Rethinking Weakly-supervised Video Temporal Grounding From a Game Perspective

Xiang Fang[1*], Zeyu Xiong[1*], Wanlong Fang[1*], Xiaoye Qu[1], Chen Chen[2], Jianfeng Dong[3], Keke Tang[4], Pan Zhou[1✉], Yu Cheng[5], and Daizong Liu[6✉]

[1] Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2] University of Central Florida
[3] Zhejiang Gongshang University
[4] Guangzhou University
[5] The Chinese University of Hong Kong
[6] Peking University
xfang9508@gmail.com, zeyuxiong@hust.edu.cn, wanlongfang@gmail.com, xiaoye@hust.edu.cn, chen.chen@crcv.ucf.edu, dongjf24@gmail.com, tangbohutbh@gmail.com, panzhou@hust.edu.cn, chengyu@cse.cuhk.edu.hk, dzliu@stu.pku.edu.cn

Abstract. This paper addresses the challenging task of weakly-supervised video temporal grounding. Existing approaches are generally based on the moment proposal selection framework that utilizes contrastive learning and reconstruction paradigm for scoring the pre-defined moment proposals. Although they have achieved significant progress, we argue that their current frameworks have overlooked two indispensable issues: 1) Coarse-grained cross-modal learning: previous methods solely capture the global video-level alignment with the query, failing to model the detailed consistency between video frames and query words for accurately grounding the moment boundaries. 2) Complex moment proposals: their performance severely relies on the quality of proposals, which are also time-consuming and complicated for selection. To this end, in this paper, we make the first attempt to tackle this task from a novel game perspective, which effectively learns the uncertain relationship between each vision-language pair with diverse granularity and flexible combination for multi-level cross-modal interaction. Specifically, we creatively model each video frame and query word as game players with multivariate cooperative game theory to learn their contribution to the cross-modal similarity score. By quantifying the trend of frame-word cooperation within a coalition via the game-theoretic interaction, we are able to value all uncertain but possible correspondence between frames and words. Finally, instead of using moment proposals, we utilize the learned query-guided frame-wise scores for better moment localization. Experiments show that our method achieves superior performance on both Charades-STA and ActivityNet Caption datasets.

* Xiang Fang, Zeyu Xiong, and Wanlong Fang contributed equally to this work.
✉ Corresponding authors: Pan Zhou and Daizong Liu.

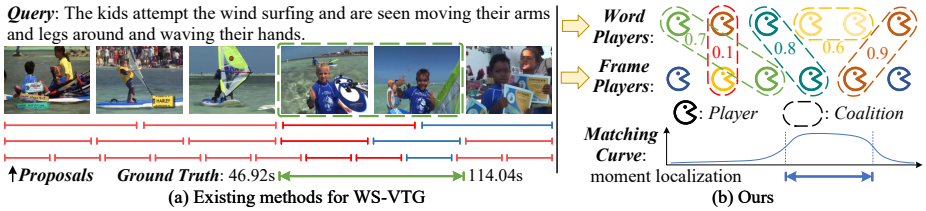

Fig. 1: (a) Most existing WS-VTG methods define moment proposals for query matching. They not only suffer from the coarse-grained alignment but also are limited to the quality of the proposals. (b) Instead of using proposals, our method reviews the WS-VTG from a brand-new game-interaction perspective to learn more fine-grained frame-aware self- and cross-modal alignment for accurate boundary localization.

1 Introduction

Video temporal grounding (VTG) aims to localize the start and end timestamps of a specific event moment described by a given language query in an untrimmed video [2,32,39,75–77,99–102,113,119,120]. This task enables us to efficiently find video moments of human interest instead of traversing the entire video, which has broad application potential in video surveillance [13–15, 18, 24–26, 50–52, 59, 90–93, 125, 126], video summarization [12, 16, 17, 21–23, 36, 48, 94, 106, 107], *etc.* Most VTG methods [27–31, 53–55, 57, 58, 60–62, 65, 66, 127] follow the fully-supervised setting, where each frame is manually annotated as query-relevant or not. Although these methods achieve significant breakthroughs, they severely rely on extensive laborious manual annotations of moment boundaries, thus limiting their scalability and practicability in real-world applications. To alleviate the dense reliance, weakly-supervised VTG (WS-VTG) has received increasing attention [10, 49, 73, 88, 89, 123], which only requires the knowledge of matched video-query pairs without detailed frame-wise annotations, which is more challenging than the fully-supervised setting.

Since there is no frame-wise annotation in WS-VTG, learning detailed frame-query alignment is difficult. As shown in Fig. 1(a), almost all existing WS-VTG solutions directly employ the proposal-based framework [10, 33, 73, 89] that first generates multiple moment proposals and then learns the scores to indicate the potential alignment between video proposals and language query. Some approaches further utilize reconstruction-based paradigm [49, 88, 123] to minimize the reconstruction loss between the partially masked query and moment proposals to identify the proposal which best reconstructs the query. Although exising WS-VTG methods have achieved great progress, we argue that they still suffer from two inescapable limitations: 1) Firstly, as frame-wise annotation is not provided, these WS-VTG methods generally generate the moment proposals for all videos via sliding windows. In this way, they only model the coarse-grained cross-modal interaction of each proposal-query pair. Actually, in most cases, we expect to capture more fine-grained information for accurate boundary localization, such as how much the semantics of a specific frame helps or harms the query-guided cross-modal alignment. Unfortunately, existing methods rely on proposal-level cross-modal learning cannot achieve this goal. 2) Secondly, the

performance of these proposal-based methods is severely limited by the quality of moment proposals as they devise moment proposals regardless of the specific contents and difficulty of each video.

To tackle the above limitations, in this paper, we propose to address the WS-VTG task from a new game-theory perspective, which creatively models the subtle video frames and query words as game players to learn their uncertain but possible cross-modal interaction with diverse granularity and flexible combination. The game learning helps to generate query-guided frame-aware knowledge for fine-grained video grounding. As shown in Fig. 1(b), if visual representations and textual representations have strong semantic correspondence, they tend to cooperate together and contribute to a high cross-modal similarity score. By forming these cooperated frame-word representations as a coalition and quantifying the trend of cooperation within a coalition via the game-theoretic interaction index, we can not only learn the coalition contribution to the cross-modal semantic similarity, but also measure the additional benefits brought by the coalition compared with the costs of the lost coalitions of these players with others. In this manner, our cooperative WS-VTG game is able to value all possible correspondence between frames and words for sensitive and explainable cross-modal contrast. By formulating these detailed cross-modal correspondences as fine-grained query-guided frame-wise scores, we can efficiently generate the moment instead of using complicated proposals. In particular, we implement the cooperative game in both self-modal and cross-modal scenarios for self-modal attentive learning and cross-modal fine-grained alignment. We further extend the cross-modal game with multi-level query-semantic interaction for more comprehensive video understanding.

Our main contributions are summarized as follows:

- We reveal the limitations of existing WS-VTG methods, which rely on complex moment proposals and fail to capture fine-grained frame-word alignment for accurate boundary grounding. To this end, we propose a novel game-theory based framework to learn the detailed frame-wise query-relevance scores for efficiently constructing the accurate moment.
- We apply the game learning to both self- and cross-modal scenarios for contextual self-enhancement and cross-alignment. The cross-modal game is also extended to multi-level for comprehensive learning.
- To verify the effectiveness of our framework, we conduct experiments on two challenging WS-VTG benchmarks, *i.e.*, Charades-STA and ActivityNet Caption. Experiments show that our method achieves superior performance. In representative cases, our method outperforms all compared methods by 4.53% on Charades-STA.

2 Related Works

Fully-supervised video temporal grounding. Existing methods generally address the video temporal grounding in a fully-supervised manner, where both the annotations of video-sentence pairs and corresponding moment boundaries

are given. Most of them [32, 35, 37, 56, 61, 64, 66, 95, 103, 110, 111, 114, 116, 121, 122] utilize the proposal-based framework, which first integrate the sentence representation with those pre-defined moment proposals individually, and then evaluate their matching relationships. The proposal with the highest matching score is selected as the prediction. Instead of utilizing the moment proposals, recent proposal-free methods [8,9,74,115] directly regress the temporal locations of the target moment.

