Curriculum-Listener: Consistency- and Complementarity-Aware Audio-Enhanced Temporal Sentence Grounding

Houlun Chen
chenhl23@mails.tsinghua.edu.cn
DCST, Tsinghua University

Xin Wang*
xin_wang@tsinghua.edu.cn
DCST, BNRist, Tsinghua University

Xiaohan Lan
lanxh20@tsinghua.org.cn
DCST, Tsinghua University

Hong Chen
h-chen20@mails.tsinghua.edu.cn
DCST, Tsinghua University

Xuguang Duan
duan_xg@outlook.com
DCST, Tsinghua University

Jia Jia*
Wenwu Zhu*
{jjia,wwzhu}@tsinghua.edu.cn
DCST, BNRist, Tsinghua University

ABSTRACT

Temporal Sentence Grounding aims to retrieve a video moment given a natural language query. Most existing literature merely focuses on visual information in videos without considering the naturally accompanied audio which may contain rich semantics. The few works considering audio simply regard it as an additional modality, overlooking that: i) it's non-trivial to explore consistency and complementarity between audio and visual; ii) such exploration requires handling different levels of information densities and noises in the two modalities. To tackle these challenges, we propose **A**daptive **D**ual-branch **P**romoted **N**etwork (ADPN) to exploit such consistency and complementarity: i) we introduce a dual-branch pipeline capable of jointly training visual-only and audio-visual branches to simultaneously eliminate inter-modal interference; ii) we design **T**ext-**G**uided **C**lues **M**iner (TGCM) to discover crucial locating clues via considering both consistency and complementarity during audio-visual interaction guided by text semantics; iii) we propose a novel curriculum-based denoising optimization strategy, where we adaptively evaluate sample difficulty as a measure of noise intensity in a self-aware fashion. Extensive experiments show the state-of-the-art performance of our method.[1]

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

audio-visual; temporal sentence grounding; curriculum learning

ACM Reference Format:
Houlun Chen, Xin Wang, Xiaohan Lan, Hong Chen, Xuguang Duan, Jia Jia, and Wenwu Zhu. 2023. Curriculum-Listener: Consistency- and Complementarity-Aware Audio-Enhanced Temporal Sentence Grounding. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612504

*Corresponding Authors.
[1]Code is available at https://github.com/hlchen23/ADPN-MM.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612504

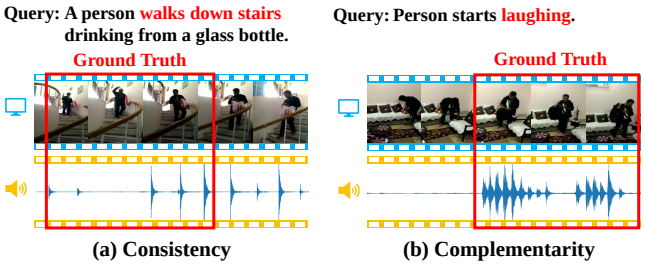

Figure 1: An illustration of (a) consistency and (b) complementarity in TSG. (a) Visual content and sound of footsteps consistently match "walks down stairs". (b) It's difficult to recognize discriminative action visually to localize "laughing" but laughter provides complementary locating clues.

1 INTRODUCTION

Temporal Sentence Grounding (TSG) [1, 11] aims to retrieve one moment from an untrimmed video that semantically matches a descriptive natural language query [21]. For a long time, existing works on TSG [39, 46, 60, 61] merely consider static frames in videos. Inspired by multimodal learning, recent TSG works [1, 9, 26] integrate multiple modalities in videos and fuse them to achieve better performance, where all the modalities derive from visual information including but not limited to RGB images, optical flows, depth etc. However, these works ignore the naturally accompanied audio signals in videos, which may contain useful and rich semantics as well. Audio signals are consistent with visual signals both temporally and semantically, providing discriminative clues as a complement when visual information is missing or unrecognizable. Research [2, 17, 37, 38, 40, 41] in many other video analysis tasks such as video object segmentation and video event recognition have proved that audio does contribute to deeper understanding of objects and activities in videos. As shown in Figure 1, TSG benefits from audio as well, so it's worth exploring to combine visual and audio modalities for better moment localization.

Though a few works [8, 31] consider the audio modality for TSG, they suffer from the following limitations. First, they merely treat audio as an additional modality homogeneous to other modalities and employ neural network architectures regardless of different

modalities, thus leaving the consistency and complementarity between audio and visual insufficiently explored. Besides, they ignore the importance of texts when modeling the audio-visual interactions, in which the textual queries may contain semantics for localization shared by other modalities.

Therefore, in this paper we study the **A**udio-enhanced **T**emporal **S**entence **G**rounding (ATSG) to overcome these limitations, where audio serves as an auxiliary modality, aiming to capture more accurate locating clues with comprehensive utilization of both visual and audio, especially when visual modality collapses. However, handling ATSG poses the two challenges: i) it's non-trivial to explore the consistency and complementarity between audio and visual; ii) such exploration requires the ability to handle different levels of information densities and noises in the two modalities since audio is usually less informative and contains noisy information.

To solve the challenges, we propose an **A**daptive **D**ual-branch **P**romoted **N**etwork (ADPN) for ATSG to better capture locating clues with the introduction of audio. i) To utilize extra semantics from audio while suppressing the inter-modal interference when more modalities are introduced, we design a dual-branch pipeline to fill the information gap by jointly training visual-only and audio-visual branches, which contributes to maintaining valid information from visual when audio is redundant or noisy. ii) To better model the audio-visual interactions, we design a transformer-based **T**ext-**G**uided **C**lues **M**iner (TGCM) to exploit both consistent and complementary components across audio and visual with the semantics of textual queries as guidance, where text works as a bridge to transfer shared semantics to audio and visual features, discovering the crucial locating clues in this process. iii) In order to effectively separate and eliminate noisy information, we consider the denoising process as handling the modality imbalance problem and design a curriculum learning strategy. More specifically, we develop a set of difficulty evaluation criteria to approximately measure the noise intensity from the outputs of two branches and adaptively adjust the optimization process of these two branches by re-weighting specific loss functions.

We conduct extensive experiments on Charades-STA [11] and ActivityNet Captions [19] benchmark datasets, showing that our ADPN achieves state-of-the-art performance against baseline methods. Besides, ablation studies indicate that our ADPN is able to eliminate noise in audio by maintaining significant semantics in visual, and can capture key locating clues from audio especially when visual information is damaged. Finally, we conduct case studies to illustrate how our ADPN benefits from the consistency and complementarity between audio and visual, providing interpretability to our method.

To sum up, our contributions can be summarized as follows:

(1) We study the Audio-enhanced Temporal Sentence Grounding (ATSG) and propose the Adaptive Dual-branch Promoted Network (ADPN) to introduce audio.
(2) We design the Text-Guided Clues Miner (TGCM) to discover crucial clues during the interactions of text, audio and visual, taking both consistency and complementarity into consideration.
(3) We design a novel curriculum learning strategy, where we measure difficulty for training samples in a self-aware fashion and adjust the optimization process adaptively to denoise the audio modality.
(4) Extensive experiments demonstrate that our ADPN achieves competitive performance against baselines and obtains significant performance improvement with the assistance of audio modality.

2 RELATED WORKS

Temporal Sentence Grounding (TSG). TSG [1, 11] aims to retrieve a video segment given a natural language query, which requires understanding correlations in multiple modalities. Existing supervised TSG methods can mainly be grouped into two categories: (1) Proposal-based methods [4, 11, 58, 61, 64] search for candidate segments and match them with the query. (2) Proposal-free methods [24, 39, 59, 60] model the prediction of start-and-end timestamps as a regression problem. Besides, a few works explore to combine these two paradigms [27, 54] to balance the performance and efficiency. An important part of all of them is the interaction or fusion between queries and videos. Most TSG works only focus on the RGB information from videos and model the interaction or fusion within queries and videos by technologies such as Attention Mechanism [4, 10, 39, 56, 58, 62, 63], Hadamard Product [39, 64], Convolutional Neural Networks (CNN) [39, 58], and Graph Neural Networks (GNN) [7, 28, 49]. For more reasonable queries-videos semantic alignment, some works [51, 60, 63] explore to adopt multi-level/-stage interaction fashions.

