ESDN: Event-Centric Semantic Differential Denoising for Long-Form Audio-Visual Video Understanding

1st Longfei Zhang
School of Computer Science and Technology
Xinjiang University
Urumqi, China
zhanglongfei@stu.xju.edu.cn

2nd Lixu Sun
School of Computer Science and Technology
Xinjiang University
Urumqi, China
sunlixu@stu.xju.edu.cn

3rd Mengen Huang
School of Computer Science and Technology
Xinjiang University
Urumqi, China
107552304065@stu.xju.edu.cn

4th Nurmemet Yolwas†
School of Computer Science and Technology
Xinjiang University
Urumqi, China
nurmemet@xju.edu.cn

***Abstract*—Multisensory temporal event localization aims to precisely localize modality-aware events and their temporal boundaries in long-form videos. However, dense event overlaps in long videos cause relevant semantics and irrelevant semantic noise to intertwine during cross-modal interaction, leading to severe semantic entanglement. Consequently, existing methods frequently suffer from missed detections and insufficient recognition precision. To address this, the Event-Centric Semantic Differential Denoising Network is proposed. Specifically, the Semantic Differential Pyramid Multimodal Transformer utilizes Cross-modal Semantic Differential Attention to reconstruct cross-modal interactions for enhancing relevant semantics fusion; its driven pyramid strategy ensures a balance between noise suppression and temporal dependency modeling for events of varying durations. Multi-Granularity Semantic Loss Guidance mitigates label noise via fine-grained supervision and utilizes an event-aware joint loss to optimize event relation modeling. Experiments demonstrate significant performance improvements.**

***Index Terms*—Audio-Visual, Long Video, Cross-modal Semantic Differential Attention.**

I. Introduction

Long videos, carrying rich event semantics and relational information, serve as a key medium for machine understanding of natural scenes. While various long-form video understanding benchmarks have emerged recently [1]–[7], most predominantly focus on the visual modality while overlooking audio, thereby limiting comprehensive understanding. In the audio-visual domain, Tian et al. [8] introduced the AVVP task for short videos, aiming to recognize both the categories and temporal boundaries of audio-visual events. Subsequent research has achieved significant performance gains through architectural advancements [8]–[14] and enhanced weak-supervision strategies [15]–[18]. However, these methods struggle to generalize effectively to long videos, owing to their complex semantic environments and long-range event dependencies.

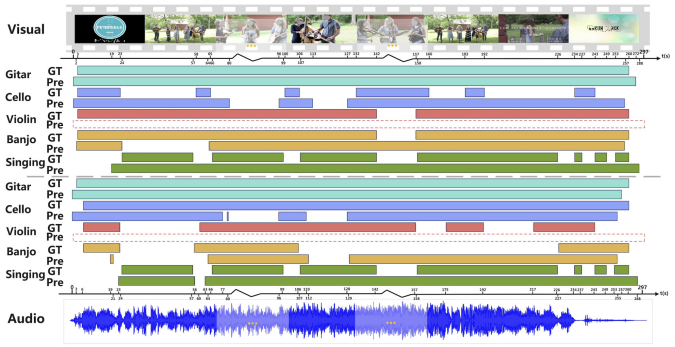

Fig. 1. Illustration of Multisensory Temporal Event Localization. Pre and GT denote predictions of the ECF model and ground truth, respectively. Ashed boxes mark undetected Violin events.

To address this gap, Hou et al. [19] proposed the Multisensory Temporal Event Localization task and the Event-Centric Framework (ECF). This framework employs a three-stage architecture utilizing a Pyramid Multimodal Transformer (PMT) with offsets, event-aware graphs, and event interaction modules to respectively capture multi-scale dependencies, refine features, and model inter-event relations. However, dense event overlaps in long videos cause relevant event semantics and irrelevant semantic noise to coexist, leading to severe semantic entanglement. Consequently, existing methods struggle to effectively identify relevant semantic information in such multi-event environments, frequently resulting in missed detections and insufficient recognition precision (e.g., the missed detection of Violin events and low recognition precision for Cello events shown in Fig 1). This exposes current technical

†denotes the Corresponding author. This work was supported by the National Key Research and Development Program of China under Grant No. 2023B01005.