Weakly-supervised video temporal grounding. As manually annotating temporal boundaries of target moments is time-consuming, recent research attentions shift to weakly-supervised video temporal grounding [10, 73, 88, 89], which only requires video-level annotations. Mithun *et al.* [73] proposed the first weakly-supervised model to learn a joint embedding space for video and query representations. Gao *et al.* [33] developed a two-stream structure to measure the moment-query consistency and conduct moment selection simultaneously. Although the above methods have achieved promising performance, they are two-stage approaches that utilize multi-scale sliding windows to generate moment candidates, therefore suffering from inferior effectiveness and efficiency. To address this issue, [49, 68, 123] score all the moments sampled at different scales in a single pass and further improve the moment-sentence matching accuracy. However, almost all of them rely on moment proposals for matching and selection, which fail to capture and distinguish more fine-grained details among visually similar frames for acquiring more accurate moment boundaries.

Cooperative game theory. Cooperative game theory focuses on the formation of coalitions and cooperation among rational players or groups to achieve common goals [1, 6, 40, 41, 47, 69, 80–82, 96]. A typical cooperative game consists of a set of players with a game function. The game function maps all possible subsets of players, called group or coalition, to a number which represents the total payoff earned by these players working cooperatively to achieve the goal [19, 20, 46, 84]. To measure the contributions of each player and allocate different payoffs to these individuals fairly, a few researchers have proposed various concepts of value by computing the average added worth that one player brings to all possible coalitions, such as Shapley value [72, 85, 108] and Banzhaf value [4, 45, 78]. Recently, some concepts of cooperative game theory have been interpreted in the field of deep learning [11, 67, 118]. For example, Li *et al.* [47] propose a fine-grained image-text semantic alignment pre-training framework based on Shapley interaction. Different from these works, we propose for the first time to model the detailed consistency between video frames and query words by recasting multivariate game-theoretic indices to address more complicated WS-VTG task.

3 Methodology

In this section, we elaborate on the proposed method, which learns the uncertain and fine-grained cross-modal alignment in WS-VTG task. Specifically, our method conceptualizes the interaction between the video and query as a game,

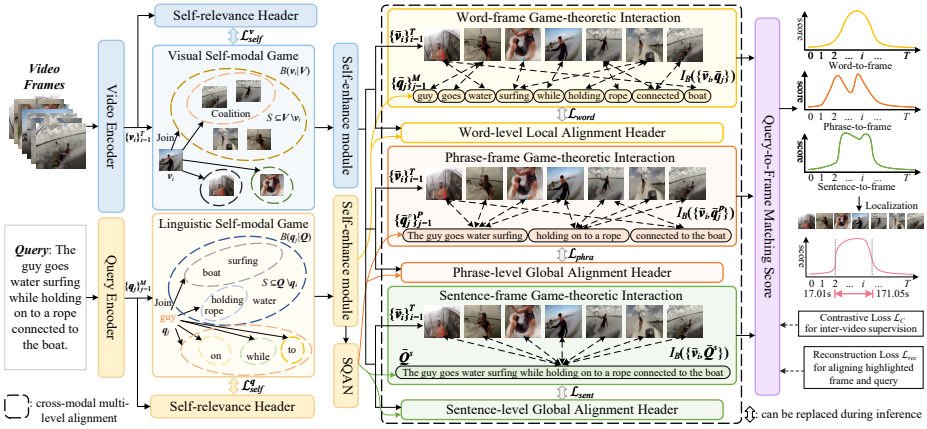

Fig. 2: Overview of our framework, where "SQAN" denotes Sequential Query Attention Network. Specifically, we first conduct self-modal game to enhance the self-modal instance semantics, then devise multi-level cross-modal games to learn the fine-grained and uncertain cross-modal correspondence. At last, we predict the moment based on the query-guided frame-wise scores.

where each frame/word acts as a player and forms coalitions. During the game learning, visual frame and linguistic word representations with strong semantic relevance will cooperate to form a new coalition, the benefits of which are quantified through the game interaction index. Our method maximizes payoff by strategically forming coalitions to achieve fine-grained semantic alignment between video and query. The overall framework is shown in Fig. 2.

3.1 Preliminaries

Problem formulation. Generally, given an untrimmed video and a sentence query, WS-VTG aims to localize the specific video moment semantically related to the query with only the video-level annotation. Let $V = \{v_1, ..., v_i..., v_T\}$ denote the untrimmed video with T frames, and $Q = \{q_1, ..., q_j, ..., q_M\}$ represent the sentence query containing M words, where v_i and q_j are the i-th frame and j-th word, respectively. The objective function $\mathcal{F}$ can be formulated as $\mathcal{F} : (V, Q) \rightarrow (\tau_s, \tau_e)$, where τ_s, τ_e denote the start and end timestamps of the video moment boundary in video V semantically corresponding to query Q.

Video and query representations. To extract contextual video representation, following [32,117], each input video V is fed into a pre-trained 3D convolutional network [5,97] with a stack of multi-head self-attention [98] layers for video encoding, then we get frame-wise video representation $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{T} \in \mathbb{R}^{T \times D_h}$, where D_h is the hidden dimension. For query representation, we also follow [59,117] to embed each word of the query by the Glove [83] model with a further multi-head self-attention module and a BiLSTM layer. Similarly, we can obtain the word-wise query representation $\boldsymbol{Q} = \{\boldsymbol{q}_j\}_{j=1}^{M} \in \mathbb{R}^{M \times D_h}$.

3.2 Game-theoretic Learning for Self-modal Semantic Enhancement

We argue that the instance relations within each modality are important to infer the consecutive visual event contents or linguistic phrase contexts. To achieve this self-modal learning, we review previous classic game-theoretic value [34, 85] to explore the weighted semantic contribution of each frame/word to the whole video/query semantics for enhancing the self-modal representation of both video and query.

Visual self-modal game. Generally, a cooperative game consists of a player set $\mathcal{P} = \{1, 2, ..., n\}$ and a game function $g(\cdot)$. In detail, function $g(\cdot)$ maps each player subset $\mathcal{S} \subseteq \mathcal{P}$ to a score value, which indicates the payoff when all players in coalition $\mathcal{S}$ work together in the game. To capture the inherent correlations within the video modality, we directly take the video frames $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{T}$ as frame players $\mathcal{P} = \boldsymbol{V}, n = T$ to play the game and evaluate frame-wise contribution to frame-level video semantics.

Since different frames contribute and weight differently to video understanding, we attempt to adopt a game-theoretic value to measure the importance of each frame player in the cooperative game. Specifically, each frame player $\boldsymbol{v}_i$ in player set $\boldsymbol{V}$ has a weighted marginal contribution to the entire game, which can be measured by the game-theoretic value (*e.g.,* Banzhaf value [34]) as:

$$B(\boldsymbol{v}_i|\boldsymbol{V}) = \frac{1}{2^{T-1}} \sum_{\mathcal{S} \subseteq \boldsymbol{V} \backslash \boldsymbol{v}_i} [g(\mathcal{S} \cup \{\boldsymbol{v}_i\}) - g(\mathcal{S})], \tag{1}$$

where $\mathcal{S} \subseteq \boldsymbol{V} \backslash \boldsymbol{v}_i$ represents a group formed without $\boldsymbol{v}_i$, $\frac{1}{2^{T-1}}$ is the likelihood of $\mathcal{S}$ being sampled, and $g(\mathcal{S} \cup \{\boldsymbol{v}_i\}) - g(\mathcal{S})$ denotes the marginal contribution brought by player $\boldsymbol{v}_i$ joining the coalition S. $B(\boldsymbol{v}_i|\boldsymbol{V})$ indicates the ability of player $\boldsymbol{v}_i$ to influence the video-level semantic by calculating the weighted marginal contribution of player $\boldsymbol{v}_i$ to all possible coalitions $\mathcal{S}$. Generally, within the context of our visual game, coalitions in the video modality represent consecutive frames of the same event or frames that are semantically similar. In this way, we can enumerate and learn all possible relations between the frames within the video for more contextual visual feature learning.