Since it's difficult to capture complicated and subtle semantics merely from RGB, recent TSG works incorporate more modalities. MCN [1] integrates RGB images and optical flows with late fusion strategy. Chen et al. [9] find the redundancy in RGB and integrate RGB images, optical flows and depth modalities from videos by a dynamic fusion mechanism with transformers. Liu et al. [26, 29] argue that low-level information from consecutive RGB frames fails to describe complicated activities and introduce optical flow information [26] and object features [26, 29] in addition to RGB images for motion-awareness.

However, these modalities still derive from visual information and share similar semantics, which are not able to provide comprehensive clues for moment localization, so a few TSG works take audio into consideration. PMI-LOC [8] adopts RGB, motion and audio and designs pairwise modality interactions in both sequence and channel levels. UMT [31] proposes a unified multimodal transformer framework to fuse visual and audio. However, they ignore the information gap between audio and visual modalities. We propose a dual-branch mutually-promoted pipeline with a carefully designed Text-Guided Clues Miner (TGCM) to fill this gap.

Curriculum Learning. Curriculum learning, first proposed by Bengio et al. [3], is a training strategy inspired by human curricula that means training a model from easier to harder data. It can be generalized to modulating the training process guided by a difficulty measure, which is not limited to changing the exposure order of training data. Wang et al. [53] argue that curriculum learning can be unified into the general framework of *Difficulty Measurer + Training Scheduler* and various curriculum learning strategies can

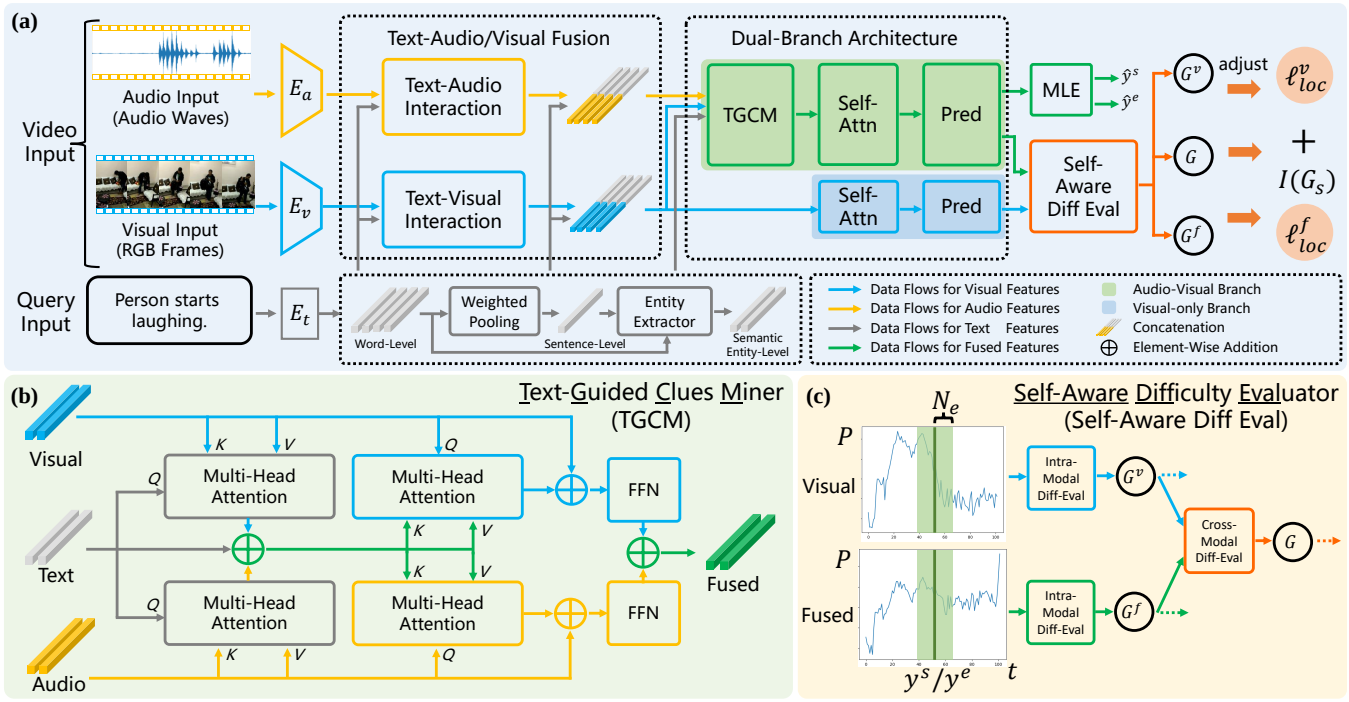

Figure 2: (a) An overview of our proposed Adaptive Dual-branch Promoted Network (ADPN) for ATSG, in which two branches, visual-only and audio-visual, are jointly trained. The output of the audio-visual branch is used to predict the start-and-end timestamps. (b) A detailed diagram of Text-Guided Clues Miner (TGCM), where text works as a bridge to transfer shared information to audio and visual modalities. (c) Our self-aware difficulty evaluator takes the probability distributions of the start-and-end timestamps from two branches as input and generates three difficulty grades to adjust the optimization process.

be categorized to Pre-defined Curriculum Learning [3, 44], Self-paced Learning [20, 36], Transfer Teacher [14, 65], Reinforcement Learning Teacher [13, 35] etc. based on this framework. Since samples with noise can be seen as hard samples, curriculum learning has been used to denoise the learning process in many machine learning problems [6, 34, 45].

Imbalanced Multimodal Learning. Audio may accompany noisy information that would destroy the audio-visual consistency and complementarity. For example, background music in some user-generated videos often has weak semantic correlations with visual content, which may mislead the model to learn from this noise and forget valid information from the visual modality. However, existing TSG methods [8, 31] neglect this problem when introducing the audio modality. One similar problem is the modality imbalance problem, which means some modalities are not perfectly trained because of interference from other modalities. It has aroused much attention in the field of multimodal and machine learning [16, 52]. Many works try to alleviate it from the perspective of optimization to coordinate the learning process for different modalities. Wu et al. [55] defines the Conditional Learning Speed for each modality and takes re-balancing optimization steps according to it. Jiang et al. [17] designs an additional Unidirectional Guiding Loss to transfer unimodal discriminative knowledge to multimodal learning branches. Peng et al. [42] proposes On-the-fly Gradient Modulation to adaptively control the optimization for each modality. Drawing inspiration from this, we design a curriculum learning strategy to coordinate audio-visual learning, where we remove defective gradients guided by a self-aware difficulty evaluator.

3 PROPOSED METHOD

In this part, we elaborate on our Adaptive Dual-branch Promoted Network (ADPN) (Figure 2 (a)). After giving the problem formulation (Section 3.1), we describe the technical details of feature encoding (Section 3.2) and text-audio / visual fusion (Section 3.3). Then, we describe our dual-branch architecture (Section 3.4) which enhances audio-visual learning by jointly-training strategy. Within the audio-visual branch, to discover shared key locating clues within text, audio and visual modalities, we design a Text-Guided Clues Miner (TGCM) (Section 3.5) to model audio-visual interaction guided by text semantics. Finally, we introduce our curriculum learning strategy (Section 3.6) where we adjust the optimization process adaptively.

3.1 Problem Formulation

In ATSG, a piece of training data can be formalized as $(\mathbf{Q}, \mathbf{V}, \mathbf{y})$, where $\mathbf{Q}$, $\mathbf{V}$, $\mathbf{y}$ are query input, video input, and the ground truth start-and-end timestamps, respectively. Query input is formalized as $\mathbf{Q} = \{\mathbf{w}_i\}_{i=1}^{L_q}$; video input $\mathbf{V}$ can be divided into audio and visual modalities i.e. $\mathbf{V} = \{\mathbf{X}^v, \mathbf{X}^a\}$, where $\mathbf{X}^v = \{\mathbf{x}_t^v\}_{t=1}^{T_v}$ and $\mathbf{X}^a =$

$\{\mathbf{x}_t^a\}_{t=1}^{T_a}$. The goal is to predict start-and-end timestamps, whose ground truth annotations are (y^s, y^e), denoted as $\mathbf{y}$.