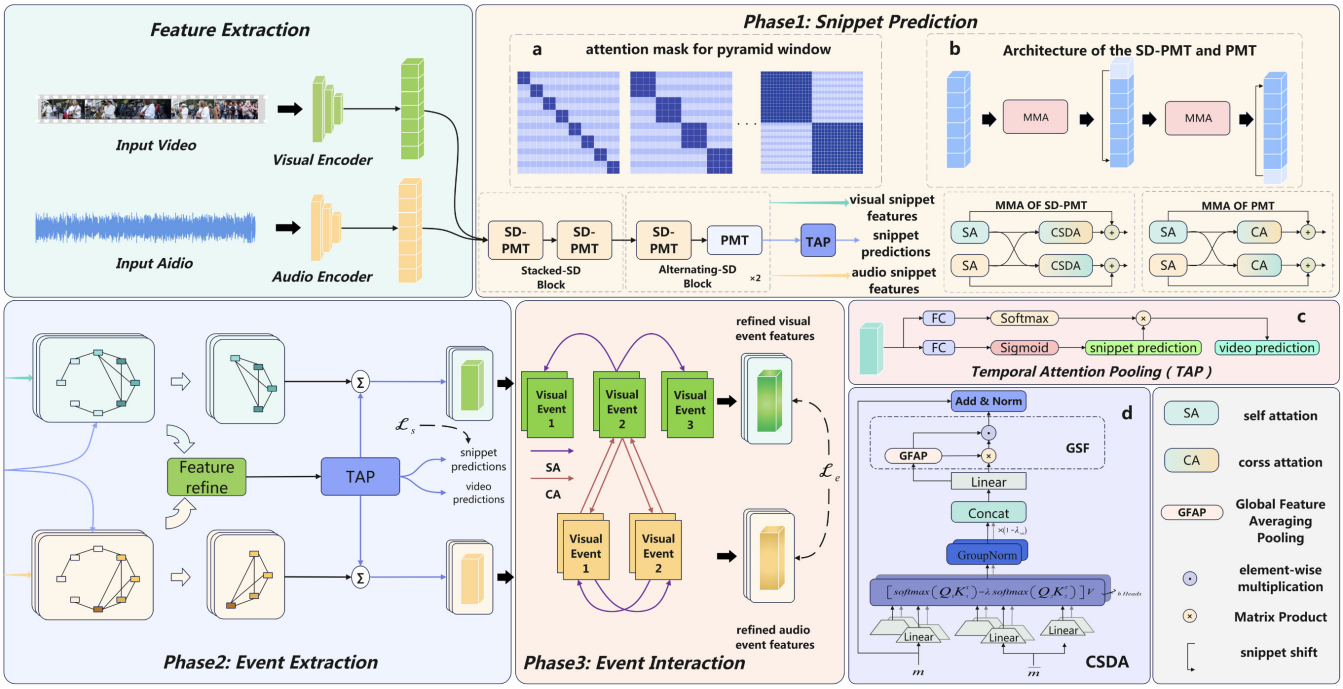

Fig. 2. The framework of ESDN. (a) Diagonal mask across layers. (b) snippet shift mechanism. (c) Temporal attention pooling. (d) CSDA, where m and $\overline{m}$ are modalities features. $\mathcal{L}_s$ denotes segment-level semantic guided loss and $\mathcal{L}_e$ denotes event-aware joint loss.

bottlenecks and highlights the urgent need to enhance semantic disentanglement and noise suppression capabilities.

To address the aforementioned challenges, the Event-Centric Semantic Differential Denoising Network (ESDN) is proposed. Specifically, the Semantic Differential Pyramid Multimodal Transformer (SD-PMT) reconstructs multimodal interaction via Cross-modal Semantic Differential Attention (CSDA) to effectively suppress noise and enhance event-relevant semantics. Addressing the phenomenon of significant variance in event durations in long videos, an SD-PMT-driven pyramid hierarchy strategy is introduced to effectively suppress noise for short-term events while achieving a balance between denoising and temporal dependency modeling for long-term events. Furthermore, the Multi-Granularity Semantic Loss Guidance (MSG) mechanism is incorporated to mitigate video-level label noise through pseudo-label-based segment-level supervision, and leverage an event-aware joint loss to enhance the modeling of event relations.