Visual soft supervision. To assist the above game-theoretic value learning, we first develop an additional learnable video-domain relevance header to predict the same self-modal relevance $\hat{R}^v = \{\hat{r}_i^v\}_{i=1}^{T} = \mathit{header}(\boldsymbol{V})$ as Banzhaf values $\{B(\boldsymbol{v}_i|\boldsymbol{V})\}_{i=1}^{T}$ among all frames for soft supervision (more discussions will be illustrated later). Specifically, the header is implemented by: 1) 1D convolutional layers for local self-consistency modeling, 2) a self-attention module for global self-relevance capturing, 3) a convolutional layer for decoding. Then, we optimize the Kullback-Leibler Divergence (KLD) between the predicted relevance $\hat{r}_i^v$ and the Banzhaf value $B(\boldsymbol{v}_i|\boldsymbol{V})$ for learning the consistent self-modal similarity. In detail, we define the game-guided frame-wise probability distribution as $D^v = [p_1^v, p_2^v, ..., p_i^v, ..., p_T^v]^\top$, where $p_i^v = \frac{\exp(B(\boldsymbol{v}_i|\boldsymbol{V}))}{\sum_{l=1}^{T} \exp(B(\boldsymbol{v}_l|\boldsymbol{V}))}$. Similarly, the predicted frame-wise probability distribution calculated by the output of the self-relevance header can be denoted as $\hat{D}^v = [\hat{p}_1^v, \hat{p}_2^v, ..., \hat{p}_i^v, ..., \hat{p}_T^v]^\top$, where

$\hat{p}_i^v = \frac{\exp(\hat{r}_i^v)}{\sum_{l=1}^{T}\exp(\hat{r}_l^v)}$. Finally, the self-relevance loss $\mathcal{L}_{self}^v$ for each video is:

$$\mathcal{L}_{self}^v = \mathrm{KLD}(\hat{D}^v||D^v) = -\frac{1}{T}\sum_{i=1}^{T} p_i^v(\log\hat{p}_i^v - \log p_i^v). \tag{2}$$

Based on the learned game-theoretic value $B(\boldsymbol{v}_i|\boldsymbol{V})$ for each player $\boldsymbol{v}_i$ in player set $\boldsymbol{V}$, we deploy a Softmax function to obtain the self-modal enhanced video features $\widetilde{\boldsymbol{v}}_i$ as:

$$\widetilde{\boldsymbol{v}}_i = \boldsymbol{v}_i + \sum_{\mathcal{S}\subseteq\boldsymbol{V}\setminus\boldsymbol{v}_i} \mathrm{Softmax}[g(\mathcal{S}\cup\{\boldsymbol{v}_i\}) - g(\mathcal{S})]\cdot(\mathcal{S}\cup\{\boldsymbol{v}_i\}). \tag{3}$$

The enhanced video features $\widetilde{\boldsymbol{V}} = \{\widetilde{\boldsymbol{v}}_i\}_{i=1}^T$ aggregates the relevant contexts from the whole video.

Linguistic self-modal game. Similarly, we deem query words $\boldsymbol{Q} = \{\boldsymbol{q}_j\}_{j=1}^M$ as players $\mathcal{P} = \boldsymbol{Q}, n = M$ and calculate each word's marginal contribution $B(\boldsymbol{q}_j|\boldsymbol{Q})$ to the whole query $\boldsymbol{Q}$ by formulating the linguistic self-modal game as:

$$B(\boldsymbol{q}_j|\boldsymbol{Q}) = \frac{1}{2^{M-1}}\sum_{\mathcal{S}\subseteq\boldsymbol{Q}\setminus\boldsymbol{q}_j}[g(\mathcal{S}\cup\{\boldsymbol{q}_j\}) - g(\mathcal{S})]. \tag{4}$$

Linguistic soft supervision. We first take the game probability distributions $D^q = [p_1^q, p_2^q, ..., p_j^q, ..., p_M^q]^\top$ as the soft label, and then supervise the predicted distributions $\hat{D}^q = [\hat{p}_1^q, \hat{p}_2^q, ..., \hat{p}_j^q, ..., \hat{p}_M^q]^\top$ of another header by KLD loss $\mathcal{L}_{self}^q$ similar to Eq. (2). At last, we can obtain the enhanced query features $\widetilde{\boldsymbol{Q}} = \{\widetilde{\boldsymbol{q}}_j\}_{j=1}^M$ by:

$$\widetilde{\boldsymbol{q}}_j = \boldsymbol{q}_j + \sum_{\mathcal{S}\subseteq\boldsymbol{Q}\setminus\boldsymbol{q}_j} \mathrm{Softmax}[g(\mathcal{S}\cup\{\boldsymbol{q}_j\}) - g(\mathcal{S})]\cdot(\mathcal{S}\cup\{\boldsymbol{q}_j\}). \tag{5}$$

Discussion on learnable header. Although we can directly utilize the game interaction value to enumerate the relations between different players via Eqs. (1) and (4), there are still two challenges: 1) No supervision signal: The performance of the game interaction severely depends on the quality of the learned player features. However, there is no supervision for learning the representative features. 2) Complex and time-consuming: The game process is complex and costs much time and resources. To this end, we develop a learnable header to mimic the game interaction by utilizing the KLD loss. This KLD function serves as a soft supervision label to not only learn the consistency between the game interaction and header, but also potentially train more distinguishing player features of each modality. During the inference, we can solely utilize the prediction of header to model the game interaction instead of Eqs. (1) and (4).

3.3 Game-theoretic Interaction for Cross-modal Multi-level Alignment

After obtaining the self-enhanced video and query features, we further develop a cross-modal game between frames and words to handle their uncertainty during their fine-grained semantic alignment.

Cross-modal game. From a game-theoretic perspective, we attempt to conceptualize the problem of fine-grained cross-modal alignment as a collaborative effort between video frames and query words. Specifically, when a video frame and a query word exhibit a high degree of semantic similarity, they are more likely to collaborate and their union will have a larger score contributed to the final moment. To this end, we take the enhanced video features $\widetilde{\boldsymbol{V}} = \{\widetilde{\boldsymbol{v}}_i\}_{i=1}^{T}$ and enhanced query features $\widetilde{\boldsymbol{Q}} = \{\widetilde{\boldsymbol{q}}_j\}_{j=1}^{M}$ as two different kinds of players $\mathcal{P} = \{\widetilde{\boldsymbol{v}}_i\}_{i=1}^{T} \cup \{\widetilde{\boldsymbol{q}}_j\}_{j=1}^{M}, n = T + M$. In a cooperative game, different players tend to work together in groups, called coalitions, to achieve a common goal, *i.e.*, representing the same semantics of the target event. In the case of two players $\widetilde{\boldsymbol{v}}_i$ and $\widetilde{\boldsymbol{q}}_j$ in set $\mathcal{P}$, it may occur that $g(\{\widetilde{\boldsymbol{v}}_i\})$ and $g(\{\widetilde{\boldsymbol{q}}_j\})$ are small individually, but at the same time the reward of their forming coalition $g(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\})$ is considerable. This is because players in the coalition interact and collaborate with each other, which may bring additional benefits (or costs) to the game. Therefore, we also need to measure the additional benefits brought by the coalition compared with the costs of the lost coalitions of these players with others. To form a better coalition $\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\} \subseteq \mathcal{P}$, we utilize the effective interaction index $I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\})$ like [34] to compute the additional benefit as follows:

$$I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\}) = \frac{1}{2^{n-2}} \sum_{\mathcal{S} \subseteq \mathcal{P} \backslash \{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\}} [g(\mathcal{S} \cup \{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\}) + g(\mathcal{S}) - g(\mathcal{S} \cup \{\widetilde{\boldsymbol{v}}_i\}) - g(\mathcal{S} \cup \{\widetilde{\boldsymbol{q}}_j\})], \quad (6)$$

where $\frac{1}{2^{n-2}}$ is the likelihood of $\mathcal{S} \subseteq \mathcal{P} \backslash \{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\}$ being sampled. Intuitively, $I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\})$ quantifies this additional benefit compared to when players $\widetilde{\boldsymbol{v}}_i$ and $\widetilde{\boldsymbol{q}}_j$ work independently. It also embodies the tendency of interaction between frame-player and word-player. A higher value of $I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\})$ signifies that players $\widetilde{\boldsymbol{v}}_i$ and $\widetilde{\boldsymbol{q}}_j$ cooperate closely with each other, and the formed coalition will bring additional high returns. By this interaction index, we can measure the closeness between plays $\widetilde{\boldsymbol{v}}_i$ and $\widetilde{\boldsymbol{q}}_j$ to achieve fine-grained alignment between video and query.

Game details. To ensure complete consistency between the cooperative game and cross-modal alignment learning, the game function g should meet the following criteria: 1) the game payoff benefits from strongly corresponding semantic pairs $\{\boldsymbol{v}_i^+, \boldsymbol{q}_j^+\}$, *i.e.*, $g(\mathcal{P}) - g(\mathcal{P} \backslash \{\boldsymbol{v}_i^+, \boldsymbol{q}_j^+\} \cup \{[\{\boldsymbol{v}_i^+, \boldsymbol{q}_j^+\}]\}) < 0$; 2) the game payoff is compromised by semantically opposite pairs $\{\boldsymbol{v}_i^-, \boldsymbol{q}_j^-\}$, *i.e.*, $g(\mathcal{P}) - g(\mathcal{P} \backslash \{\boldsymbol{v}_i^-, \boldsymbol{q}_j^-\} \cup \{[\{\boldsymbol{v}_i^-, \boldsymbol{q}_j^-\}]\}) > 0$; 3) when there are no players to cooperate, the payoff is zero, *i.e.*, $g(\{\boldsymbol{v}_i\}_{i=1}^{T}) = g(\{\boldsymbol{q}_j\}_{j=1}^{M}) = g(\phi) = 0$, where ϕ is the empty set. It should be noted that any function that satisfies the aforementioned conditions can be employed as the game function $g(\cdot)$. To simplify matters, we utilize the cosine similarity as g in all games.