3.2 Feature Encoding

Queries. We encode the queries in multiple granularities. The word-level features $\mathbf{Q} = \{\mathbf{w}_i\}_{i=1}^{L_q} \in \mathbb{R}^{L_q \times D}$ are generated from both word and character embeddings. Then we pass $\mathbf{Q}$ into a self-weighted pooling layer to get the sentence-level features $\mathbf{Q}^s \in \mathbb{R}^D$. To balance expressiveness and computational costs in the two features, we adopt a recurrent approach inspired by [39] to compress a sentence into several semantic entity-level features $\mathbf{Q}^e = \{\mathbf{e}_i\}_{i=1}^{L_e} \in \mathbb{R}^{L_e \times D}$ where $L_e \ll L_q$. We calculate L_e guidance vectors $\mathbf{Q}^g = \{\mathbf{g}_i\}_{i=1}^{L_e} \in \mathbb{R}^{L_e \times D}$ with different focuses in a recurrent way and use $\mathbf{Q}^g$ as query vectors to extract semantic entity-level features $\mathbf{Q}^e$ by attention mechanism. In the n-th step, we first calculate g_n:

$$\mathbf{g}_n = \text{ReLU}(\mathbf{W}^g([\mathbf{W}_n^{gq}\mathbf{Q}^s; \mathbf{e}_{n-1}])) \tag{1}$$

where $\mathbf{W}^g \in \mathbb{R}^{D \times 2D}$ and $\mathbf{W}_n^{gq} \in \mathbb{R}^{D \times D}$ are learnable parameters and $\mathbf{W}_n^{gq}$ is step-specific; [;] denotes concatenation operation. The semantic entity in the previous step $\mathbf{e}_{n-1}$ is incorporated to introduce historical information during recurrences. Then we use additive attention on word-level features $\mathbf{Q}$ to generate $\mathbf{e}_n$:

$$\begin{aligned} \mathbf{C}^{(n)} &= \text{softmax}(\mathbf{w}^{c\top}(\tanh(\mathbf{W}^{cg}\mathbf{g}_n + \mathbf{W}^{cq}\mathbf{Q}^\top))) \\ \mathbf{e}_n &= \sum_{i=1}^{L_q} c_i^{(n)}\mathbf{w}_i \end{aligned} \tag{2}$$

where $\mathbf{w}^c \in \mathbb{R}^{\frac{D}{2}\times 1}$, $\mathbf{W}^{cg} \in \mathbb{R}^{\frac{D}{2}\times D}$ and $\mathbf{W}^{cq} \in \mathbb{R}^{\frac{D}{2}\times D}$ are learnable parameters; $\mathbf{C}^{(n)} = \{c_i^{(n)}\}_{i=1}^{L_q}$.

To extract more diversified semantic entity-level features, we follow the regularization technique in [25, 39] as our loss ℓ_e to force attention weights less similar to each other:

$$\ell_e = \|(\mathbf{C}^\top\mathbf{C}) - \eta\mathbf{I}_e\|_F^2 \tag{3}$$

where $\mathbf{C} \in \mathbb{R}^{L_q \times L_e}$ is the attention weights when generating $\mathbf{Q}^e$; $\|\cdot\|_F$ denotes Frobenius norm; $\mathbf{I}_e \in \mathbb{R}^{L_e \times L_e}$ is an identity matrix; η is a hyperparameter to control the overlapping extent across different weight distributions.

Videos. We extract audio and visual features by pretrained models and project them to the same dimension. After that, we add positional encoding to them and finally obtain $\mathbf{X}^a = \{\mathbf{x}_t^a\}_{t=1}^{T_a} \in \mathbb{R}^{T_a \times D}$ and $\mathbf{X}^v = \{\mathbf{x}_t^v\}_{t=1}^{T_v} \in \mathbb{R}^{T_v \times D}$.

3.3 Text-Audio/Visual Fusion

To highlight the part in audio/visual that's semantically relevant to the query, we modulate audio/visual features by query semantics in a fine-grained manner inspired by [58]. We linearly scale and shift audio/visual features by a dynamically computed sentence representation attended by each temporal feature unit in audio/visual. We use semantic entity-level features other than word-level features to improve efficiency. For the convenience of notation, we omit the superscript/subscript $(\cdot)^m/(\cdot)_m$ ($m \in \{a, v\}$) in this section since we model the text-audio (T-A) and text-visual (T-V) interactions in the same way. In detail, we compute a condensed representation $\overline{\mathbf{e}}^{(i)}$ from semantic entity-level features $\mathbf{Q}^e$ attended by the i-th feature unit of $\mathbf{x}_i$:

$$\begin{aligned} \mathbf{A}^{(i)} &= \text{softmax}(\mathbf{w}^{\alpha\top}\tanh(\mathbf{W}^{\alpha q}\mathbf{Q}^{e\top} + \mathbf{W}^{\alpha x}\mathbf{x}_i + \mathbf{b}^\alpha)) \\ \overline{\mathbf{e}}^{(i)} &= \sum_{n=1}^{L_e} \alpha_n^{(i)}\mathbf{e}_n \end{aligned} \tag{4}$$

where $\mathbf{w}^\alpha \in \mathbb{R}^{D\times 1}$, $\mathbf{W}^{\alpha q} \in \mathbb{R}^{D\times D}$, $\mathbf{W}^{\alpha x} \in \mathbb{R}^{D\times D}$ and $\mathbf{b}^\alpha \in \mathbb{R}^D$ are learnable parameters; $\mathbf{A}^{(i)} = \{\alpha_n^{(i)}\}_{n=1}^{L_e}$. Then we scale and shift $\mathbf{x}_i$ by coefficients β_i and γ_i:

$$\begin{aligned} \beta_i/\gamma_i &= \tanh(\mathbf{W}^{\beta/\gamma}\overline{\mathbf{e}}^{(i)} + \mathbf{b}^{\beta/\gamma}) \\ \mathbf{x}_i' &= \beta_i \cdot \mathbf{x}_i + \gamma_i \end{aligned} \tag{5}$$

where $\mathbf{W}^\beta \in \mathbb{R}^{1\times D}$, $\mathbf{W}^\gamma \in \mathbb{R}^{1\times D}$, $\mathbf{b}^\beta \in \mathbb{R}$ and $\mathbf{b}^\gamma \in \mathbb{R}$ are learnable parameters. We denote $\{\mathbf{x}_i'\}_{i=1}^{T}$ as $\mathbf{X}'$.

After that, we adopt Context-Query Attention [47, 57, 62] to further model text-audio / -visual interaction. We first compute a similarity matrix $\mathbf{S} \in \mathbb{R}^{T\times L_e}$ in the same way as [62], in which $\mathbf{S}_{i,j}$ indicates the similarity between i-th audio/visual feature and j-th semantic entity-level feature.

$$\mathbf{S} = \mathbf{X}'\mathbf{W}^{sx} + (\mathbf{Q}^e\mathbf{W}^{sq})^\top + (\mathbf{X}' \odot \mathbf{W}^s)\mathbf{Q}^{e\top} \tag{6}$$

where $\mathbf{W}^{sx} \in \mathbb{R}^{D\times 1}$, $\mathbf{W}^{sq} \in \mathbb{R}^{D\times 1}$ and $\mathbf{W}^s \in \mathbb{R}^{1\times D}$ are learnable parameters; $\odot$ denotes Hadamard product. $\mathbf{S}$ is computed with dimension expansion when calculated. Then we calculate context-to-query ($\mathbf{S}^{c2q} \in \mathbb{R}^{T\times D}$) and query-to-context ($\mathbf{S}^{q2c} \in \mathbb{R}^{T\times D}$) attention weights as:

$$\mathbf{S}^{c2q} = \mathbf{S}^r\mathbf{Q}^e,\ \mathbf{S}^{q2c} = \mathbf{S}^r\mathbf{S}^{c\top}\mathbf{X}' \tag{7}$$

where $\mathbf{S}^r \in \mathbb{R}^{T\times L_e}$ and $\mathbf{S}^c \in \mathbb{R}^{T\times L_e}$ are row- and column-wise normalization of $\mathbf{S}$ by softmax. We then compute audio/visual features $\overline{\mathbf{X}} \in \mathbb{R}^{T\times D}$ fused by text:

$$\overline{\mathbf{X}} = \text{FFN}([\mathbf{X}'; \mathbf{S}^{c2q}; \mathbf{X}' \odot \mathbf{S}^{c2q}; \mathbf{X}' \odot \mathbf{S}^{q2c}]) \tag{8}$$

where FFN is a feed-forward network; $\odot$ denotes Hadamard product.

