AxIoU: An Axiomatically Justified Measure for Video Moment Retrieval

Riku Togashi
Cyberagent, Inc., Waseda University

Mayu Otani
Cyberagent, Inc.

Yuta Nakashima
Osaka University

Esa Rahtu
Tampere University

Janne Heikkilä
University of Oulu

Tetsuya Sakai
Waseda Univeristy

Abstract

Evaluation measures have a crucial impact on the direction of research. Therefore, it is of utmost importance to develop appropriate and reliable evaluation measures for new applications where conventional measures are not well suited. Video Moment Retrieval (VMR) is one such application, and the current practice is to use R@K, θ for evaluating VMR systems. However, this measure has two disadvantages. First, it is rank-insensitive: It ignores the rank positions of successfully localised moments in the top-K ranked list by treating the list as a set. Second, it binarizes the Intersection over Union (IoU) of each retrieved video moment using the threshold θ and thereby ignoring fine-grained localisation quality of ranked moments.

We propose an alternative measure for evaluating VMR, called Average Max IoU (AxIoU), which is free from the above two problems. We show that AxIoU satisfies two important axioms for VMR evaluation, namely, ***Invariance against Redundant Moments*** *and* ***Monotonicity with respect to the Best Moment****, and also that R@K, θ satisfies the first axiom only. We also empirically examine how AxIoU agrees with R@K, θ, as well as its stability with respect to change in the test data and human-annotated temporal boundaries.*

1. Introduction

Video Moment Retrieval (VMR) has been explored to find relevant fragments of videos (*i.e.* video moments) based on a user's textual query [10, 15]. Most existing VMR systems [10, 22, 38, 39, 42] cast the problem of finding video moments into a ranking problem. For evaluating ranked lists of video moments, R@K, θ is widely adopted in the literature [10]. R@K, θ for a query q is defined as 1 if at least one relevant video moment in the top K of the ranked list has an Intersection over Union (IoU) larger than θ with the ground truth for q.

R@K, θ has two disadvantages as illustrated in Figure 1.

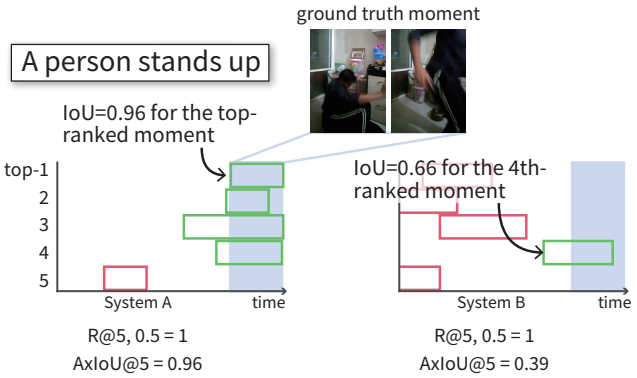

Figure 1. The system on the left shows a moment with a large overlap with the ground truth (blue band) in the top of the ranked list, and the system on the right illustrates a moment with a much smaller overlap with the ground truth at rank 4. According to R@$5, 0.5$, the two systems are equally effective. We propose AxIoU whose measurements reflect the localisation quality (*i.e.* IoU) and the rank of successfully retrieved video moments. The photo in the figure is taken from Charades-STA [10].

First, it is *rank-insensitive*, as the video moments in the top-K ranked list are treated as a set, and their ranks are not considered. Second, it is *localisation-insensitive*, *i.e.*, the exact position (start and end points) of the video moment does not affect the measurement as it binarizes the IoU of each video moment using threshold θ. Thus, R@K, θ only provides a binary measurement for a ranked list in an all-or-nothing manner, ignoring the ranking and localisation quality of top-K predicted video moments. As we shall demonstrate in this paper, these properties of R@K, θ are problematic for reliable evaluation. R@K, θ cannot distinguish ranked lists with different quality due to the binary property, while leading to instability under a small number of evaluation samples and label ambiguity [2, 15, 26, 35]. Moreover, R@K, θ evaluates rather different aspects of system quality depending on a parameter setting and can conflict with each other; for example, a ranked list whose second moment achieves IoU $= 0.71$ is measured to be 1.0 in terms of

R@$2, 0.7$ but to be 0.0 in terms of R@$1, 0.7$ when the first moment has IoU $= 0.69$. These undesirable properties of R@K, θ should be carefully considered for future studies because conclusions drawn from potentially unstable measures may not generalise well. In practice, the instability of R@K, θ implies that we may underestimate a VMR method by adopting a non-best model based on validation R@K, θ.