During the game, when players $\boldsymbol{v}_i$ and $\boldsymbol{q}_j$ form a coalition $\{\boldsymbol{v}_i, \boldsymbol{q}_j\}$, we deem $[\{\boldsymbol{v}_i, \boldsymbol{q}_j\}]$ as a singular hypothetical player, which is the union of the player in $\{\boldsymbol{v}_i, \boldsymbol{q}_j\}$. Then, the reduced game is formed by removing the individual players in $\{\boldsymbol{v}_i, \boldsymbol{q}_j\}$ from the game and adding $[\{\boldsymbol{v}_i, \boldsymbol{q}_j\}]$ to the game.

Word-level alignment with local game interaction. Similar to the self-modal game, we develop an alignment header to predict the word-frame alignment matrix $A = [a_{i,j}]^{T \times M}$ between frame $\widetilde{\boldsymbol{v}}_i$ and word $\widetilde{\boldsymbol{q}}_j\}$, supervised by

$I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\})$. Then, we optimize the Kullback-Leibler Divergence (KLD) between the $I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\})$ and $a_{i,j}$, to bring the probability distributions of the output of the alignment header and interaction index close together to establish word-level semantic alignment between frame players and word players.

In detail, we define the word-to-frame probability distribution d^j_{w2f} of the j-th word-to-frame alignment as $d^j_{w2f} = [p_{1,j}, ..., p_{t,j}]$, where $p_{i,j} = \frac{\exp\left(I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\})\right)}{\sum_{l=1}^{T} \exp\left(I_B(\{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{q}}_j\})\right)}$, and the total query-to-video probability distribution $D_{q2v} = [d^1_{w2f}, ..., d^M_{w2f}]^\top$. Similarly, the predicted fine-grained probability distribution can be represented as $\hat{d}^j_{w2f} = [\hat{p}_{1,j}, ..., \hat{p}_{t,j}]$, where $\hat{p}_{i,j} = \frac{\exp\left(a_{i,j}\right)}{\sum_{l=1}^{t} \exp\left(a_{l,j}\right)}$, and the total predicted probability distribution is defined as $\hat{D}_{q2v} = [\hat{d}^1_{w2f}, ..., \hat{d}^M_{w2f}]^\top$. At last, the word-level alignment loss $\mathcal{L}_{word}$ is formulated as:

$$\mathcal{L}_{word} = \mathrm{KLD}(\hat{D}_{q2v} || D_{q2v}) = -\frac{1}{TM} \sum\nolimits_{i=1}^{T} \sum\nolimits_{j=1}^{M} p_{i,j} (\log \hat{p}_{i,j} - \log p_{i,j}). \tag{7}$$

Phrase- and sentence-level alignment with global game interaction. Although the above cross-modal game-theoretic interaction between each possible player-pair $\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\}$ addresses the problem of word-level alignment between word player $\widetilde{\boldsymbol{v}}_i$ and frame player $\widetilde{\boldsymbol{q}}_j$, it is still limited to aligning frames with isolated words, and fails to explore the relations between frames with more contextual phrase- or sentence-level semantics for better understanding temporal events. To this end, we introduce a multi-level game interaction to measure the semantic consistency between frames with different-grained textual semantics. Specifically, we follow [63, 74] to utilize Sequential Query Attention Network (SQAN) to extract the phrase-level textual semantics $\widetilde{\boldsymbol{Q}}^p = \{\widetilde{\boldsymbol{q}}^p_j\}_{j=1}^{P}$, and P is the phrase number. The sentence-level semantics $\widetilde{\boldsymbol{Q}}^s$ is obtained by concatenating the last hidden states in both the forward and backward BiLSTM. Similar to worl-level alignment, we compute both phrase- and sentence-level text-to-frame game interaction as $I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}^p_j\}), I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{Q}}^s\})$ via Eq. (6). We design corresponding alignment headers (phrase-level cross-modal header and sentence-level cross-modal header) to predict their alignment matrix via $\mathcal{L}_{phra}, \mathcal{L}_{sent}$.

3.4 Model Training and Inference

Weak supervision for video-query matching. We first interact each frame with the semantics of words, phrases and the whole sentence to calculate the overall query-to-frame matching degree by:

$$m_i = \frac{1}{3}(I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{Q}}^s\}) + \frac{1}{M} \sum\nolimits_{j=1}^{M} I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}_j\}) + \frac{1}{P} \sum\nolimits_{l=1}^{P} I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{q}}^p_l\})), \tag{8}$$

where m_i is the overall query-to-frame matching degree of i-th frame.

Then, the whole video-sentence matching score is calculated by:

$$\phi^c = \sum\nolimits_{i=1}^{T} w_i \cdot m_i, \; w_i = \frac{\exp\left(I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{Q}}^s\})\right)}{\sum_{i=1}^{T} \exp(I_B(\{\widetilde{\boldsymbol{v}}_i, \widetilde{\boldsymbol{Q}}^s\}))}, \tag{9}$$

where we utilize the sentence-level similarity to represent the contribution of each frame to the target moment among the entire video. For matched video V_k and query Q_k, we calculate the video-sentence matching score $\phi^c_{V_k,Q_k}$. The scores of unmatched pairs are expressed as $\phi^c_{V_k,Q_l}$ and $\phi^c_{Q_k,V_l}$ where $l \neq k$.

Finally, we utilize the cross-modal contrastive loss [79] for weak supervision:

$$\mathcal{L}_C = -\frac{1}{N_b}\sum_{k=1}^{N_b} \log \frac{\exp\left(\phi^c_{V_k,Q_k}/\tau\right)}{\sum_{l=1}^{N_b}\exp(\phi^c_{V_k,Q_l}/\tau)} - \frac{1}{N_b}\sum_{k=1}^{N_b} \log \frac{\exp\left(\phi^c_{Q_k,V_k}/\tau\right)}{\sum_{l=1}^{N_b}\exp(\phi^c_{Q_k,V_l}/\tau)}, \quad (10)$$

where τ is the temperature parameter, N_b is the batch size.

Reconstruction loss. To further enhance the semantic relevance of the highlighted frames to the sentence, we reconstruct the query conditioned on highlighted frames. Following popular reconstruction strategy [49,124], we randomly replace the 1/3 words in the original query with a specific symbol, and predict the next word by given a prefix of the query and highlighted frames feature. Specifically, we embed the masked query $\boldsymbol{Q}^m$ by Glove and predict the next word $\hat{\boldsymbol{q}}_{i+1}$ using the conditioned transformer following CNM [124] with input $(\widetilde{\boldsymbol{V}}, C_m, \boldsymbol{Q}^m_{1:i})$, where $C_m = \{m_1, m_2, ..., m_T\}$ represents the highlighted frame mask, $\boldsymbol{Q}^m_{1:i}$ denotes the masked query words from the 1-st to i-th.

Then, we introduce the cross-entropy loss $\mathcal{L}_{rec}$ to calculate the probability distribution $\mathcal{P}^q$ different between the predict word $\hat{\boldsymbol{q}}_{i+1}$ and the real word as:

$$\mathcal{L}_{rec} = -\sum_{i=1}^{M-1} \log \mathcal{P}^q(\boldsymbol{q}_{i+1}|\hat{\boldsymbol{q}}_{i+1}). \quad (11)$$

Overall training. To train the whole model, we combine game-guided self-relevance losses $\mathcal{L}_{self} = \mathcal{L}^v_{self} + \mathcal{L}^q_{self}$, game alignment loss $\mathcal{L}_{align} = \mathcal{L}_{word} + \mathcal{L}_{phra} + \mathcal{L}_{sent}$, video-level contrastive loss $\mathcal{L}_C$, and reconstruction loss $\mathcal{L}_{rec}$ with trade-off hyper-parameters ($\alpha_1, \alpha_2, \alpha_3$ and α_4) to define the overall loss as:

$$\mathcal{L} = \alpha_1\mathcal{L}_{self} + \alpha_2\mathcal{L}_{align} + \alpha_3\mathcal{L}_C + \alpha_4\mathcal{L}_{rec}. \quad (12)$$

Inference. During inference, we directly utilize the self-relevance headers to enhance the video and query features, and exploit the multi-level alignment headers to predict the overall query-to-frame matching degree. We locate the frame with the highest score as the basic predicted moment, and add the left/right frames into the moment if the ratio of their scores to the frame score of the closest moment boundary is higher than a threshold. We repeat this step until no frame can be added. We can obtain more moments by locating different initial frames with different scores.