Finally, we concatenate the sentence-level embedding $\mathbf{Q}^s$ to $\overline{\mathbf{X}}$ and pass it through a linear layer to keep the same dimension. For convenience, we still use $\overline{\mathbf{X}}$ for notation.

3.4 Dual-Branch Architecture

To maintain the key locating clues in visual while taking advantage of the audio-visual interaction, we split the data flow into two branches: audio-visual and visual-only, and train them jointly. The audio-visual branch takes audio, visual and text as input and fuses them to obtain prediction, while the visual-only branch gives prediction only by visual features.

When we get $\overline{\mathbf{X}}^a = \{\overline{\mathbf{x}}_i^a\}_{i=1}^{T_a} \in \mathbb{R}^{T_a\times D}$ and $\overline{\mathbf{X}}^v = \{\overline{\mathbf{x}}_i^v\}_{i=1}^{T_v} \in \mathbb{R}^{T_v\times D}$, we re-sample $\overline{\mathbf{X}}^a$ to the same length of $\overline{\mathbf{X}}^v$ and we pass them through TGCM in the audio-visual branch to perform their interaction, which will be elaborated on in Section 3.5. After that, we get audio-visual fused features $\overline{\mathbf{X}}^f = \{\overline{\mathbf{x}}_i^f\}_{i=1}^{T_f} \in \mathbb{R}^{T_f\times D}$ where $T_f = T_v$.

The following operations are the same for the two branches, so we omit the superscript/subscript $(\cdot)^m/(\cdot)_m$ ($m \in \{f, v\}$) in the following part of this section. We adopt self-attention with residual

connection to capture global correlations across audio and visual for $\overline{\mathbf{X}}^f$ and capture intra-modal correlations in visual for $\overline{\mathbf{X}}^v$.

$$\widetilde{\mathbf{X}} = \text{self-attn}(\overline{\mathbf{X}}) + \overline{\mathbf{X}} \tag{9}$$

where self-attn denotes the self-attention layer.

Finally, we use a transformer-based predictor following [62] to generate the probability distributions of start-and-end timestamps i.e. $\hat{P}^{s/e} = \text{pred}(\widetilde{\mathbf{X}})$, which can be detailed as follows:

$$\begin{aligned} \mathbf{H}^s &= \text{multi-head\_self-attn}(\text{conv1d}(\widetilde{\mathbf{X}})) \\ \mathbf{H}^e &= \text{multi-head\_self-attn}(\text{conv1d}(\mathbf{H}^s)) \\ \hat{P}^{s/e} &= \text{softmax}(\text{FFN}([\mathbf{H}^{s/e}; \widetilde{\mathbf{X}}])) \end{aligned} \tag{10}$$

where multi-head_self-attn, conv1d and FFN denote the multi-head self-attention layer, channel-wise separable 1D convolution and feed-forward network, respectively; $\hat{P}^{s/e}$ are the probability distributions of the start/end point predicted from the fused or visual features. We adopt the moment localization loss following [62]:

$$\ell_{loc} = \text{CE}(\hat{P}^s, Y^s) + \text{CE}(\hat{P}^e, Y^e) \tag{11}$$

where CE denotes cross-entropy loss; $Y^{s/e} = \{Y_i^{s/e}\}_{i=1}^{T_f} \in \{0,1\}^{T_f}$ represents the supervision where $Y_i^{s/e}$ is set to 1 only at the start/end point. Combining the two branches, the moment localization loss of prediction is:

$$\ell_{loc} = \ell_{loc}^f + \ell_{loc}^v \tag{12}$$

The total loss can be expressed as:

$$\ell = \ell_{loc} + \lambda \ell_e \tag{13}$$

During inference, we use Maximum Likelihood Estimation (MLE) to obtain the predicted $(\hat{y}^s, \hat{y}^e)$ with the constraint $\hat{y}^s \le \hat{y}^e$ from the audio-visual branch.

3.5 Text-Guided Clues Miner (TGCM)

To capture complicated correlations between audio and visual, we propose TGCM to model audio-visual interaction guided by text semantics, which contains two steps: *extracting* and *propagating*. Refer to Figure 2 (b) for details.

First, we use semantic entity-level features to extract shared semantics from audio and visual by attention mechanism with $\mathbf{Q}^e$ as query and $\overline{\mathbf{X}}^{a/v}$ as key and value vectors.

$$\mathbf{Q}^{e(a/v)} = \text{multi-head\_attn}(q = \mathbf{Q}^e, k = \overline{\mathbf{X}}^{a/v}, v = \overline{\mathbf{X}}^{a/v}) \tag{14}$$

where multi-head_attn denotes the multi-head attention layer with the specified query (q), key (k) and value (v). After extraction, we add $\mathbf{Q}^{e(a)}$ and $\mathbf{Q}^{e(v)}$ with residual connection to integrate consistent and complementary components:

$$\overline{\mathbf{Q}}^e = \mathbf{Q}^{e(a)} + \mathbf{Q}^{e(v)} + \mathbf{Q}^e \tag{15}$$

Then we propagate $\overline{\mathbf{Q}}^e$ to audio and visual features with $\overline{\mathbf{X}}^{a/v}$ as query and $\overline{\mathbf{Q}}^e$ as key and value vectors.

$$\overline{\mathbf{X}}^{q(a/v)} = \text{multi-head\_attn}(q = \overline{\mathbf{X}}^{a/v}, k = \overline{\mathbf{Q}}^e, v = \overline{\mathbf{Q}}^e) + \overline{\mathbf{X}}^{a/v} \tag{16}$$

Finally, we add $\overline{\mathbf{X}}^{q(a)}$ and $\overline{\mathbf{X}}^{q(v)}$ to obtain the fused features $\overline{\mathbf{X}}^f$:

$$\overline{\mathbf{X}}^f = \overline{\mathbf{X}}^{q(a)} + \overline{\mathbf{X}}^{q(v)} \tag{17}$$

3.6 Curriculum Optimization Strategy

In this section, we evaluate the difficulty of each sample as a measure of noise intensity in audio modality, as shown in Figure 2 (c). Considering the outputs of audio-visual and visual-only branches $\hat{P}^{s/e(f)}$ and $\hat{P}^{s/e(v)}$, we expand the boundary-level supervision $Y^{s/e}$ to $Y'^{s/e}$ such that $Y'^{s/e} = \{Y_i'^{s/e}\}_{i=1}^{T_f} \in \{0,1\}^{T_f}$ in which $Y_i'^{s/e} = 1$ where $\max\{y^{s/e} - N_e, 0\} \le i \le \min\{y^{s/e} + N_e, T_f\}$. N_e is the expansion coefficient. The difficulty grades of each sample for the audio-visual ($G^f \in (0,1)$) and visual-only branches ($G^v \in (0,1)$) are as follows:

$$G^{f/v} = \frac{1}{2}\left(\sum_{i=1}^{T_f} Y_i'^s \hat{P}_i^{s(f/v)} + \sum_{i=1}^{T_f} Y_i'^e \hat{P}_i^{e(f/v)}\right) \tag{18}$$

where lower $G^{f/v}$ means it's a harder sample for the corresponding branch. Then a relative difficulty grade across these two branches $G \in (0,1)$ can be calculated as follows:

$$G = \sigma(\log \frac{G^f}{G^v}) \tag{19}$$

where σ denotes the sigmoid function. G reflects the difficulty of a sample when audio modality is introduced and lower G means the introduction of audio makes learning harder compared to learning from visual individually, which can approximately be a measure of noise in audio when the model is trained enough. We denote G_s for (G, G^f, G^v).