II. PROPOSED METHOD

A. Problem Definition

The Multisensory Temporal Event Localization task generates segment-level predictions via multi-instance learning and aggregates consecutive segments to produce event-level predictions, aiming to identify modality-specific events in long videos.Specifically, given a video of length T seconds, it is split into audio segments $\{A_t\}_{t=1}^{T}$ and visual segments $\{V_t\}_{t=1}^{T}$, generating segment-level predictions $\{p_t^a\}_{t=1}^{T}$ and $\{p_t^v\}_{t=1}^{T}$, where $p_t^a, p_t^v \in \mathbb{R}^C$ denote audio and visual predictions for the t-th segment, and C is the number of event categories. Audio-visual event labels are obtained by element-wise multiplication of the two modality labels. During training, only video-level event labels $y^m \in \{0,1\}^C, m \in \{a,v\}$ are accessible.

B. SD-PMT and Its Driven Pyramid Hierarchical Strategy

To address the pervasive noise induced by highly entangled event semantics in long-form videos, the SD-PMT is proposed. This architecture integrates CSDA to facilitate purified multimodal interaction, while utilizing a driven pyramid hierarchical strategy to adaptively reconcile noise suppression with temporal dependency modeling across diverse scales.

Cross-Modal Semantic Differential Attention. Addressing the challenge of highly entangled relevant semantics and irrelevant noise in long videos, a cross-modal semantic differential denoising paradigm is proposed, inspired by the differential attention mechanism [20]. Specifically, CSDA functionally constructs a mechanism analogous to a differential amplifier by computing the difference between two independent attention maps. This operation actively cancels out redundant semantic noise in cross-modal interactions, forcing the model to focus on sparse and highly discriminative cross-modal correlations. Furthermore, CSDA introduces a Global Semantic Fusion mechanism to align local features with global contextual information, thereby ensuring long-term temporal semantic consistency.

Following this paradigm, the computation of CSDA is formulated as follows. First, audio and visual features extracted per ECF settings are denoted as $\mathbf{F}^m, \mathbf{F}^{m'} \in \mathbb{R}^{T\times d}$ (where $m, m' \in \{a, v\}$). To enable differential modeling, $\mathbf{F}^m$ is

linearly projected into two sets of queries $\mathbf{Q}_{1,2} \in \mathbb{R}^{T\times d}$, while $\mathbf{F}^{m'}$ is projected into two sets of keys $\mathbf{K}_{1,2} \in \mathbb{R}^{T\times d}$ and common values $\mathbf{V} \in \mathbb{R}^{T\times 2d}$. CSDA then employs a multi-head mechanism to perform differential denoising in parallel. For each head $j \in \{1,\dots,h\}$, let $\mathbf{A}_k = \sigma(\mathbf{Q}_k\mathbf{K}_k^\top/\sqrt{d})$ for $k \in \{1,2\}$, where $\sigma(\cdot)$ denotes the softmax function. The differential output is computed as:

$$\begin{aligned} \mathrm{DA}_j &= (\mathbf{A}_1 - \lambda\mathbf{A}_2)\mathbf{V}, \\ \mathbf{X}_j &= (1-\lambda_{\text{init}})\cdot \mathrm{RMSNorm}(\mathrm{DA}_j). \end{aligned} \tag{1}$$

Here, λ is a learnable scalar parameterized as $\lambda = \exp(\boldsymbol{\alpha}_1^\top\boldsymbol{\beta}_1) - \exp(\boldsymbol{\alpha}_2^\top\boldsymbol{\beta}_2) + \lambda_{\text{init}}$, where $\boldsymbol{\alpha}_{1,2}, \boldsymbol{\beta}_{1,2} \in \mathbb{R}^d$ are learnable vectors initialized from a normal distribution. This specific parameterization synchronizes the learning dynamics of λ with the attention gradients, ensuring optimization stability. λ_{init} is a constant initialized to 0.35. The outputs from all heads are concatenated and projected to obtain the intermediate feature $\widehat{\mathbf{F}}^m = \mathrm{Concat}(\mathbf{X}_1,\dots,\mathbf{X}_h)\mathbf{W}^O \in \mathbb{R}^{T\times d}$. Subsequently, a global semantic vector $\mathbf{G} = \mathrm{AvgPool}(\widehat{\mathbf{F}}^m) \in \mathbb{R}^d$ is derived to capture the video-level context. Local-global feature alignment is then achieved by computing a time-wise cosine similarity vector $\mathbf{S} \in \mathbb{R}^T$ to gate the global enhancement:

$$\mathbf{S}_t = \frac{\widehat{\mathbf{F}}_t^m \cdot \mathbf{G}}{\|\widehat{\mathbf{F}}_t^m\|\|\mathbf{G}\|}, \quad \widetilde{\mathbf{F}}_t^m = \widehat{\mathbf{F}}_t^m + \mathbf{S}_t \cdot \mathbf{G}. \tag{2}$$

Finally, the SD-PMT block reconstructs the Multimodal Attention (MMA) unit by integrating intra-modality Self-Attention (SA) with CSDA, formally defined as:

$$\Phi_{\text{MMA}}(\mathbf{F}^m, \mathbf{F}^{m'}) = \phi_{\text{CSDA}}\Big(\phi_{\text{SA}}(\mathbf{F}^m),\ \phi_{\text{SA}}(\mathbf{F}^{m'})\Big). \tag{3}$$

Combined with the snippet shift mechanism to prevent event fragmentation, as illustrated in Fig. 2-(b), the two-stage computation is given by:

$$\begin{aligned} \mathbf{H}^m &= \Phi_{\text{MMA}}\Big(\mathrm{Shift}\big(\Phi_{\text{MMA}}(\mathbf{F}^m, \mathbf{F}^{m'})\big),\ \mathrm{Shift}(\mathbf{F}^{m'})\Big), \\ \mathbf{F}_{\text{out}}^m &= \phi_{\text{SA}}(\mathbf{F}^m) + \mathrm{Shift}^{-1}(\mathbf{H}^m). \end{aligned} \tag{4}$$

Specifically, for a pyramid window of size W, the Shift operation cyclically displaces the feature sequence by $W/2$ steps. This reconfiguration enables cross-window interaction in the second stage, mitigating the boundary limitations of non-overlapping partitions, after which Shift^{-1} reverses the displacement to preserve temporal alignment.

Pyramid Hierarchical Strategy Driven by SD-PMT. In a pyramid architecture where the temporal window size grows exponentially (2^l) as shown in Fig. 2-(a), features at different scales face distinct theoretical challenges: short-term windows are dominated by high-frequency local noise accumulation, whereas long-term windows suffer from signal attenuation due to repeated aggregation. To resolve this conflict, a differentiated block design is proposed tailored to signal characteristics.

Specifically, for shallow layers dealing with local details, noise suppression is prioritized. The Stacked-SD Block (SDB) is employed, which stacks two consecutive SD-PMT layers to aggressively filter out local semantic noise and purify fine-grained semantics. Conversely, for deeper layers modeling long-range dependencies, preserving signal integrity becomes critical. To this end, the Alternating-SD Block (ADB) is designed by alternating the denoising SD-PMT with the original PMT unit. This hybrid structure effectively balances noise removal with the preservation of long-term temporal dependencies, thereby preventing feature degradation. In the proposed hierarchy, SDBs are followed by ADBs (e.g., with a 1:2 ratio), enabling the differential mechanism to adaptively process events of varying durations.

C. Multi-Granularity Semantic Loss Guidance

To synergistically improve fine-grained discrimination and inter-event relation modeling, a Multi-Granularity Semantic Loss Guidance mechanism is designed to establish a hierarchical supervision paradigm spanning from local segments to global events.

Segment-Level Semantic Guided Loss. In the event extraction stage, the quality of event graphs relies heavily on the discriminative power of snippet features. However, relying solely on coarse video-level labels often limits fine-grained semantic learning, leading to *shortcut learning* where the model overfits to only the most discriminative frames. This results in incomplete feature activations, which in turn degrades the construction of event graphs. To rectify this, inspired by prior work [15], [21], [22], a segment-level semantic guided loss $\mathcal{L}_{\text{s}}$ is introduced as an explicit fine-grained supervisory signal. By leveraging segment-level pseudo-labels $\mathbf{p}_{\text{pseudo}}^m$ generated from pre-trained open-vocabulary models (CLIP [23] and CLAP [24]), this loss forces the model to learn semantically complete feature representations, thereby enhancing the reliability of feature nodes for subsequent event graph construction.