4 Experiments

4.1 Datasets and Evaluation Metrics

Charades-STA. The Charades-STA dataset [32] is built based on the Charades [86], which contains 6,672 videos of indoor activities and involves 16,128 query-video pairs. There are 12,408 pairs used for training and 3,720 used for testing. The average duration of each video is 29.76 seconds.

ActivityNet Caption. The ActivityNet Caption dataset [43] contains 20,000 videos with 100,000 queries, where 37,421 query-video pairs are used for training and 34,536 are used for testing. On average, each video in ActivityNet Caption has 3.65 annotated moments and each annotated moment lasts for 36 seconds. For fair comparison, we follow [10, 73, 88, 89, 123, 124] to use val1 as valuation and report the test performance on val2.
Evaluation metrics. Following previous works, we adopt the metrics "R@n, IoU=m" to evaluate our model, which presents the proportion of the top n predicted moments with IoU larger than m.

4.2 Implementation Details

For fair comparison, we follow previous works to apply C3D [97] to encode the videos on ActivityNet Caption and I3D [5] on Charades-STA. We uniformly downsample the length of video feature sequences to $T = 200$ for ActivityNet Caption and $T = 64$ for Charades-STA. As for query encoding, we also follow previous works to utilize set the length of word feature sequences to $M = 20$, and utilize Glove [83] to embed each word. The dimension D_h is set to 512. These trade-off hyper-parameters are set to $\alpha_1 = 1.0, \alpha_2 = 1.0, \alpha_3 = 1.0, \alpha_4 = 10$. Since the calculation of the exact game-theoretic interaction is an NP-hard problem [71], we follow previous works to utilize sampling-based method [3, 44] to obtain unbiased estimates for approximating it. We train the whole model for 100 epochs with batch size of 16 and early stopping strategy. Parameter optimization is performed by Adam [42] optimizer with learning rate 3×10^{-4}, and linear decay of learning rate and gradient clipping of 1.0. The inference threshold is set to 0.8 in ActivityNet Caption and 0.9 in Charades-STA.

4.3 Main Results

Quantitative comparison. We compare our method with the state-of-the-art methods, including *fully-supervised* (FS) methods and *weakly-supervised* (WS) methods. Best results are in **bold**. We introduce two types of cooperative game indices, *i.e.*, Banzhaf index [34] and Shapley index [85], to our model. As summarized in Tab. 1, both two different variants of our method surpass all existing methods on both Charades-STA and ActivityNet Captions. Particularly, on Charades-STA in terms of "R@5, IoU=0.5", Ours(Banzhaf) outperforms the best compared method CPL by 4.53%. The main reason is that our game-theoretic-based approach can learn the local alignment between words and frames for accurately predicting the moment boundaries. Besides, the Shapley variant performs similar to the Banzhaf variant for this specific WS-VTG task, since they share the same game interaction formulation with different normalization. Overall, this table verifies the effectiveness of our game-based framework.
Generalization on cross-dataset evaluation. At last, we explore the generalization of models by cross-dataset evaluation in Tab. 2. Concretely, we first train a model on one dataset and then evaluate its performance on the other dataset. We can find that: 1) For all WS-VTG methods, the performance of

Table 1: Comparisons with state-of-the-art methods. "WS" and "FS" denotes weakly-supervised and fully-supervised methods, respectively.

Type	Method	Charades-STA				ActivityNet Caption			
		R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
FS	CTRL [32]	23.63	8.89	58.92	29.52	-	29.01	-	59.17
	2DTAN [121]	39.81	23.25	79.33	51.15	59.45	44.51	85.53	77.13
	DRN [115]	53.09	31.75	89.06	60.05	-	45.45	-	77.97
WS	SCN [49]	23.58	9.97	71.80	38.87	47.23	29.22	71.45	55.69
	WSTAN [104]	29.35	12.28	76.13	41.53	52.45	30.01	79.38	63.42
	ICVC [7]	31.02	16.53	77.53	41.91	46.62	29.52	80.92	66.61
	MARN [87]	33.87	15.54	73.90	41.94	48.52	31.37	75.91	60.00
	CRM [38]	34.76	16.37	-	-	55.26	32.19	-	-
	CNM [124]	35.15	14.95	-	-	55.68	33.33	-	-
	ACN [109]	37.02	15.26	-	-	57.66	34.18	-	-
	VCA [105]	38.13	19.57	78.75	37.75	50.45	31.00	71.79	53.83
	LCNet [112]	39.19	18.17	80.56	45.24	48.49	26.33	82.51	62.66
	CPL [125]	49.24	22.39	84.71	52.37	55.73	31.37	63.05	43.13
	DM2 [70]	51.39	23.72	-	-	54.73	31.85	-	-
	Ours(Banzhaf)	54.17	**26.46**	88.93	**55.79**	58.94	35.65	86.41	**70.26**
	Ours(Shapley)	**54.31**	26.28	**89.24**	55.62	**59.17**	**35.80**	**86.59**	70.03

Table 2: Cross-dataset evaluation. "$A \rightarrow C$" denotes training on ActivityNet Caption and evaluation on Charades-STA, and vice versa.

Method	$A \rightarrow C$		$C \rightarrow A$	
	R@1, IoU=0.7	R@5, IoU=0.7	R@1, IoU=0.5	R@5, IoU=0.5
LCNet [112]	12.06	31.47	10.65	31.76
CPL [125]	14.32	35.81	11.59	29.88
Ours(Banzhaf)	**18.74**	42.58	**20.37**	**42.96**
Ours(Shapley)	18.26	**43.13**	20.04	41.42

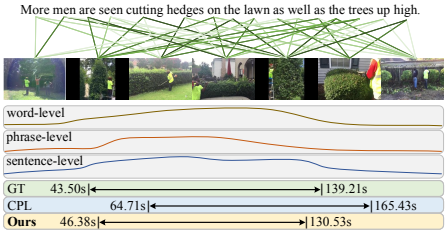

Fig. 3: Qualitative results. The colors of green lines represent different degrees of confidence.

A → C achieves better than C → A. This is because that ActivityNet Caption is larger and more complex than Charades-STA, thus the former is able to bring more generalized knowledge to the latter while the latter fails to provide general knowledge to the former. 2) Our two variants outperform state-of-the-art methods (*e.g.*, LCNet) with clear margins on cross-dataset evaluation. We speculate it to: Previous works severely rely on the moment proposals, however, LCNet defines different sliding windows on different datasets for proposal generation. Therefore, its non-uniform proposal structure limits its generalization on cross-dataset evaluation. Instead, we learn multi-level frame-aware alignment and utilize frame-wise scores for predicting moment, which is more flexible and adaptable, thus improving the generalization-ability.

Qualitative comparison. In Fig. 3, we provide a qualitative example on ActivityNet Caption. The colors of green lines represent different degrees of interaction confidence, demonstrating that the game interaction can well learn the uncertain and fine-grained frame-word correspondence in the WS setting. Besides, all three-level interaction (word-level, phrase-level and sentence-level) contributes

Table 3: Comparison on speed per sample (s) and GPU memory cost (MB).

Metric	Stage	LCNet [112]	CPL [125]	Ours(Banzhaf)	Ours(Shapley)
Speed	Training	0.79s	**0.27s**	0.94s	0.98s
	Inference	0.62s	0.21s	**0.10s**	0.11s
GPU Memory	Training	8647MB	**4156MB**	5237MB	5482MB
	Inference	7462MB	3428MB	**2058MB**	2165MB

Table 4: Main ablation study.