To prevent defective gradients caused by noise in audio from back-propagating in the network, we adjust the loss function under the guidance of G_s. We modify ℓ_{loc} in Equation (12) for each sample as:

$$\ell_{loc} = \mathbb{I}(G_s)\ell_{loc}^f + \ell_{loc}^v \tag{20}$$

where $\mathbb{I}(\cdot)$ is an indicator function, i.e. $\mathbb{I}(G_s) = 0$ when $G < \overline{G}$ and $G^v > \overline{G^v}$; $\mathbb{I}(G_s) = 1$ in other situations. $\overline{G}$ and $\overline{G^v}$ are threshold hyperparameters. This condition here means that we clear the gradients from the audio-visual branch when the model performs well enough on the visual-only branch and much better than that on the audio-visual branch, which indicates a high likelihood of significant noise in the audio modality. Clearing gradients on the audio-visual branch when audio is noisy makes the model memorize valid information in the visual modality better.

4 EXPERIMENTS

4.1 Datasets and Metrics

We conduct our experiments on benchmark datasets for TSG task: Charades-STA [11] and ActivityNet Captions [19].

Charades-STA. Charades-STA [11] contains short videos about indoor activities. The videos are not post-edited and accompany the original soundtrack of the videos. We use 12,408 and 3,720 annotations for training and test split, respectively.

ActivityNet Captions. ActivityNet Captions [19] contains user-generated videos with much longer duration than Charades-STA. The Videos accompany audio but some of them are post-edited such as replacing the original soundtrack with background music.

Table 1: Performance (%) comparison on Charades-STA and ActivityNet Captions dataset. "w/o audio" means the model is trained without audio modality and "$_{\uparrow}$" denotes performance improvement when audio modality is introduced. Values highlighted by bold and underline represent the top-2 methods (Variants of "w/o audio" are not within the scope of comparison).

Method	Charades-STA				ActivityNet Captions			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
CTRL	-	23.63	8.89	-	-	29.01	10.34	-
ACRN	-	20.26	7.64	-	-	31.67	11.25	-
SCDM	-	54.44	33.43	-	54.80	36.75	19.86	-
BPNet	65.48	50.75	31.64	46.34	<u>58.98</u>	<u>42.07</u>	<u>24.69</u>	<u>42.11</u>
DEBUG	54.95	37.39	17.69	-	55.91	39.72	-	39.51
GDP	54.54	39.47	18.49	-	56.17	39.27	-	39.80
PfTML-GA	<u>67.53</u>	52.02	33.74	-	51.28	33.04	19.26	37.78
DRN	-	53.09	31.75	-	-	**42.49**	22.25	-
Moment-DETR	-	55.65	34.17	-	-	-	-	-
CPNET	-	**60.27**	<u>38.74</u>	<u>52.00</u>	-	40.56	21.63	40.65
PMI-LOC w/o audio	56.84	41.29	20.11	-	60.16	39.16	18.02	-
PMI-LOC	$58.08_{1.24\uparrow}$	$42.63_{1.34\uparrow}$	$21.32_{1.21\uparrow}$	-	$\mathbf{61.22}_{1.06\uparrow}$	$40.07_{0.91\uparrow}$	$18.29_{0.27\uparrow}$	-
UMT	-	48.31	29.25	-	-	-	-	-
ADPN w/o audio	70.35	55.32	37.47	51.13	55.72	39.56	25.20	41.55
ADPN	$\mathbf{71.99}_{1.64\uparrow}$	$\underline{57.69}_{2.37\uparrow}$	$\mathbf{41.10}_{3.63\uparrow}$	$\mathbf{52.86}_{1.73\uparrow}$	$57.16_{1.44\uparrow}$	$41.40_{1.84\uparrow}$	$\mathbf{26.31}_{1.11\uparrow}$	$\mathbf{42.31}_{0.76\uparrow}$

We follow the commonly adopted setup [59] for training/test partition. Actually, we use 33,721 and 15,753 annotations in our training and test sets for the absence of a number of videos on YouTube since we extract audio features from raw videos.

Metrics. We use "R{n}@{m}" (%) and "mIoU" (%) as our metrics. "R{n}@{m}" is defined as the percentage of queries having at least one result whose Intersection-over-Union (IoU) with ground truth is larger than m in the top-n recalled predictions. We use "R1@0.3", "R1@0.5" and "R1@0.7" in our experiments. "mIoU" is defined as average IoU with ground truth when testing.

4.2 Implementation Details

For textual queries, we use 300d GloVe [43] vectors as our initial word embeddings. For Charades-STA, we apply I3D [5] features for visual and PANN [18] features for audio. PANN is a network pretrained on AudioSet [12] dataset. For ActivityNet Captions, we apply C3D [50] features for visual and VGGish [15] features for audio, which are extracted by a VGG [48] network pretrained on YouTube-100M [15] dataset. We set the initial learning rate as 0.00015 and 0.0005 for Charades-STA and ActivityNet Captions and use AdamW [32] optimizer with linear learning rate decay and gradient clipping of 1.0; 3 semantic entities are extracted; η and λ are fixed on 0.3 and 25; we set N_e, $\overline{G}$ as 3, 0.3 and $\overline{G^v}$ as 0.25 and 0.5 for Charades-STA and ActivityNet Captions; we train the model for 300 epochs with batch size 32 and 64 for Charades-STA and ActivityNet Captions and adopt early stopping strategy. All experiments are implemented on a single NVIDIA TITAN X GPU.

4.3 Overall Performance

In Table 1, we evaluate our ADPN in two benchmark datasets and compare it with: (1) *Proposal-based*: CTRL [11], ACRN [30], SCDM [58], BPNet [56]. (2) *Proposal-free*: DEBUG [33], GDP [7], PfTML-GA [46], DRN [60], Moment-DETR [22], CPNET [23]. Particularly, we also compare our ADPN with UMT [31] and PMI-LOC [8], which incorporate audio for TSG solutions. Furthermore, we list the results for PMI-LOC and our ADPN when trained without audio, where we only train the visual-only branch for ADPN.

On Charades-STA, our ADPN achieves the best performance on most metrics. Furthermore, it's worth noting that our ADPN performs even better on harder metrics. We achieve a superior performance of 2.36% over CPNET on R1@0.7 although showing comparatively lower results on R1@0.5. And we outperform SCDM and Moment-DETR by 3.25% and 2.04% on R1@0.5, while 7.67% and 6.93% on R1@0.7. More significant improvement on R1@0.7 proves our ADPN excels in capturing subtle clues for precise moment retrieval. On ActivityNet Captions, our method still achieves comparable performance and reaches the best on R1@0.7 and mIoU, which is consistent with its performance on Charades-STA.

Compared with PMI-LOC and UMT, our ADPN shows higher performance improvement when audio is introduced, especially on harder metrics R1@0.5 and R1@0.7, where the improvement of our method is 176.87%, 300.00% for Charades-STA and 202.20%, 411.11% for ActivityNet Captions than that of PMI-LOC, indicating our ADPN can better leverage audio's potential. We observe more significant improvement on samples that are not as hard when using visual modality individually, leading us to consider audio's role more as *refining* other than *rectifying*. This implies that the model primarily attains better performance by refining the prediction from the correlations between audio and visual. Therefore, it's more important to pay attention to the audio-visual interaction than just utilize extra information from audio individually, since audio often contains sparse and noisy information and serves more as an auxiliary modality.