In this paper, we propose an alternative measure for evaluating VMR systems, called *Average Max IoU* (AxIoU), which does not suffer from the problems above with R@K, θ. To evaluate evaluation measures, we take an axiomatic approach [3, 9, 33, 34] and introduce two important axioms that an effectiveness measure for VMR must satisfy, namely, *invariance against redundant moments* and *monotonicity with respect to the best moment*. We show that R@K, θ only satisfies the first axiom. We also empirically investigate the properties of AxIoU in practical terms, namely, agreement with conventional R@K, θ and the stability to the size of a dataset and label ambiguity.

2. Related Work

Most of the prior studies of VMR adopted the R@K, θ measure [7, 10, 11, 19, 22, 37, 38, 39, 41, 42]. Gao *et al*. [10] proposed the use of this measure for VMR by referring to the work of Hu *et al*. [16], which is an early study on an object retrieval task with textual queries. The values of K and θ are chosen for each dataset. For example, the combinations of $K = 1, 5, 10$ and $\theta = 0.3, 0.5, 0.7$ are widely adopted in Charades-STA [10], ActivityNet [5, 19], and DiDeMo [15]. In the TACoS dataset [28], relatively relaxed values of θ (*i.e*. $\theta = 0.1, 0.3, 0.5$) are used. For evaluating methods that output only one moment per query [12, 14, 40], R@K, θ with $K = 1$ is often adopted. Lei *et al*. recently proposed a new retrieval task called video corpus moment retrieval [20], in which a system requires to retrieve relevant moments from multiple videos. Owing to a large number of candidate moments, they utilise large values of K such as $K = 100$. However, the common practice of reporting multiple settings of R@K, θ is controversial. As we shall demonstrate in this paper, different parameter settings often lead to different system rankings, from which it may be difficult to draw useful conclusions from the evaluation.

Prior studies suggested that the inter-rater agreement of human-annotated temporal boundaries is often not strong [2, 15, 26, 35]. Hendricks *et al*. found that there are multiple video moments, which can be described by a textual query [15]; to alleviate this label ambiguity, they developed a user interface. Sigurdsson *et al*. and Alwassel *et al*. also reported that human-annotated temporal regions do not agree well with each other [2, 35]. Otani *et al*. observed high label ambiguity in Charades-STA and ActivityNet [26]. Nevertheless, the binarization of IoU values in R@K, θ introduces potential instability to the change of labels. In particular, a large value of θ requires exactly located temporal regions and thereby being noisy, inheriting label ambiguity.

In the context of object detection, in which evaluation measures often rely on a threshold parameter for spatial IoU, prior studies have discussed the disadvantages of a fixed threshold [13, 25, 27]. On the MSCOCO [21] dataset, an average of measures over IoU threshold values is adopted for evaluating fine-grained localisation quality; the measure is called COCO mean average precision (mAP). Oksuz *et al*. proposed an object detection measure to directly quantify the bounding box tightness by introducing IoU values without thresholding in their measure [25]. Hall *et al*. have explored a way to improve the spatial quality evaluation of detected regions beyond the conventional box-based IoU while reducing the parameters in evaluation measures [13]. In temporal localisation tasks for videos (*e.g*. action detection), Alwassel *et al*. used a COCO mAP-like measure for evaluation [2].

In contrast to rank-insensitive set retrieval measures (*e.g*. precision and recall), ranked retrieval measures have been explored for evaluating the quality of a list of ranked items, such as normalised discounted cumulative gain (nDCG) [17]. Such measures often have weights for the rank positions in a retrieval result; for example, the discount function in nDCG can be regarded as the importance of each position. Based on the interpretation of the position weights from the viewpoint of *user models*, prior studies have developed various evaluation measures [6, 23, 29, 31].

The evaluation of evaluation measures is often challenging as it requires the true evaluation results a priori. One approach to verify the experiments based on an evaluation measure is to collect human manual assessments for *search engine result pages (SERPs)* [32]. For new applications such as VMR, it is often costly to establish a reliable environment to collect the gold data that aligns well with the "true" quality; we may need to study such as human effects on the reliability of the gold data [18]. An axiomatic approach is another direction for the verification of evaluation measures [3, 9, 33, 34]. By formally defining requirements that a measure should satisfy, we can analytically confirm the validity of measures. Such requirements inevitably depend on a number of assumptions. However, this is also true for the assessment-based approach because guidelines for assessors implicitly involve assumptions on users' behaviours [32].

