context of the entire video, which can effectively model both short-term and long-term event dependencies.

# 3 Experimental Results

## 3.1 Experimental Setup

**Dataset and Implementation Details.** We use the LLP dataset [1] to evaluate our framework, which includes 11849 videos with 25 categories and has been widely used in the AVVP task. Each video is divided into 10 segments and each segment lasts 1 second. We utilize the pre-trained CLAP [9] to extract 768-D audio features from the audio signal. We use pre-trained CLIP [10] and 3D ResNet to extract 768-D and 512-D visual features from the visual signal, then fuse the concatenated 2D and 3D visual features. Finally, a linear layer is used to project audio and visual features into the same feature space to facilitate subsequent operations. We set the number of hops $K = 4$ for both audio and visual temporal graphs to ensure a balanced temporal dependency modeling across modalities. In addition, we performed our experiments using PyTorch on an NVIDIA A100 GPU.

**Evaluation Metrics** Following [1, 6], we use F1-score to evaluate audio **(A)**, visual **(V)**, and audio-visual **(AV)** events, with mIoU $\geqq$ 0.5 as the threshold. F1-scores are computed at both segment and event levels: the former compares predictions per segment, while the latter considers sequences of segments as complete events. **Type@AV** averages F1-scores across A, V, and AV, and **Event@AV** jointly evaluates all events in a video.

## 3.2 Overall Performance Analysis

Table 1 shows the experimental results of the comparison between our method and the existing SOTA methods in the LLP dataset. From the results, it can be seen that TeMTG has achieved the best

**Table 1: The performance of TeMTG and comparative methods in AVVP, with the best results highlighted in bold and the second results highlighted in text.**

| Model | Venue | Segment-level (%) | | | | | Event-level (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | V | AV | Type@AV | Event@AV | A | V | AV | Type@AV | Event@AV |
| HAN [1] | ECCV'20 | 60.1 | 52.9 | 48.9 | 54.0 | 55.4 | 51.3 | 48.9 | 43.0 | 47.7 | 48.0 |
| MGN [13] | NeurIPS'22 | 60.8 | 55.4 | 50.0 | 55.1 | 57.6 | 52.7 | 51.8 | 44.4 | 49.9 | 50.0 |
| MA [20] | CVPR'21 | 60.3 | 60.0 | 55.1 | 58.9 | 57.9 | 53.6 | 56.4 | 49.0 | 53.0 | 50.6 |
| CMPAE [6] | CVPR'23 | 64.2 | 66.2 | 59.2 | 63.3 | 62.8 | 56.6 | 63.7 | 51.8 | 57.4 | 55.7 |
| CoLeaF [7] | ECCV'24 | 64.2 | 67.1 | 59.8 | 63.8 | 61.9 | 57.1 | 64.8 | 52.8 | 58.2 | 55.5 |
| LEAP [11] | ECCV'24 | 64.8 | 67.7 | 61.8 | 64.8 | 63.6 | 59.2 | 64.9 | <u>56.5</u> | 60.2 | 57.4 |
| VALOR++ [8] | NeurIPS'23 | 68.1 | 68.4 | 61.9 | 66.2 | 66.8 | 61.2 | 64.7 | 55.5 | 60.4 | 59.0 |
| LSLD+ [12] | NuerIPS'23 | 68.7 | <u>71.3</u> | <u>63.4</u> | 67.8 | 68.2 | 61.5 | <u>67.4</u> | 55.9 | <u>61.6</u> | 60.6 |
| NREP [16] | TNNLS'24 | <u>70.2</u> | 70.9 | **64.4** | <u>68.5</u> | <u>68.8</u> | **62.8** | 67.3 | **57.6** | **62.6** | <u>61.1</u> |
| TeMTG (Ours) | - | **74.4** (+4.2) | **72.9** (+1.6) | 62.0 | **69.8** (+1.3) | **74.1** (+5.3) | <u>61.9</u> | **69.0** (+1.6) | 53.2 | 61.4 | **62.2** (+1.1) |

**Table 2: Ablation study for TeMTG. w/o TE and w/o MTG mean without TE and MTG respectively.**

| | Method | A | V | AV | Type@AV | Event@AV |
|---|---|---|---|---|---|---|
| Segment-level | CoLeaF$^\dagger$ | 64.2 | 67.4 | 59.9 | 63.8 | 63.3 |
| | *w/o* TE | 64.8 | 68.9 | 60.6 | 64.8 | 64.2 |
| | *w/o* MTG | 76.5 | 72.9 | 62.4 | 70.6 | 75.7 |
| | TeMTG | 74.4 | 72.9 | 62.0 | 69.8 | 74.1 |
| | Method | A | V | AV | Type@AV | Event@AV |
| Event-level | CoLeaF$^\dagger$ | 53.2 | 64.1 | 52.4 | 56.6 | 52.7 |
| | *w/o* TE | 53.5 | 65.6 | 52.4 | 57.1 | 53.3 |
| | *w/o* MTG | 66.6 | 69.0 | 53.4 | 63.1 | 66.0 |
| | TeMTG | 61.9 | 69.0 | 53.2 | 61.4 | 62.2 |

performance in multiple key indicators, especially in segment-level parsing, which significantly surpasses existing methods.