4.4 Ablation Studies

Table 2: Ablation Studies on Charades-STA. "$\cdot_{\uparrow}$" denotes the performance improvement of the audio-visual branch compared to the visual-only one within the same model.

Method	R1@0.3	R1@0.5	R1@0.7	mIoU
(1) ADPN w/ V-only	70.35	55.32	37.47	51.13
(2) ADPN w/ F-only	**72.02**	56.34	39.30	52.04
(3) ADPN w/o TG	70.62	<u>57.31</u>	39.41	51.74
(4) ADPN w/o TGCM	70.65	56.26	38.41	51.30
(5) ADPN w/o CL	70.67	56.72	<u>39.62</u>	<u>52.24</u>
(6) ADPN-V (m)	32.39	20.81	10.46	23.78
(7) ADPN-F (m)	$34.57_{2.18\uparrow}$	$22.34_{1.53\uparrow}$	$11.48_{1.02\uparrow}$	$25.22_{1.44\uparrow}$
(8) ADPN-V	70.81	56.72	39.46	51.68
(9) ADPN-F	$\underline{71.99}_{1.18\uparrow}$	$\mathbf{57.69}_{0.97\uparrow}$	$\mathbf{41.10}_{1.64\uparrow}$	$\mathbf{52.86}_{1.18\uparrow}$

We conduct ablation studies on Charades-STA to evaluate the crucial factors in our proposed ADPN, as shown in Table 2. Here we give some implementation details. **(1)** "w/ V-only": we only train the visual-only branch with $\ell_{loc} = \ell^v_{loc}$ and use the prediction of this branch during inference. **(2)** "w/ F-only": we only train the audio-visual branch with $\ell_{loc} = \ell^f_{loc}$ and use the prediction of this branch during inference. **(3)** "w/o TG": we remove the guidance of text in TGCM by replacing the text features $\mathbf{Q}^e$ with a randomly initialized learnable tensor $\mathbf{R}^e$ which has the same shape as $\mathbf{Q}^e$. **(4)** "w/o TGCM": we remove TGCM by making $\overline{\mathbf{X}}^f = \overline{\mathbf{X}}^a + \overline{\mathbf{X}}^v$ after re-sampling $\overline{\mathbf{X}}^a$. **(5)** "w/o CL": we remove our curriculum learning strategy of adaptive adjustment on the optimization process, i.e. $\mathbb{I}(G_s) \equiv 1$. **(6)** "-V (m)": we jointly train visual-only and audio-visual branches and use the prediction of the visual-only branch during inference. Particularly, we mask the ground truth part of the visual input by Gaussian noise with the same mean value and standard deviation as the corresponding visual features during inference. **(7)** "-F (m)": we do the same operation as (6), except we use the prediction of the audio-visual branch during inference. **(8)** "-V": we jointly train visual-only and audio-visual branches and use the prediction of the visual-only branch during inference. **(9)** "-F" **(our standard model)**: we do the same operation as (8), except we use the prediction of the audio-visual branch during inference.

Jointly-Training Strategy. Observing (1,2,8,9), the prediction accuracy of both visual-only and audio-visual branches improves significantly when trained jointly compared to when they are trained individually, verifying the validity of our jointly-training strategy. Taking one step further, (9) achieves 1.35%, 1.80%, and 0.82% performance improvement on R1@0.5, R1@0.7, and mIoU compared to (2), which is close to 1.02%, 1.83% and 0.91% from (1) to (2). It indicates our jointly-training strategy can alleviate the information gap between audio and visual and maintain more valid information in visual, thus further improving the performance from audio-visual collaboration, which is as important as just introducing audio modality.

Text-Guided Clues Miner (TGCM). Comparing (3) with (9), there are dramatic drops of 1.37%, 0.38% and 1.69% on R1@0.3, R1@0.5 and R1@0.7 when text guidance is removed, verifying its importance for more precise predictions. As shown in (4), without TGCM, performance degenerates even more on R1@0.5 and R1@0.7, indicating it's essential to discover and amplify the shared crucial locating clues within text, audio and visual modalities.

Curriculum Learning Strategy. Comparing (5) with (9), the performance increases by 1.32%, 0.97%, 1.48% and 0.62% on four metrics when our curriculum learning strategy is implemented. To further prove the effectiveness of our curriculum learning strategy, we conduct extra experiments on the two threshold hyperparameters $\overline{G^v}$ and $\overline{G}$, as shown in Figure 3. As we can see, the performance improvement is stable with suitable hyperparameters. Interestingly, insufficient (when $\overline{G^v}$ is high or $\overline{G}$ is small) or excessive (when $\overline{G^v}$ is small or $\overline{G}$ is high) adjustment in the optimization process both weaken the performance. Insufficient adjustment fails to fully suppress defective gradients from noise, while excessive adjustment discards some valid audio information in the early stage of training when the model is not trained well enough.

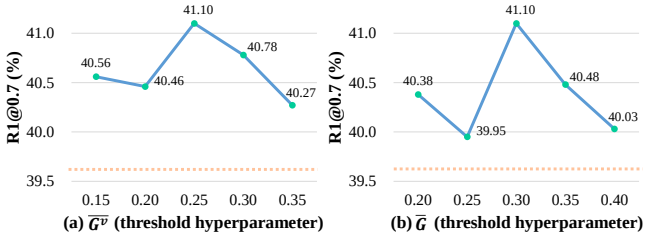

Figure 3: Ablation on different $\overline{G^v}$ and $\overline{G}$ of our curriculum learning strategy. The dashed line in orange denotes the baseline performance when the curriculum learning strategy is disabled. $\overline{G}$ is 0.3 in (a) and $\overline{G^v}$ is 0.25 in (b).

To further verify that our ADPN truly captures locating clues from audio, we design such an experiment where visual input is partially masked within the ground truth moment during inference and we observe the performance gap of the visual-only and audio-visual branches. As shown in (6)~(9) of Table 2, the audio-visual branch still outperforms the visual-only one despite the destruction of visual information, indicating our ADPN does capture complementary clues for localization from audio modality and works even without visual information in some scenarios.

The performances consistently drop by a similar amount when removing any of our crucial parts, suggesting that all our designs are equally indispensable for effective audio-visual joint learning. Key units can be individually and flexibly transferred to more common scenes when handling multiple diverse and unbalanced modalities. This is especially suitable for the jointly-training strategy and TGCM, which are not reliant on specific hyperparameters, significantly reducing the barriers to easily applying these techniques.

4.5 Qualitative Analysis

We conduct case studies and obtain some interesting findings, as shown in Figure 4. Audio-visual interaction helps generate accurate prediction especially when the query words prominently correlate with audio, *e.g.* "laugh" and "discuss". We visualize attention weight distributions on audio/visual features attended by text's semantic entities (*c.f.* Equation (14)), and we discover attention over