In this paper, we propose an alternative VMR measure, AxIoU, which is an instantiation of normalised cumulative utility (NCU) [30, 31], which is a wide class of information retrieval measures including AP. Our proposed measure considers the rank positions and IoU values of video moments. To confirm the properties of measures, we take

an axiomatic approach. The derivation of AxIoU is related to COCO mAP, whereas AxIoU analytically reduces the binarization process for IoU values. Through empirical experiments, we confirm the numerical properties of AxIoU while showing the undesirable behaviours of R@K, θ.

3. Preliminaries

3.1. Notations

Our goal is to develop a measure $\mu(q, \sigma)$ that estimates the retrieval effectiveness of a system σ based on a test query q. We also denote by $\mu(\mathcal{Q}, \sigma) = (1/|\mathcal{Q}|) \sum_{q \in \mathcal{Q}} \mu(q, \sigma)$ the mean of the measurements based on a test query set $\mathcal{Q}$. For a query $q \in \mathcal{Q}$, the system σ sorts the set $\mathcal{M}_q$ of candidate moments and creates the ranked list σ_q. We also denote by $\sigma_q(k) \in \mathcal{M}_q$ the moment ranked at position k in σ_q. Let $r_q(m) \in [0, 1]$ be the relevance score of a moment $m \in \mathcal{M}_q$, computed as the temporal IoU (Intersection over Union) between m and the ground truth region for q. Where there is no ambiguity, we will also denote it by $r(m)$.

3.2. R@K, θ

First, we formally define the conventional measure, R@K, θ [10], and clarify what it quantifies as well as its limitations. Here, we denote by $\mathbb{1} : \mathbb{B} \mapsto \{0, 1\}$ the indicator function for the boolean variable X that takes 1 if X is true and 0 if X is false. We express R@K, θ and Mean R@K, θ as follows.

$$\begin{aligned} \text{R@}K, \theta(q, \sigma) &\coloneqq \mathbb{1}\left\{ \sum_{k=1}^{K} \mathbb{1}\left\{ r(\sigma_q(k)) > \theta \right\} > 0 \right\} \\ &= \mathbb{1}\left\{ \max_{1 \le k \le K} r(\sigma_q(k)) > \theta \right\}. \end{aligned} \tag{1}$$

The value of R@K, θ depends entirely on whether the most relevant moment in the top-K retrieved results exceeds the θ threshold. It is clear from this that R@K, θ does not reward redundancy: the retrieved moments other than the most relevant one in the SERP do not count, even if they also exceed θ. We shall refer to such relevant moments as *redundant* moments.

The above property of R@K, θ is a desirable feature, since real VMR system users probably do not care about redundant moments in their SERPs. However, it is clear from Eq. 1 that R@K, θ has two potential shortcomings. First, R@K, θ is unchanged by the rank positions of the relevant moments: it is a set retrieval measure rather than a ranked retrieval measure. For $K > 1$, it cannot distinguish between a system that retrieves a perfectly relevant moment at rank 1, and a system that retrieves the same moment at rank K. Second, it binarizes the IoU of each moment using the θ threshold, and thereby ignores the degree of relevance of each retrieved moment. Choosing an appropriate value of θ is practically problematic, especially given that K also needs to be chosen at the same time.

4. Proposed Measure

4.1. Average Max IoU Measure

To design a measure for VMR, we adopt the framework of a wide class of retrieval effectiveness measures, *normalised cumulative utility (NCU)* [30, 31]. NCU assumes that there is a population of users who scan a ranked list, starting from the top, and abandons the ranked list on a certain rank position k. Here, NCU for a query q and a system σ can be expressed as follows:

$$\text{NCU}(q, \sigma) = \sum_{k=1}^{|\mathcal{M}_q|} P_A(k) U(\sigma_q, k), \tag{2}$$

where $P_A(k)$ is the abandonment probability at rank position k (*i.e.* the population of users who stop at k), and $U(\sigma_q, k)$ is the utility of the ranked list σ_q at k. As we do not want to reward redundancy in VMR, we follow the approach of R@K, θ (Eq. 1) to instantiate our utility function:

$$U(\sigma_q, k) = \max_{1 \le j \le k} r(\sigma_q(j)). \tag{3}$$

Based on this, we can obtain *normalised cumulative max IoU (NCxIoU)* measure as follows:

$$\text{NCxIoU}(q, \sigma) \coloneqq \sum_{k=1}^{|\mathcal{M}_q|} P_A(k) \max_{1 \le j \le k} r(\sigma_q(j)). \tag{4}$$