In segment-level evaluation, TeMTG achieved the best results for both audio (A) and visual (V) event parsing, outperforming NREP by 4.2% and 1.6%, respectively, highlighting the effectiveness of text-enhanced feature fusion. For Event@AV, TeMTG exceeded NREP by 5.3%, showing that our multi-hop temporal graph better captures cross-segment temporal dependencies. However, for AV event parsing, TeMTG scored 62.0%, lower than NREP (64.4%), likely due to challenges in modeling audio-visual co-occurrence under weak supervision, despite textual enhancements.

In event-level evaluation, TeMTG achieved 69.0% in visual event (V) parsing, 1.6% higher than LSLD+, showing better capture of visual features. TeMTG achieved the highest Event@AV score (62.2%), demonstrating strong overall parsing ability under weak supervision. However, for audio event (A) parsing, TeMTG scored 61.9%, slightly below NREP (62.8%), likely due to the lack of semantic constraints in our temporal graph model, which may leave residual noise from background sounds. For AV event parsing, TeMTG yielded a lower score (53.2%) compared to others, as pseudo labels from CLAP and CLIP still carry noise and uncertainty, affecting the accuracy in detecting the AV event boundaries.

## 3.3 Ablation Experiment Analysis

To show the effectiveness of text enhancement (**TE**) and multi-hop temporal graph modeling module (**MTG**), we performed ablation experiments by removing these two modules from TeMTG. For a fair comparison with CoLeaF, which was based on the audio and visual features extracted by VGGish and ResNet, we first train CoLeaF using the same input features as TeMTG, namely CoLeaF$^\dagger$.

As shown in Table 2, when only using **MTG** module at the segment level, compared to CoLeaF†, the performance for detecting audio events (A) has increased from 64.2% to 64.8%, visual events (V) increased from 67.4% to 68.9%, and Event@AV increased from 63.3% to 64.2%. At the event level, audio events (A) increased by 1.3% and visual events (V) increased by 1.5%. This indicates that temporal modeling can improve the ability to integrate local temporal information and improve single-modal feature analysis.

With only the TE mechanism enabled, the model shows a significant improvement in unimodal event parsing, especially for audio, with over 10% gains at both segment and event levels. It also achieves a 2.5% gain in AV event parsing, indicating that text enhancement benefits cross-modal parsing. Furthermore, Event@AV improves by more than 10%, suggesting that fusing modal features with text embeddings enriches information of the event category, helps the model learn temporal consistency, and reduces event segmentation errors.

Some TeMTG indicators are slightly lower than using only the TE mechanism, possibly due to temporal modeling's smoothing effect, reducing the discriminative power of text enhancement. Multi-hop temporal aggregation may spread features across segments, leading potentially to event overlap and reduced feature distinctiveness.

# 4 Conclusion

We have presented TeMTG, a multimodal framework combining text enhancement and multi-hop temporal graph modeling for weakly supervised AVVP. Text enhancement improves event classification between similar segments, while temporal modeling improves reasoning across time. Experiments on the LLP dataset show that TeMTG achieves SOTA performance on multiple metrics. However, smoothing effects from temporal modeling remain a limitation, which we plan to explore further in the future.


# Acknowledgements

This work was partially supported by a research scholarship from the China Scholarship Council (CSC) and a studentship from University of Surrey. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

# References


[1] 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.

[2] Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Yi Yu. Anchor-aware Deep Metric Learning for Audio-visual Retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 211–219, 2024.

[3] 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.

[4] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-Vocabulary Audio-Visual Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7533–7541, 2024.

[5] Fei Li, Lingfeng Shen, Yang Mi, and Zhenbo Li. Drcnet: Dynamic image restoration contrastive network. In *European Conference on Computer Vision*, pages 514–532. Springer, 2022.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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*, pages 1–22, 2024.

[16] 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.

[17] Langyu Wang, Bingke Zhu, Yingying Chen, and Jinqiao Wang. LINK: Adaptive Modality Interaction for Audio-Visual Video Parsing. *arXiv preprint arXiv:2412.20872*, 2024.

[18] Hind Taud and Jean-Franccois Mas. Multilayer perceptron (MLP). In *Geomatic approaches for modeling land change scenarios*, pages 451–455. Springer, 2017.

[19] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, 2018.

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