Specifically, for segment-level predictions $\mathbf{p}^m$ obtained in the event extraction stage, the loss $\mathcal{L}_{\text{s}}$ is computed as follows:

$$\mathcal{L}_{\text{s}} = \sum_{m\in\{a,v\}} \mathrm{BCE}(\mathbf{p}^m, \mathbf{p}_{\text{pseudo}}^m). \tag{5}$$

Here, $\mathbf{p}^m, \mathbf{p}_{\text{pseudo}}^m \in \mathbb{R}^{T\times K}$, where T denotes the temporal length and K the total number of event classes.

Event-Aware Joint Loss. During the event interaction stage, individual event features are refined to capture complex dependencies. However, the classification loss $\mathcal{L}_{\text{cls}}$ often exhibits high volatility in long-form video scenarios with dense event overlaps. This instability acts as an optimization bottleneck, directly interfering with and restricting the capacity of $\mathcal{L}_{\text{rel}}$ to effectively guide the modeling of event relations. To mitigate this interference and unleash the relational modeling potential, an Event-Aware Joint Loss Weighting Strategy ($\mathcal{L}_e$) is proposed:

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE LFAV DATASET.

Method	mAP (%) (snippet-level)				F1-score (%) (event-level)			
	Audio	Visual	Audio-visual	Avg.	Audio	Visual	Audio-visual	Avg.
HAN [8]	48.27	63.26	47.42	52.98	20.96	24.13	18.07	20.87
DHHN [10]	49.95	59.59	47.74	52.42	21.44	18.82	14.23	18.16
CMPAE [25]	45.23	59.18	44.79	49.73	20.52	27.53	21.86	23.30
CM-PIE [11]	47.67	63.42	45.81	52.30	24.08	30.20	22.75	25.67
MM-CSE [16]	52.44	65.00	49.63	55.69	25.37	29.22	22.16	25.58
ECF [19]	51.66	64.10	50.11	55.29	25.36	32.44	26.61	28.14
Ours (ESDN)	**52.48**	**66.13**	**50.98**	**56.53**	**27.94**	**32.91**	**26.96**	**29.27**

$$\begin{aligned} \mathcal{L}_e &= \alpha \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{rel}}, \\ \mathcal{L}_{\text{cls}} &= \text{BCE}(\hat{\mathbf{p}}^a, \mathbf{y}^a) + \text{BCE}(\hat{\mathbf{p}}^v, \mathbf{y}^v), \\ \mathcal{L}_{\text{rel}} &= \sum_{m \in \{a,v\}} \frac{1}{|S_m|} \sum_{i \in S_m} \text{BCE}(\tilde{\mathbf{p}}^{m,i}, \mathbf{y}^m[i]). \end{aligned} \tag{6}$$

Here, $\mathcal{L}_{\text{cls}}$ ensures overall recognition based on final video-level predictions $\hat{\mathbf{p}}^m$ and ground-truth labels $\mathbf{y}^m$. The relation-aware loss $\mathcal{L}_{\text{rel}}$ provides direct supervision for the refined features of active event classes $S_m = \{i \mid \mathbf{y}^m[i] = 1\}$, where $\tilde{\mathbf{p}}^{m,i}$ denotes the prediction derived from the i-th refined event feature. By incorporating calibration coefficients α and β, this strategy adaptively modulates the optimization focus to dampen the impact of volatile global signals. This ensures that the learning of inter-event dependencies is not suppressed by classification instability, thereby allowing the interaction module to fully exploit deep relational semantics.

Total Training Objective. The total training objective $\mathcal{L}$ integrates stage-wise supervision across all granularities to achieve global synergistic optimization. At each stage $j \in \{1, 2\}$, the multi-modal video-level loss $\mathcal{L}_j$ is defined as:

$$\mathcal{L}_j = \text{BCE}(\bar{\mathbf{p}}^a_{(j)}, \mathbf{y}^a) + \text{BCE}(\bar{\mathbf{p}}^v_{(j)}, \mathbf{y}^v), \tag{7}$$

where $\bar{\mathbf{p}}^m_{(j)}$ denotes the video-level prediction for modality m at stage j. The comprehensive final objective is formulated as:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_{\text{s}} + \mathcal{L}_e. \tag{8}$$