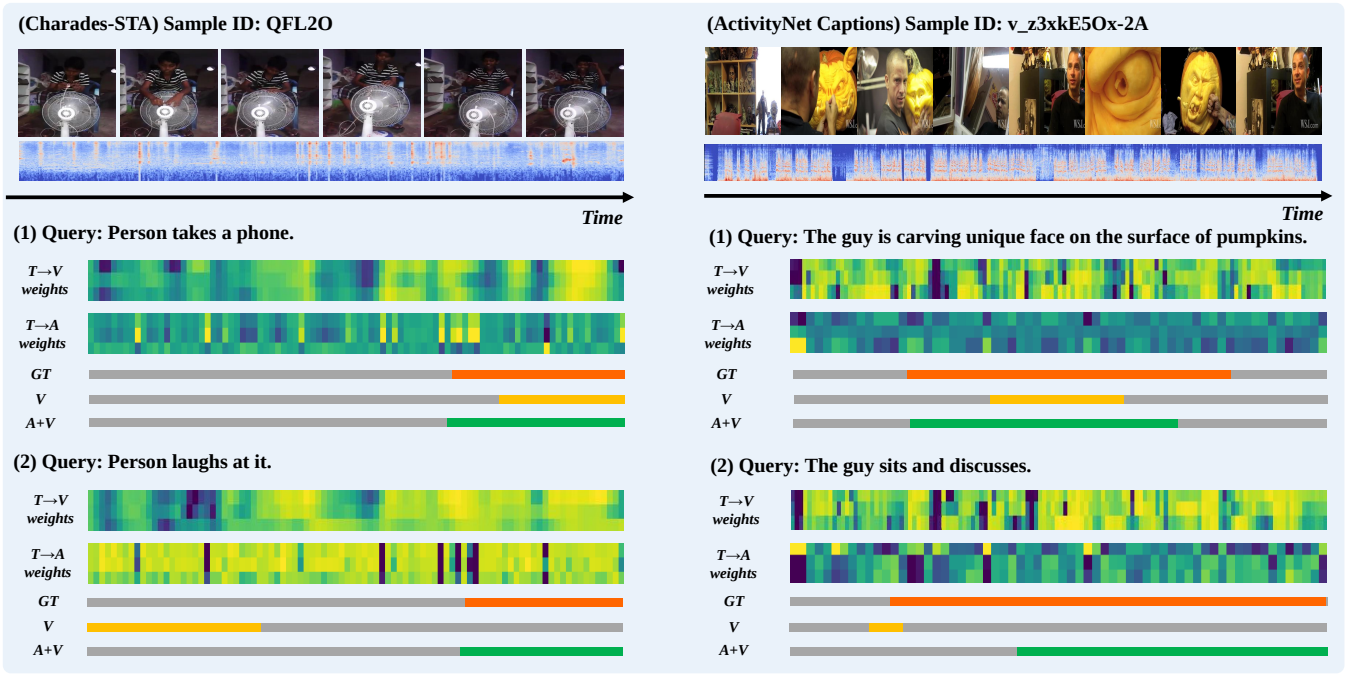

Figure 4: Sample results on Charades-STA (left) and ActivityNet Captions (right). Orange, yellow and green rectangular bars represent the ground truth, prediction of the visual-only branch, and prediction of the audio-visual branch. "T→V/A weights" show the attention weight distributions on visual/audio features for every semantic entity in TGCM. The darker, the more.

visual and audio are usually consistent (sample (1)) but sometimes attention on audio provides key complementary clues when visual collapses (sample (2)). For example, in the sample of "Person laughs at it", the model wrongly focuses more on the early part of the video from visual information, but pays more attention around the ground truth segment guided by audio and makes correct predictions by combining audio and visual.

To provide more interpretability, we visualize the weight distributions over query words when extracting semantic entity-level features (*c.f.* Equation (2)) for sample (2) on ActivityNet Captions, as shown in Figure 5. We discover that our model can capture fine-grained correlations between text and video. In the query "The guy sits and discusses", the first semantic entity pays the most attention to the word "discusses" and it guides the model to pay more attention to the area around the ground truth in audio modality compared to the other two semantic entities, which corrects the problem of poor attention in visual to some extend. This indicts our model does capture some correlations between the meaning of the word "discuss" and the audio signal of people's speech.

Figure 5: Weight distributions on words for every semantic entity in sample "v_z3xkE5Ox-2A"-(2). The darker, the more.

5 CONCLUSION

We introduce a novel Adaptive Dual-branch Promoted Network (ADPN) to solve Audio-enhanced Temporal Sentence Grounding (ATSG). We design a dual-branch pipeline to jointly train visual-only and audio-visual branches to fill the information gap between audio and visual, which outperforms any of the branches when it's trained individually. Furthermore, we propose a Text-Guided Clues Miner (TGCM) to model audio-visual interaction with text semantics as guidance, which is proven to benefit from the consistency and complementarity between audio and visual. Finally, we design a curriculum-based optimization strategy to further eliminate noises, where we evaluate the sample difficulty as a measure of noise intensity in a self-aware fashion and adjust the optimization process adaptively. We become the first to handle ATSG with real audio-awareness and our method achieves competitive performance in comparison with the state-of-the-art methods. In the future, we would like to build more suitable datasets for ATSG benchmarks to encourage more insightful research in this area.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China No. 2020AAA0106300, NSFC (No. 62250008, 62222209, 6210-2222), BNRist under Grant No. BNR2023-RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*. 5803–5812.

[2] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. 2022. Hear Me out: Fusional Approaches for Audio Augmented Temporal Action Localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2022, Volume 5: VISAPP, Online Streaming, February 6-8, 2022*. SCITEPRESS, 144–154. https://doi.org/10.5220/0010832700003124

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[4] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On Pursuit of Designing Multi-modal Transformer for Video Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 9810–9823. https://doi.org/10.18653/v1/2021.emnlp-main.773

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

[6] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. 2021. Curriculum disentangled recommendation with noisy multi-feedback. *Advances in Neural Information Processing Systems* 34 (2021), 26924–26936.

[7] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10551–10558.

[8] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. 2020. Learning modality interaction for temporal sentence localization and event captioning in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 333–351.

[9] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. 2021. End-to-end multi-modal video temporal grounding. *Advances in Neural Information Processing Systems* 34 (2021), 28442–28453.

[10] Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia* (2022).

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.

[12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[13] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *international conference on machine learning*. Pmlr, 1311–1320.

[14] Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *International conference on machine learning*. PMLR, 2535–2544.

[15] 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. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[16] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *International Conference on Machine Learning*. PMLR, 9938–9964.

[17] Xun Jiang, Xing Xu, Zhiguo Chen, Jingran Zhang, Jingkuan Song, Fumin Shen, Huimin Lu, and Heng Tao Shen. 2022. Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*. 719–727.

[18] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[20] M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems* 23 (2010).

[21] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. 2023. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2 (2023), 1–33.

[22] Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems* 34 (2021), 11846–11858.

[23] Kun Li, Dan Guo, and Meng Wang. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1902–1910.

[24] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, Shiliang Pu, and Fei Wu. 2022. End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 8707–8717. https://doi.org/10.18653/v1/2022.acl-long.596

[25] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-Attentive Sentence Embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=BJC_jUqxe

[26] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. 2023. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *IEEE Transactions on Multimedia* (2023).

[27] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2021. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 9292–9301. https://doi.org/10.18653/v1/2021.emnlp-main.732

[28] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4070–4078.

[29] Daizong Liu, Xiaoye Qu, Pan Zhou, and Yang Liu. 2022. Exploring motion and appearance information for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1674–1682.

[30] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 15–24.

[31] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3042–3051.

[32] Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR* abs/1711.05101 (2017). arXiv:1711.05101 http://arxiv.org/abs/1711.05101

[33] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. Debug: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5144–5153.

[34] Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022. Curriculum contrastive context denoising for few-shot conversational dense retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 176–186.

[35] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2020. Teacher-Student Curriculum Learning. *IEEE Trans. Neural Networks Learn. Syst.* 31, 9 (2020), 3732–3740. https://doi.org/10.1109/TNNLS.2019.2934906

[36] Deyu Meng, Qian Zhao, and Lu Jiang. 2017. A theoretical understanding of self-paced learning. *Information Sciences* 414 (2017), 319–328.

[37] Otniel-Bogdan Mercea, Thomas Hummel, A Sophia Koepke, and Zeynep Akata. 2022. Temporal and cross-modal attention for audio-visual zero-shot learning. In *European Conference on Computer Vision*. Springer, 488–505.

[38] Otniel-Bogdan Mercea, Lukas Riesch, A. Sophia Koepke, and Zeynep Akata. 2022. Audio-Visual Generalised Zero-Shot Learning With Cross-Modal Attention and Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10553–10563.

[39] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10810–10819.

[40] Wenwen Pan, Haonan Shi, Zhou Zhao, Jieming Zhu, Xiuqiang He, Zhigeng Pan, Lianli Gao, Jun Yu, Fei Wu, and Qi Tian. 2022. Wnet: Audio-Guided Video Object Segmentation via Wavelet-Based Cross-Modal Denoising Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1320–1331.

[41] Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. 2021. Adamml: Adaptive multi-modal

learning for efficient video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7576–7585.

[42] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8238–8247.