Self-modal Game		Multi-level Cross-modal Game			Banzhaf-based Game Charades-STA		Banzhaf-based Game ActivityNet Caption		Shapley-based Game Charades-STA		Shapley-based Game ActivityNet Caption	
Video Game	Query Game	Word-level	Phrase-level	Sentence-level	R@1, IoU=0.7	R@5, IoU=0.7	R@1, IoU=0.5	R@5, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.7	R@1, IoU=0.5	R@5, IoU=0.5
×	×	✓	×	×	19.34	44.67	29.83	61.72	19.29	44.31	30.15	61.56
✓	×	✓	×	×	21.25	47.86	31.39	63.98	21.10	47.73	31.54	63.67
✓	✓	✓	×	×	22.97	50.64	32.72	65.80	22.69	50.38	32.97	65.65
✓	✓	✓	✓	×	24.81	53.28	34.46	68.03	24.64	53.03	34.59	67.82
✓	✓	✓	✓	✓	**26.46**	**55.79**	**35.65**	**70.26**	**26.28**	**55.62**	**35.80**	**70.03**

differently to the grounding, and our final method is able to localize more accurate moment boundaries than state-of-the-art method CPL.

Complexity comparison. In Tab. 3, our method costs more training time/memory as we need to enumerate coalition samples. Specifically, we sampled 5500 samples. We find that, the model instability decreases along with the increase of the sampling number, and when the sample size ranges from 4500 to 5500, the error with the true value of banzhaf gradually decreases from 6.9% to 1.1%. After exceeding 5500 samples, more samples will not affect the result. During inference, since we can directly utilize the learned game headers to generate game value instead of using complex game learning, our model is faster with lower memory.

4.4 Ablation Study

In this section, we perform in-depth ablation studies to evaluate the effect of each component. We implement our model with different game settings.

Main ablation. As shown in Tab. 4, we conduct ablation studies regarding the components (*i.e.*, self-modal games in two modalities, and multi-level cross-modal games). We observe the following findings: 1) Our baseline model (only contains word-level cross-modal game) is powerful and outperforms some WS methods, demonstrating that it is effective to capture the uncertain yet fine-grained alignment for better determining the boundaries. 2) Both video and query self-modal games bring significant improvement to the overall model, indicating that self-enhancement is crucial for context integration. 3) Each query level of cross-modal game further boost the performance, verifying that phrase- and sentence-level semantics help to better understand temporal events. Overall, all components contribute a lot to final performance, showing their effectiveness.

Effect of game headers. The headers in both self- and cross-modal games are utilized to provide soft supervision for better game interaction and player-representation learning. To explore the impact of the structure of these headers, we compare several popular structures in Tab. 5. We find that the combination of

Table 5: Ablation study on each component, where "LL" means "Linear Layer", "CL" means "Convolution Layer", "SA" means "Self-attention"; "MSE" means "Mean Squared Error", "CE" means "Cross Entropy", "KLD" means "Kullback-Leibler Divergence"; "RL" means "Reconstruction Loss".

Component	Setting	Banzhaf-based Game				Shapley-based Game			
		Charades-STA		ActivityNet Caption		Charades-STA		ActivityNet Caption	
		R@1, IoU=0.7	R@5, IoU=0.7	R@1, IoU=0.5	R@5, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.7	R@1, IoU=0.5	R@5, IoU=0.5
Game Headers	LL	25.28	53.90	34.15	68.03	24.97	53.68	34.25	67.94
	CL	25.52	54.31	34.56	68.64	25.26	54.03	34.82	68.47
	LL + SA	25.94	54.83	35.27	69.18	25.71	54.59	35.53	69.06
	CL + SA	**26.46**	**55.79**	**35.65**	**70.26**	**26.28**	**55.62**	**35.80**	**70.03**
Soft Supervision	None	23.21	51.64	32.36	66.40	22.99	51.45	32.71	66.18
	MSE	24.42	53.19	33.85	67.57	24.24	52.89	40.06	67.33
	CE	24.73	53.58	34.16	68.04	24.52	53.26	34.37	67.80
	KLD	**26.46**	**55.79**	**35.65**	**70.26**	**26.28**	**55.62**	**35.80**	**70.03**
Grounding Losses	w/o RL	24.35	53.57	34.14	68.42	24.08	53.31	34.45	68.19
	w/ RL	**26.46**	**55.79**	**35.65**	**70.26**	**26.28**	**55.62**	**35.80**	**70.03**

convolution layer and self-attention can capture both local and global interaction, so it achieves the best performance and is beneficial for cross-modal multi-level interaction, illustrating the significance of our game headers.

Necessity of soft supervision. To explore the effect of soft supervision, we conduct the ablation study to investigate the contribution of the KLD loss during the header learning in Tab. 5. The KLD loss achieves better performance than both MSE and CE, showing that it is more suitable for game headers to mimic game-theoretic interaction learning. Moreover, without soft supervision, the performance will drop a lot, showing the importance of the soft supervision.

Analysis on reconstruction loss. Most previous WS-VTG works also additionally utilize the reconstruction paradigm to assist the model training. As shown in Tab. 5, we further conduct such ablation study to investigate the impact of reconstruction loss. It shows that the reconstruction loss can improve significantly the performance, demonstrating its effectiveness.

5 Conclusion

In this paper, we rethink the limitations of previous weakly-supervised video temporal grounding works, and creatively propose to utilize the game-theoretic interaction to learn the uncertain relationship between video-query pairs with diverse granularity. Specifically, we first introduce two self-modal games to correlate the frames/words to enhance their contextual semantics. Then, we propose a cross-modal game to value fine-grained correspondence between frames and words with multiple levels for better determining the query-related moment boundaries. Extensive experiments demonstrate the effectiveness of our game-based framework on two challenging datasets.

References

1. Albarelli, A., Rodola, E., Torsello, A.: A game-theoretic approach to fine surface registration without initial motion estimation. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 430–437. IEEE (2010)
2. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5803–5812 (2017)
3. Bachrach, Y., Markakis, E., Resnick, E., Procaccia, A.D., Rosenschein, J.S., Saberi, A.: Approximating power indices: theoretical and empirical analysis. Autonomous Agents and Multi-Agent Systems **20**, 105–122 (2010)
4. Banzhaf III, J.F.: Weighted voting doesn't work: A mathematical analysis. Rutgers L. Rev. **19**, 317 (1964)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Chalkiadakis, G., Elkind, E., Wooldridge, M.: Computational aspects of cooperative game theory. Synthesis Lectures on Artificial Intelligence and Machine Learning **5**(6), 1–168 (2011)
7. Chen, J., Luo, W., Zhang, W., Ma, L.: Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 267–275 (2022)
8. Chen, J., Ma, L., Chen, X., Jie, Z., Luo, J.: Localizing natural language in videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8175–8182 (2019)
9. Chen, L., Lu, C., Tang, S., Xiao, J., Zhang, D., Tan, C., Li, X.: Rethinking the bottom-up framework for query-based video localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10551–10558 (2020)
10. Chen, Z., Ma, L., Luo, W., Tang, P., Wong, K.Y.K.: Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. arXiv preprint arXiv:2001.09308 (2020)
11. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE symposium on security and privacy. pp. 598–617. IEEE (2016)
12. Deng, S., Wen, J., Liu, C., Yan, K., Xu, G., Xu, Y.: Projective incomplete multiview clustering. IEEE Transactions on Neural Networks and Learning Systems pp. 1–13 (2023). https://doi.org/10.1109/TNNLS.2023.3242473
13. Dong, J., Chen, X., Zhang, M., Yang, X., Chen, S., Li, X., Wang, X.: Partially relevant video retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 246–257 (2022)
14. Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia **20**(12), 3377–3388 (2018)
15. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4065–4080 (2022)
16. Dong, J., Peng, X., Ma, Z., Liu, D., Qu, X., Yang, X., Zhu, J., Liu, B.: From region to patch: Attribute-aware foreground-background contrastive learning for fine-grained fashion retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1273–1282 (2023)