III. EXPERIMENTAL RESULTS

A. Dataset and Evaluation Metrics

The LFAV [19] dataset comprises 5,175 videos totaling 302 hours, covering 35 categories, with 52% containing three or more event types. All videos are longer than 30 seconds and include at least one second of audio or visual events; 74% exceed two minutes, and 2% exceed ten minutes. The dataset is randomly split into 3,721 training, 486 validation, and 968 test videos, with 24,875 video-level and 23,666 secondary event annotations (validation and test sets only). Each event category appears at least 146 times, with significant inter-category label correlations. Event-level predictions for audio, visual, and audio-visual modalities are evaluated using the F1-score, and segment-level predictions using mAP.

B. Implementation Details

Our experimental settings follow ECF [19]. using a batch size of 16 and Adam optimization (initial learning rate 1×10^{-4}) for 30 epochs, with the learning rate reduced by 0.1 every 10 epochs. In the snippet prediction stage, number of heads 4, dropout 0.2, and SDB-to-ADB ratio 1:2. In the event extraction stage, the semantic edge confidence threshold is 0.5. For the $\mathcal{L}_e$ loss, the $\mathcal{L}_{cls}$ and $\mathcal{L}_r$ weights α and β are set to 0.5 and 0.3, respectively.

C. Comparison with State-of-the-art Methods

Quantitative results on the LFAV dataset are summarized in Table I. It is observed that ESDN achieves superior performance across all evaluation metrics, notably surpassing the ECF baseline. The systemic gains across all event categories—highlighted by a significant 10.2% relative improvement in the event-level Audio F1-score (from 25.36% to 27.94%) and consistent leads in both snippet-level mAP (56.53% avg.) and event-level F1-score (29.27% avg.)—demonstrate that ESDN robustly addresses the inherent challenges of semantic entanglement and noise interference. Such comprehensive improvements validate the synergistic effect of integrating a denoising architecture with multi-granularity supervision for long-form video understanding.

To qualitatively verify these advantages, a representative comparison is provided in Fig. 3. In this dense temporal overlap scenario, the ECF model fails to detect the guitar event and suffers from noise interference during singing recognition, resulting in low recognition precision. In contrast, ESDN accurately localizes the previously missed events and attains significantly higher precision. These results further indicate that the proposed framework effectively decouples relevant event semantics from the noise induced by overlapping events and intertwined multimodal signals, thereby substantially enhancing overall recognition and localization capabilities in complex environments.

D. Ablation Studies

Ablation study of CSDA and MSG. Table II assesses the respective contributions of CSDA and MSG. Comparing Model #1 (baseline) with Model #3 reveals that incorporating CSDA alone yields substantial gains, confirming its pivotal role in suppressing cross-modal noise and disentangling event semantics. In contrast, Model #2 (MSG only) results in limited

TABLE II
ABLATION STUDY OF CSDA AND MSG OF OUR PROPOSED METHOD.

Model	CSDA	MSG	mAP (%) (snippet-level)				F1-score (%) (event-level)			
			Audio	Visual	Audio-visual	Avg.	Audio	Visual	Audio-visual	Avg.
#1	×	×	51.66	64.10	50.11	55.29	25.36	32.44	26.61	28.14
#2	×	✓	49.55	64.50	48.20	54.09	25.92	32.68	25.74	28.11
#3	✓	×	51.74	65.35	50.23	55.77	26.67	32.55	**27.16**	28.76
#4	✓	✓	**52.48**	**66.13**	**50.98**	**56.53**	**27.94**	**32.91**	26.96	**29.27**

TABLE III
EFFECTS OF DIFFERENT COMPONENTS, λ_{init}, AND SDB:ADB STACKING RATIO ON MODEL PERFORMANCE. ONLY AVG-MAP AND AVG-F1 ARE REPORTED FOR COMPACTNESS.