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[44] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. 2019. Competence-based Curriculum Learning for Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 1162–1172. https://doi.org/10.18653/v1/n19-1119

[45] Yuhan Quan, Jingtao Ding, Chen Gao, Lingling Yi, Depeng Jin, and Yong Li. 2023. Robust Preference-Guided Denoising for Graph based Social Recommendation. In *Proceedings of the ACM Web Conference 2023*. 1097–1108.

[46] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2464–2473.

[47] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=HJ0UKP9ge

[48] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[49] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. 2021. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3224–3234.

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[51] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured Multi-Level Interaction Network for Video Moment Localization via Language Query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7026–7035.

[52] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What Makes Training Multi-Modal Classification Networks Hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[53] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 4555–4576.

[54] Sangmin Woo, Jinyoung Park, Inyong Koo, Sumin Lee, Minki Jeong, and Changick Kim. 2022. Explore-And-Match: Bridging Proposal-Based and Proposal-Free With Transformer for Sentence Grounding in Videos. *arXiv preprint arXiv:2201.10168* (2022).

[55] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*. PMLR, 24043–24055.

[56] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2986–2994.

[57] Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic Coattention Networks For Question Answering. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rJeKjwvclx

[58] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems* 32 (2019).

[59] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9159–9166.

[60] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10287–10296.

[61] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1247–1257.

[62] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 6543–6554. https://doi.org/10.18653/v1/2020.acl-main.585

[63] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. 2021. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12669–12678.

[64] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12870–12877.

[65] Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739* (2018).

Table 3: Performance (mIoU) gain of the audio-visual branch compared to the visual-only branch on different kinds of activities on Charades-STA. We demonstrate the top- and bottom-20 ones.

Activity Category	mIoU Gain (%)
Throwing food somewhere	53.72
Laughing at television	49.77
Fixing a doorknob	38.15
Washing a window	34.97
Throwing a broom somewhere	27.61
Watching something/someone/ themselves in a mirror	22.76
Taking shoes from somewhere	17.27
Holding a picture	16.71
Putting a blanket somewhere	13.97
Tidying some clothes	12.34
Taking a blanket from somewhere	12.02
Holding some food	11.94
Putting a cup/glass/bottle somewhere	11.08
Washing a dish/dishes	10.70
Turning off a light	10.61
Throwing a book somewhere	10.18
Playing with a phone/camera	9.98
Throwing shoes somewhere	9.94
Holding some clothes	9.75
Putting a towel/s somewhere	9.63
...	...
Holding a box	-4.25
Sitting in a bed	-4.30
Sitting on the floor	-4.45
Taking a laptop from somewhere	-4.63
Sitting on sofa/couch	-4.84
Tidying up a table	-6.06
Putting a picture somewhere	-6.10
Closing a box	-6.57
Taking a box from somewhere	-7.31
Watching a laptop or something on a laptop	-7.67
Holding a bag	-9.54
Washing some clothes	-9.95
Taking a bag from somewhere	-11.80
Fixing a door	-12.72
Taking/consuming some medicine	-14.20
Holding a mirror	-14.25
Taking a dish/es from somewhere	-15.06
Working on paper/notebook	-17.31
Holding a dish	-18.91
Holding a vacuum	-42.96

A SUPPLEMENTARY MATERIAL

This material presents supplementary experiments on our proposed method, which consist of ablation studies on the ActivityNet Captions dataset (A.1), a qualitative analysis from the perspective of various activity categories (A.2) and a demonstration of more representative cases on Charades-STA for consistency and complementarity (A.3).

A.1 Ablation Studies on ActivityNet Captions

To further verify the general effectiveness of the crucial contributions in our proposed ADPN, we conduct supplementary ablation studies on more challenging ActivityNet Captions with the same settings of Charades-STA, as shown in Table 4.

Following the same analytical approach as before, we can draw similar conclusions to prove the validity of our jointly-training strategy, Text-Guided Clues Miner (TGCM) and curriculum optimization strategy. Nevertheless, several noteworthy points should be mentioned, which can inspire further discussion on our findings and encourage inspiring insights.

Our jointly-training strategy takes better effect for the audio-visual branch compared to the visual-only branch, especially on R1@0.5 and R1@0.7, emphasizing its importance for multimodal learning when handling the information gap of a dominant and a weak modality, i.e. visual and audio in our settings. Such a strategy can eliminate inter-modal interference without sacrificing uni-modal learning.

The performance gain still remains on R1@0.5 and R1@0.7 although the assistance of audio modality is weakened from (6)~(9) in Table 4. Audio accompanies more noisy information on ActivityNet Captions compared to Charades-STA. This indicates that audio generally provides complementary information on different datasets and implies that it's vital to exploit a noisy modality during its interaction with a cleaner one. Interestingly, the overall performance doesn't drop as dramatically as that on Charades-STA when the ground truth moment of visual features is masked during inference. Since the average duration of videos on ActivityNet Captions is much longer than that on Charades-STA, the ground truth moment on ActivityNet Captions is shorter in a relative sense. We speculate more unmasked visual features also provide more valid information for boundary prediction, thus it may be promising to explore context reasoning in ATSG for intra- and inter-modal to better utilize the contextual coherence of multiple modalities.

Table 4: Ablation Studies on ActivityNet Captions. "$_{\cdot\uparrow/\downarrow}$" means the performance gain of the audio-visual branch compared to the visual-only one when training jointly.

Method	R1@0.3	R1@0.5	R1@0.7	mIoU
(1) ADPN w/ V-only	55.72	39.56	25.20	41.55
(2) ADPN w/ F-only	**57.23**	40.67	25.69	<u>42.23</u>
(3) ADPN w/o TG	56.45	<u>41.01</u>	<u>26.01</u>	41.90
(4) ADPN w/o TGCM	56.26	40.15	25.58	41.65
(5) ADPN w/o CL	57.13	40.80	25.34	42.01
(6) ADPN-V (m)	52.80	38.15	24.54	39.26
(7) ADPN-F (m)	$52.75_{0.05\downarrow}$	$39.09_{0.94\uparrow}$	$25.02_{0.48\uparrow}$	$39.31_{0.05\uparrow}$
(8) ADPN-V	56.32	39.66	24.93	41.50
(9) ADPN-F	$\underline{57.16}_{0.84\uparrow}$	$\mathbf{41.40}_{1.74\uparrow}$	$\mathbf{26.31}_{1.38\uparrow}$	$\mathbf{42.31}_{0.81\uparrow}$

A.2 Category-Wise Analysis

To further analyze how our method benefits from audio modality, we conduct activity category-wise analysis on our dual-branch pipeline to observe the learning differences between the two branches in a fine-grained manner. In concrete, we use the activity category annotations on original Charades dataset and observe the average performance (mIoU) gain of the audio-visual branch compared to the visual-only one on each activity category. We totally capture 142 types of activities on 3,080 samples of 3,720 test samples. It is worth noting that such classification doesn't incorporate comprehensive information for the query-video pair and thus it doesn't seem rigorous but offers a straightforward and effective approach for intuitive analysis.

We observe that though the assistance from audio doesn't work in all scenarios, the audio-visual branch achieves much better performance on 88 of 142 activities. Especially, we demonstrate activity categories that exhibit the top- and bottom-20 performance gain in Table 3. As we can see, audio modality works well in activities of "throwing something", "laughing at something", "putting something" etc., which correlates natural audio signals intuitively, and the model refines its predictions with the assistance of audio modality. However, audio fails to boost performance in cases of "sitting", "taking something", "holding something" etc., since these activities have weak acoustic semantics such as "sitting" or the pattern across audio and visual is too ambiguous to learn for the model in activities like "taking or holding something". The learning differences between the two branches inspire that it's worth exploring to design modality selection strategies with confidence awareness of different modalities.

A.3 More Case Study on Charades-STA

To further show that consistency and complementarity can be deeply mined by our method, we demonstrate more cases on Charades-STA. These videos have varied scenarios, durations and moment temporal locations, which are fairly representative to some extend.

Please visit here for details with raw videos.