17. Dong, J., Sun, S., Liu, Z., Chen, S., Liu, B., Wang, X.: Hierarchical contrast for unsupervised skeleton-based action representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 525–533 (2023)
18. Dong, J., Wang, Y., Chen, X., Qu, X., Li, X., He, Y., Wang, X.: Reading-strategy inspired visual representation learning for text-to-video retrieval. IEEE Transactions on Circuits and Systems for Video Technology **32**(8), 5680–5694 (2022)
19. Donoser, M., Bischof, H.: Diffusion processes for retrieval revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1320–1327 (2013)
20. Dowdall, J., Pavlidis, I.T., Tsiamyrtzis, P.: Coalitional tracking in facial infrared imaging and beyond. In: 2006 Conference on Computer Vision and Pattern Recognition Workshop. pp. 134–134. IEEE (2006)
21. Fang, X., Easwaran, A., Genest, B.: Uncertainty-guided appearance-motion association network for out-of-distribution action detection. In: IEEE International Conference on Multimedia Information Processing and Retrieval (2024)
22. Fang, X., Fang, W., Liu, D., Qu, X., Dong, J., Zhou, P., Li, R., Xu, Z., Chen, L., Zheng, P., Cheng, Y.: Not all inputs are valid: Towards open-set video moment retrieval using language. In: Proceedings of the 32th ACM International Conference on Multimedia (2024)
23. Fang, X., Hu, Y.: Double self-weighted multi-view clustering via adaptive view fusion. arXiv preprint arXiv:2011.10396 (2020)
24. Fang, X., Hu, Y., Zhou, P., Wu, D.: Animc: A soft approach for autoweighted noisy and incomplete multiview clustering. IEEE Transactions on Artificial Intelligence **3**(2), 192–206 (2021)
25. Fang, X., Hu, Y., Zhou, P., Wu, D.O.: V3h: View variation and view heredity for incomplete multiview clustering. IEEE Transactions on Artificial Intelligence **1**(3), 233–247 (2020)
26. Fang, X., Hu, Y., Zhou, P., Wu, D.O.: Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. IEEE Transactions on Emerging Topics in Computational Intelligence **6**(4), 913–927 (2021)
27. Fang, X., Liu, D., Fang, W., Zhou, P., Cheng, Y., Tang, K., Zou, K.: Annotations are not all you need: A cross-modal knowledge transfer network for unsupervised temporal sentence grounding. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 8721–8733 (2023)
28. Fang, X., Liu, D., Fang, W., Zhou, P., Xu, Z., Xu, W., Chen, J., Li, R.: Fewer steps, better performance: Efficient cross-modal clip trimming for video moment retrieval using language. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1735–1743 (2024)
29. Fang, X., Liu, D., Zhou, P., Hu, Y.: Multi-modal cross-domain alignment network for video moment retrieval. IEEE Transactions on Multimedia **25**, 7517–7532 (2022)
30. Fang, X., Liu, D., Zhou, P., Nan, G.: You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2448–2460 (2023)
31. Fang, X., Liu, D., Zhou, P., Xu, Z., Li, R.: Hierarchical local-global transformer for temporal sentence grounding. IEEE Transactions on Multimedia (2023)
32. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5267–5275 (2017)

33. Gao, M., Davis, L., Socher, R., Xiong, C.: Wslln: Weakly supervised natural language localization networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 1481–1487 (2019)
34. Grabisch, M., Roubens, M.: An axiomatic approach to the concept of interaction among players in cooperative games. International Journal of game theory **28**, 547–565 (1999)
35. Guo, C., Liu, D., Zhou, P.: A hybird alignment loss for temporal moment localization with natural language. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2022)
36. Guo, D., Li, K., Hu, B., Zhang, Y., Wang, M.: Benchmarking micro-action recognition: Dataset, method, and application. IEEE Transactions on Circuits and Systems for Video Technology **34**(7), 6238–6252 (2024)
37. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with temporal language. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing p. 1380–1390 (2018)
38. Huang, J., Liu, Y., Gong, S., Jin, H.: Cross-sentence temporal and semantic relations in video activity localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7199–7208 (2021)
39. Jiang, L., Wang, C., Ning, X., Yu, Z.: Lttpoint: A mlp-based point cloud classification method with local topology transformation module. In: 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT). pp. 783–789. IEEE (2023)
40. Jin, P., Huang, J., Xiong, P., Tian, S., Liu, C., Ji, X., Yuan, L., Chen, J.: Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2472–2482 (2023)
41. Jin, S., Wang, S., Fang, F.: Game theoretical analysis on capacity configuration for microgrid based on multi-agent system. International Journal of Electrical Power & Energy Systems **125**, 106485 (2021)
42. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
43. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 706–715 (2017)
44. Leech, D.: Computation of power indices (2002)
45. Lehrer, E.: An axiomatization of the banzhaf value. International Journal of Game Theory **17**, 89–99 (1988)
46. Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., Zou, Y.: G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12032–12042 (2023)
47. Li, J., HE, X., Wei, L., Qian, L., Zhu, L., Xie, L., Zhuang, Y., Tian, Q., Tang, S.: Fine-grained semantically aligned vision-language pre-training. In: Advances in Neural Information Processing Systems (2022)
48. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2804 (2023)

49. Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H.: Weakly-supervised video moment retrieval via semantic completion network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11539–11546 (2020)
50. Liu, C., Wen, J., Luo, X., Huang, C., Wu, Z., Xu, Y.: Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 8807–8815 (2023)
51. Liu, C., Wen, J., Luo, X., Xu, Y.: Incomplete multi-view multi-label learning via label-guided masked view- and category-aware transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 8816–8824 (2023)
52. Liu, C., Wen, J., Wu, Z., Luo, X., Huang, C., Xu, Y.: Information recovery-driven deep incomplete multiview clustering network. IEEE Transactions on Neural Networks and Learning Systems pp. 1–11 (2023)
53. Liu, D., Fang, X., Hu, W., Zhou, P.: Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. IEEE Transactions on Multimedia **25**, 8539–8553 (2023)
54. Liu, D., Fang, X., Qu, X., Dong, J., Yan, H., Yang, Y., Zhou, P., Cheng, Y.: Unsupervised domain adaptative temporal sentence localization with mutual information maximization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3567–3575 (2024)
55. Liu, D., Fang, X., Zhou, P., Di, X., Lu, W., Cheng, Y.: Hypotheses tree building for one-shot temporal sentence localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1640–1648 (2023)
56. Liu, D., Hu, W.: Learning to focus on the foreground for temporal sentence grounding. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 5532–5541 (2022)
57. Liu, D., Hu, W.: Skimming, locating, then perusing: A human-like framework for natural language video localization. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4536–4545 (2022)
58. Liu, D., Qu, X., Dong, J., Nan, G., Zhou, P., Xu, Z., Chen, L., Yan, H., Cheng, Y.: Filling the information gap between video and query for language-driven moment retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4190–4199 (2023)
59. Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y.: Context-aware biaffine localizing network for temporal sentence grounding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11235–11244 (2021)
60. Liu, D., Qu, X., Dong, J., Zhou, P., Xu, Z., Wang, H., Di, X., Lu, W., Cheng, Y.: Transform-equivariant consistency learning for temporal sentence grounding. ACM Transactions on Multimedia Computing, Communications and Applications **20**(4), 1–19 (2024)
61. Liu, D., Qu, X., Fang, X., Dong, J., Zhou, P., Nan, G., Tang, K., Fang, W., Cheng, Y.: Towards robust temporal activity localization learning with noisy labels. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 16630–16642 (2024)
62. Liu, D., Qu, X., Hu, W.: Reducing the vision and language bias for temporal sentence grounding. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4092–4101 (2022)

63. Liu, D., Qu, X., Liu, X.Y., Dong, J., Zhou, P., Xu, Z.: Jointly cross-and self-modal graph attention network for query-based moment localization. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4070–4078 (2020)
64. Liu, D., Zhou, P.: Jointly visual-and semantic-aware graph memory networks for temporal sentence localization in videos. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
65. Liu, D., Zhou, P., Xu, Z., Wang, H., Li, R.: Few-shot temporal sentence grounding via memory-guided semantic learning. IEEE Transactions on Circuits and Systems for Video Technology **33**(5), 2491–2505 (2022)
66. Liu, D., Zhu, J., Fang, X., Xiong, Z., Wang, H., Li, R., Zhou, P.: Conditional video diffusion network for fine-grained temporal sentence grounding. IEEE Transactions on Multimedia (2023)
67. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 4768–4777 (2017)
68. Ma, M., Yoon, S., Kim, J., Lee, Y., Kang, S., Yoo, C.D.: VLANet: Video-language alignment network for weakly-supervised video moment retrieval. In: Proceedings of the European Conference on Computer Vision. pp. 156–171 (2020)
69. Ma, W.C., Huang, D.A., Lee, N., Kitani, K.M.: Forecasting interactive dynamics of pedestrians with fictitious play. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 774–782 (2017)
70. Ma, Y., Liu, Y., Wang, L., Kang, W., Qiao, Y., Wang, Y.: Dual masked modeling for weakly-supervised temporal boundary discovery. IEEE Transactions on Multimedia (2023)
71. Matsui, Y., Matsui, T.: Np-completeness for calculating power indices of weighted majority games. Theoretical Computer Science **263**(1-2), 305–310 (2001)
72. Michalak, T.P., Aadithya, K.V., Szczepanski, P.L., Ravindran, B., Jennings, N.R.: Efficient computation of the shapley value for game-theoretic network centrality. Journal of Artificial Intelligence Research **46**, 607–650 (2013)
73. Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11592–11601 (2019)
74. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10810–10819 (2020)
75. Ning, E., Wang, C., Zhang, H., Ning, X., Tiwari, P.: Occluded person re-identification with deep learning: a survey and perspectives. Expert Systems with Applications p. 122419 (2023)
76. Ning, E., Wang, Y., Wang, C., Zhang, H., Ning, X.: Enhancement, integration, expansion: Activating representation of detailed features for occluded person re-identification. Neural Networks **169**, 532–541 (2024)
77. Ning, E., Zhang, C., Wang, C., Ning, X., Chen, H., Bai, X.: Pedestrian re-id based on feature consistency and contrast enhancement. Displays **79**, 102467 (2023)
78. Nowak, A.S.: On an axiomatization of the banzhaf value without the additivity axiom. International Journal of Game Theory **26**, 137–141 (1997)
79. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
80. Osborne, M.J., Rubinstein, A.: A course in game theory. MIT press (1994)