Setting	Avg-map (%)	Avg-F1 (%)
Component Ablation		
full	**56.53**	**29.27**
w/o GSF	54.40	27.45
w/o Diff	54.36	28.85
w/o $\mathcal{L}_s$	54.35	27.96
w/o $\mathcal{L}_e$	54.07	28.02
SDB:ADB		
3:0	55.69	29.04
2:1	55.77	28.76
1:2	**56.53**	**29.27**
0:3	55.59	28.49
λ_{init}		
$0.8\text{-}0.6\exp(-0.3l)$	56.14	28.74
0.8	55.50	28.28
0.35	**56.53**	**29.27**

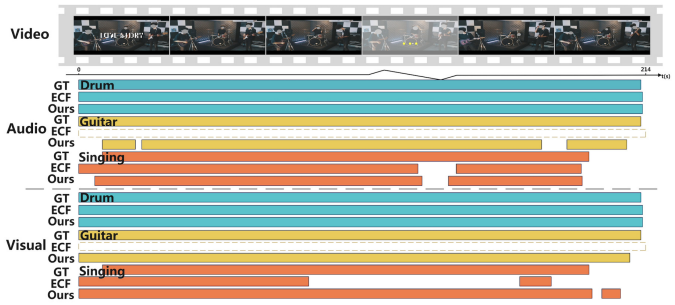

Fig. 3. Qualitative comparisons with the previous best method ECF.dashed boxes mark undetected Guitar events.

or even negative performance gains. This observation indicates that in long-form videos with severe semantic entanglement, fine-grained supervision cannot be effectively leveraged without a robust denoising architecture. The best performance achieved by Model #4 (CSDA + MSG) validates a critical synergy: CSDA purifies the feature representations, thereby providing a solid foundation for MSG to perform both precise segment-level semantic calibration and deep event-level relation modeling. These results confirm that the synergy between CSDA and MSG effectively resolves the core challenges of semantic entanglement and noise interference in long-form scenarios.

Ablation Analysis of Model Components, Noise Initialization, and Pyramid Stacking. As summarized in Table III, within the Cross-modal Semantic Differential Attention (CSDA) module, the removal of Global Semantic Fusion (w/o GSF) or the replacement of the Cross-modal Differential Mechanism (w/o Diff.) with standard cross-modal attention results in a significant performance decline. These results validate the necessity of GSF for maintaining long-term semantic consistency, while demonstrating that the cross-modal differential mechanism is essential for actively decoupling noise and enhancing relevant features. Furthermore, the performance decline resulting from the ablation of either $\mathcal{L}_s$ or $\mathcal{L}_e$ confirms that neither component can be neglected. This demonstrates that fine-grained segment-level semantic guidance and effective inter-event relation modeling are both indispensable for the MSG mechanism. Specifically, $\mathcal{L}_s$ facilitates precise local feature characterization, while $\mathcal{L}_e$ ensures that the model can adequately exploit relational semantics without being hindered by the volatility of global classification targets.

Analysis of the pyramid hierarchy reveals that both the standalone SDB and hybrid SDB-ADB configurations yield substantial improvements over the baseline. Notably, an SDB-to-ADB ratio of 1:2 provides the optimal trade-off, demonstrating that aggressive denoising in shallow layers effectively curtails local noise accumulation, while the alternating units in deeper layers successfully achieve a balance between noise suppression and temporal dependency modeling. Furthermore, experiments on noise initialization indicate that excessively high λ_{init} values trigger over-denoising and subsequent signal attenuation. Conversely, maintaining $\lambda_{\text{init}} = 0.35$—approximating the starting point of the dynamic schedule $0.8-0.6\exp(-0.3l)$ [20]—facilitates an effective suppression of irrelevant noise while preserving the richness of integrated features.

IV. CONCLUSION

To resolve the challenges of semantic entanglement and noise in long-form videos, the Event-Centric Semantic Differential Denoising Network (ESDN) is proposed. Specifically, SD-PMT utilizes Cross-modal Semantic Differential Attention to actively decouple noise from event semantics, while employing a pyramid hierarchy strategy to reconcile the conflict between local denoising and long-range temporal modeling. Furthermore, the Multi-Granularity Semantic Loss Guidance mechanism is incorporated to mitigate label noise and enhance inter-event relation modeling. Extensive experiments confirm that ESDN significantly outperforms existing methods.

REFERENCES

[1] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 961–970.

[2] M. Soldan, A. Pardo, J. L. Alcázar, F. C. Heilbron, C. Zhao, S. Giancola, and B. Ghanem, "MAD: A Scalable Dataset for Language Grounding in Videos from Movie Audio Descriptions," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5016–5025.