With VMR, we do not have any prior knowledge on $P_A(k)$. Therefore, given a SERP containing K moments, we assume that the users are uniformly distributed over the K moments: that is, that $1/K$ of the user population abandons the list at rank k ($1 \le k \le K$). Note that the Average Precision (AP), an NCU measure widely used in information retrieval evaluation with manual relevance assessments, assumes that the users are uniformly distributed over *all relevant* documents [29]. In the case of VMR, we consider only the top-K items (following R@K, θ), and assume that each retrieved moment is at least somewhat relevant, where the degree of relevance is represented by the IoU of each moment.

Our proposed measure for VMR is also an instantiation of NCU, which we call average max IoU (AxIoU):

$$\text{AxIoU@}K(q, \sigma) \coloneqq \frac{1}{K} \sum_{k=1}^{K} \max_{1 \le j \le k} r(\sigma_q(j)). \tag{5}$$

As the uniform assumption on $P_A(k)$ may not hold when with a large K, we can use a more realistic distribution for $P_A(k)$ such as the expected reciprocal rank (another NCU measure) [6], although we leave this as future work.

4.2. Interpretation of the AxIoU Measure

In this section, we describe the relationship between our proposed measure and R@K, θ. We first consider the marginalisation of R@K, θ in terms of K and θ. In practical terms, because we do not have any knowledge regarding the distribution of θ for each dataset, each query, or each set of systems to be evaluated, we assume that $\theta \sim \text{Uni}(0, 1)$ and then obtain the following equation:

$$\begin{aligned}
&\mathbb{E}_k \mathbb{E}_\theta \left[\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \text{R@}k, \theta(q, \sigma) \right] \\
&= \mathbb{E}_k \mathbb{E}_\theta \left[\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1} \left\{ \max_{1 \le j \le k} r(\sigma_q(j)) > \theta \right\} \right] \\
&= \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{E}_k \mathbb{E}_\theta \left[\mathbb{1} \left\{ \max_{1 \le j \le k} r(\sigma_q(j)) > \theta \right\} \right] \\
&= \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{E}_k \left[\max_{1 \le j \le k} r(\sigma_q(j)) \right]. \qquad (6)
\end{aligned}$$

In Eq (6), by assuming $\theta \sim \text{Uni}(0, 1)$, we can obtain the following:

$$\mathbb{E}_\theta \left[\mathbb{1} \left\{ \max_{1 \le j \le k} r(\sigma_q(j)) > \theta \right\} \right] = \max_{1 \le j \le k} r(\sigma_q(j)). \qquad (7)$$

Because we assume a uniform distribution for k on $1 \le k \le K$ and $P_A(k) = 1/K$, we obtain the following:

$$\begin{aligned}
\text{(RHS of Eq. (6))} &= \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{K} \sum_{k=1}^{K} \max_{1 \le j \le k} r(\sigma_q(j)) \\
&= \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \text{AxIoU@}K(q, \sigma). \qquad (8)
\end{aligned}$$

That is, the mean AxIoU@K can be considered as a marginalisation of the mean R@K, θ without any assumption for θ and with a weak assumption for K. The mean R@K, θ with a fixed value for each K and θ evaluates certain aspects of systems' behaviour and thus requires to examine multiple settings of the parameters for evaluation. We argue that AxIoU is a reasonable approach to avoiding the dependence on the θ threshold while considering the rank position of the best moment in a top-K ranked list.

5. Requirements for Effectiveness Measures

To evaluate the evaluation measures, we take an axiomatic approach. We first set the following requirements for the design of our VMR measure based on the properties of R@K, θ in Section 3.2: (1) It should ignore redundant moments in a ranked list, (2) it should consider the IoU value between a ranked moment and the ground truth moment, and (3) it should consider the rank position of relevant moments for evaluating top-K retrieval effectiveness of a system. We show that our Mean AxIoU satisfies all requirements while Mean R@K, θ satisfies only Requirement (1). To investigate VMR measures based on these requirements, we define two axioms for an effectiveness measure for VMR.