81. Patel, R., Garnelo, M., Gemp, I., Dyer, C., Bachrach, Y.: Game-theoretic vocabulary selection via the shapley value and banzhaf index. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2789–2798 (2021)
82. Pavan, M., Pelillo, M.: A new graph-theoretic approach to clustering and segmentation. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 1, pp. I–I. IEEE (2003)
83. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1532–1543 (2014)
84. Rodola, E., Bronstein, A.M., Albarelli, A., Bergamasco, F., Torsello, A.: A game-theoretic approach to deformable shape matching. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 182–189. IEEE (2012)
85. Shapley, L.S., et al.: A value for n-person games (1953)
86. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526 (2016)
87. Song, Y., Wang, J., Ma, L., Yu, J., Liang, J., Yuan, L., Yu, Z.: Marn: Multi-level attentional reconstruction networks for weakly supervised video temporal grounding. Neurocomputing **554**, 126625 (2023)
88. Song, Y., Wang, J., Ma, L., Yu, Z., Yu, J.: Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. arXiv preprint arXiv:2003.07048 (2020)
89. Tan, R., Xu, H., Saenko, K., Plummer, B.A.: Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 2083–2092 (2021)
90. Tang, K., Chen, Y., Peng, W., Zhang, Y., Fang, M., Wang, Z., Song, P.: Reppvconv: attentively fusing reparameterized voxel features for efficient 3d point cloud perception. The Visual Computer **39**(11), 5577–5588 (2023)
91. Tang, K., Lou, T., Peng, W., Chen, N., Shi, Y., Wang, W.: Effective single-step adversarial training with energy-based models. IEEE Transactions on Emerging Topics in Computational Intelligence (2024). https://doi.org/10.1109/TETCI.2024.3378652
92. Tang, K., Ma, Y., Miao, D., Song, P., Gu, Z., Tian, Z., Wang, W.: Decision fusion networks for image classification. IEEE Transactions on Neural Networks and Learning Systems (2022). https://doi.org/10.1109/TNNLS.2022.3196129
93. Tang, K., Shi, Y., Lou, T., Peng, W., He, X., Zhu, P., Gu, Z., Tian, Z.: Rethinking perturbation directions for imperceptible adversarial attacks on point clouds. IEEE Internet of Things Journal **10**(6), 5158–5169 (2022)
94. Tang, K., Zhao, W., Peng, W., Fang, X., Cui, X., Zhu, P., Tian, Z.: Reparameterization head for efficient multi-input networks. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6190–6194 (2024). https://doi.org/10.1109/ICASSP48485.2024.10447574
95. Tang, K., Zhao, W., Peng, W., Fang, X., Cui, X., Zhu, P., Tian, Z.: Reparameterization head for efficient multi-input networks. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6190–6194. IEEE (2024)
96. Torsello, A., Bulo, S.R., Pelillo, M.: Grouping with asymmetric affinities: A game-theoretic perspective. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1, pp. 292–299. IEEE (2006)

97. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497 (2015)
98. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
99. Wang, C., Ning, X., Li, W., Bai, X., Gao, X.: 3d person re-identification based on global semantic guidance and local feature aggregation. IEEE Transactions on Circuits and Systems for Video Technology (2023)
100. Wang, C., Ning, X., Sun, L., Zhang, L., Li, W., Bai, X.: Learning discriminative features by covering local geometric space for point cloud analysis. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022)
101. Wang, C., Wang, C., Li, W., Wang, H.: A brief survey on rgb-d semantic segmentation using deep learning. Displays **70**, 102080 (2021)
102. Wang, C., Wang, H., Ning, X., Shengwei, T., Li, W.: 3d point cloud classification method based on dynamic coverage of local area. Journal of Software **34**(4), 1962–1976 (2022)
103. Wang, J., Ma, L., Jiang, W.: Temporally grounding language queries in videos by contextual boundary-aware prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12168–12175 (2020)
104. Wang, Y., Deng, J., Zhou, W., Li, H.: Weakly supervised temporal adjacent network for language grounding. IEEE Transactions on Multimedia **24**, 3276–3286 (2021)
105. Wang, Z., Chen, J., Jiang, Y.G.: Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1459–1468 (2021)
106. Wen, J., Liu, C., Deng, S., Liu, Y., Fei, L., Yan, K., Xu, Y.: Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. IEEE Transactions on Neural Networks and Learning Systems pp. 1–13 (2023). https://doi.org/10.1109/TNNLS.2023.3260349
107. Wen, J., Zhang, Z., Li, Z.J.: A survey on incomplete multiview clustering. IEEE transactions on systems, man, and cybernetics. Systems **53**(2 Pt.2), 1136–1149 (2023)
108. Winter, E.: The shapley value. Handbook of game theory with economic applications **3**, 2025–2054 (2002)
109. Wu, H., Lyu, Y., Shen, X., Zhao, X., Wang, M., Zhang, X., Luo, Z.: Atomic-action-based contrastive network for weakly supervised temporal language grounding. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 1523–1528. IEEE (2023)
110. Xiong, Z., Liu, D., Zhou, P.: Gaussian kernel-based cross modal network for spatio-temporal video grounding. In: IEEE International Conference on Image Processing (ICIP). pp. 2481–2485 (2022)
111. Xiong, Z., Liu, D., Zhou, P., Zhu, J.: Tracking objects and activities with attention for temporal sentence grounding. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
112. Yang, W., Zhang, T., Zhang, Y., Wu, F.: Local correspondence network for weakly supervised temporal sentence grounding. IEEE Transactions on Image Processing **30**, 3252–3262 (2021)

113. Yu, Z., Li, L., Xie, J., Wang, C., Li, W., Ning, X.: Pedestrian 3d shape understanding for person re-identification via multi-view learning. IEEE Transactions on Circuits and Systems for Video Technology (2024)
114. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 536–546 (2019)
115. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10287–10296 (2020)
116. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1247–1257 (2019)
117. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6543–6554 (2020)
118. Zhang, H., Xie, Y., Zheng, L., Zhang, D., Zhang, Q.: Interpreting multivariate shapley interactions in dnns. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10877–10886 (2021)
119. Zhang, H., Wang, C., Tian, S., Lu, B., Zhang, L., Ning, X., Bai, X.: Deep learning-based 3d point cloud classification: A systematic survey and outlook. Displays **79**, 102456 (2023)
120. Zhang, H., Wang, C., Yu, L., Tian, S., Ning, X., Rodrigues, J.: Pointgt: A method for point-cloud classification and segmentation based on local geometric transformation. IEEE Transactions on Multimedia (2024)
121. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12870–12877 (2020)
122. Zhang, Z., Lin, Z., Zhao, Z., Xiao, Z.: Cross-modal interaction networks for query-based moment retrieval in videos. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 655–664 (2019)
123. Zhang, Z., Zhao, Z., Lin, Z., He, X., et al.: Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems **33**, 18123–18134 (2020)
124. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3517–3525 (2022)
125. Zheng, M., Huang, Y., Chen, Q., Peng, Y., Liu, Y.: Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15555–15564 (2022)
126. Zheng, Q., Dong, J., Qu, X., Yang, X., Wang, Y., Zhou, P., Liu, B., Wang, X.: Progressive localization networks for language-based moment localization. ACM Transactions on Multimedia Computing, Communications and Applications **19**(2), 1–21 (2023)
127. Zhu, J., Liu, D., Zhou, P., Di, X., Cheng, Y., Yang, S., Xu, W., Xu, Z., Wan, Y., Sun, L., et al.: Rethinking the video sampling and reasoning strategies for temporal sentence grounding. arXiv preprint arXiv:2301.00514 (2023)