[3] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1207–1216.

[4] C.-Y. Wu and P. Krähenbühl, "Towards Long-Form Video Understanding," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1884–1894.

[5] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2405–2414.

[6] Z. Zhao, Z. Zhang, S. Xiao, Z. Xiao, X. Yan, J. Yu, D. Cai, and F. Wu, "Long-Form Video Question Answering via Dynamic Hierarchical Reinforced Networks," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5939–5952, 2019.

[7] Y. Sun, H. Xue, R. Song, B. Liu, H. Yang, and J. Fu, "Long-form video-language pre-training with multimodal temporal contrastive learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[8] Y. Tian, D. Li, and C. 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*. Berlin, Heidelberg: Springer-Verlag, 2020, pp. 436–454. [Online]. Available: https://doi.org/10.1007/978-3-030-58580-8_26

[9] J. Yu, Y. Cheng, R.-W. Zhao, R. Feng, and Y. 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*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 6241–6249. [Online]. Available: https://doi.org/10.1145/3503161.3547869

[10] X. Jiang, X. Xu, Z. Chen, J. Zhang, J. Song, F. Shen, H. Lu, and H. T. Shen, "DHHN: Dual Hierarchical Hybrid Network for Weakly-Supervised Audio-Visual Video Parsing," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 719–727. [Online]. Available: https://doi.org/10.1145/3503161.3548309

[11] Y. Chen, R. Guo, X. Liu, P. Wu, G. Li, Z. Li, and W. 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)*, 2024, pp. 8421–8425.

[12] H. Cheng, Z. Liu, H. Zhou, C. Qian, W. Wu, and L. Wang, "Joint-Modal Label Denoising for Weakly-Supervised Audio-Visual Video Parsing," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*. Berlin, Heidelberg: Springer-Verlag, 2022, pp. 431–448. [Online]. Available: https://doi.org/10.1007/978-3-031-19830-4_25

[13] J. Zhang and W. Li, "Multi-modal and multi-scale temporal fusion architecture search for audio-visual video parsing," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 3328–3336. [Online]. Available: https://doi.org/10.1145/3581783.3611947

[14] J. Fu, J. Gao, B.-K. Bao, and C. Xu, "Multimodal imbalance-aware gradient modulation for weakly-supervised audio-visual video parsing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4843–4856, 2024.

[15] Y.-H. Lai, Y.-C. Chen, and Y.-C. F. Wang, "Modality-independent teachers meet weakly-supervised audio-visual event parser," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[16] P. Zhao, J. Zhou, Y. Zhao, D. Guo, and Y. Chen, "Multimodal class-aware semantic enhancement network for audio-visual video parsing," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i10.33134

[17] P. Zhao, Y. Chen, D. Guo, and Y. Yao, "Text-Infused Audio-Visual Video Parsing with Semantic-Aware Multimodal Contrastive Learning," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[18] L. Wang, B. Zhu, Y. Chen, and J. Wang, "Link: Adaptive modality interaction for audio-visual video parsing," 2024. [Online]. Available: https://arxiv.org/abs/2412.20872

[19] W. Hou, G. Li, Y. Tian, and D. Hu, "Toward Long Form Audio-Visual Video Understanding," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 9, Sep. 2024. [Online]. Available: https://doi.org/10.1145/3672079

[20] T. Ye, L. Dong, Y. Xia, Y. Sun, Y. Zhu, G. Huang, and F. Wei, "Differential Transformer," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=OvoCm1gGhN

[21] J. Zhou, D. Guo, Y. Zhong, and M. Wang, "Advancing Weakly-Supervised Audio-Visual Video Parsing via Segment-Wise Pseudo Labeling," *Int. J. Comput. Vision*, vol. 132, no. 11, p. 5308–5329, Jun. 2024. [Online]. Available: https://doi.org/10.1007/s11263-024-02142-3

[22] Y. Fan, Y. Wu, B. Du, and Y. Lin, "Revisit weakly-supervised audio-visual video parsing from the language perspective," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[24] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253510826

[25] J. Gao, M. Chen, and C. Xu, "Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio- Visual Event Perception," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 827–18 836.