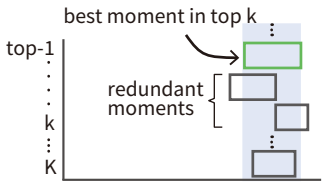

Figure 2. **INV-k** requires that a measure should be invariant to redundant moments which have smaller IoU and lower rank position than the best moment in the top-k ($1 \le k \le K$) ranked list.

Invariance against Redundant Moments A measure should be unchanged to redundant moments in a ranked list. We define this requirement as the following axiom.

Axiom 1 (Invariance against Top-k Non-Best Moment (**INV-k**).). *Suppose that two systems σ and σ' such that σ' differs from σ only for the k-th moment in the ranked lists for q. The measurement of σ' must not change from that of σ (i.e. $\mu(\mathcal{Q}, \sigma) = \mu(\mathcal{Q}, \sigma')$) when the k-th moment in σ' has a better IoU value than the k-th moment in σ but is not the most relevant within the top k of σ'.*

Figure 2 depicts the concept of **INV-k**. R@K, θ satisfies this requirement because it utilises only the moment with the maximum IoU value in a ranked list (see Eq. (1)). AxIoU can also handle the redundant moments by inheriting the property of R@K, θ. On the other hand, AP@K, θ, which is a ranked retrieval measure widely adopted in computer vision [8], does not satisfy **INV-k**. Similarly, while an information retrieval measure for graded relevance such as DCG [17] would be a straightforward choice for evaluating ranked lists while avoiding the binarization by θ, it does not satisfy **INV-k** either. The formal definition of the axiom and proofs are given in our supplementary material.

Monotonicity with respect to the Best Moment The VMR measure score should monotonically increase with the maximum IoU value in a ranked list. More specifically, we require that at any rank k, the measurement based on the top-k moments of the SERP should monotonically increase with the maximum IoU observed within the top k. This requirement can be defined through the following axiom.

Axiom 2 (Strict Monotonicity for Top-k Best Moment (**MON-k**).). *Suppose that two systems σ and σ' such that*

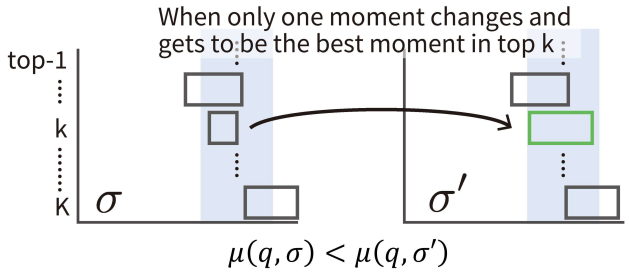

Figure 3. **MON-k** requires that a measure should be sensitive to the IoU value of the best moment in a top-k ranked list.

σ' differs from σ only for the k-th moment in the ranked lists for q. The measurement of σ' strictly increases from that of σ (i.e. $\mu(\mathcal{Q},\sigma) < \mu(\mathcal{Q},\sigma')$) when the k-th moment in σ' has a better IoU value than the k-th moment in σ and is the most relevant within the top k of σ'.

Figure 3 depicts the concept of **MON-k**. R@K,θ with a fixed parameter setting for K and θ does not satisfy this requirement. $\mu(\mathcal{Q},\sigma) < \mu(\mathcal{Q},\sigma')$ is not guaranteed since R@K,θ binarizes the relevance using θ. By contrast, ranked retrieval measures for graded relevance, such as DCG@K and our AxIoU@K, satisfies this property because these consider the ranked position and IoU value of each moment in a ranked list. The formal definition of the axiom and proofs are provided in the supplementary material.

6. Experiments

While we analytically showed the properties of AxIoU@K in terms of the axioms, we also examine the measures empirically in this section. We first investigate the agreement between the evaluation results based on the measures to confirm the compatibility of AxIoU@K with R@K,θ. To examine the effect of θ, we also discuss the stability of the measures with respect to change in the test data. Moreover, we demonstrate the advantages of AxIoU@K as the criterion for model selection.

6.1. Experimental Setup

Datasets Following the experimental settings of Otani *et al*. [26], we utilise two popular datasets for our experiments, Charades-STA [10] and ActivityNet [5, 19]. Each dataset contains a set of manually annotated temporal regions for query-video pairs that indicate the relevant moment in a video as ground truth. Charades-STA is built upon Charades [36] and contains 9,848 videos, each of which is associated with multiple natural language sentences. The number of test queries is 3,720. ActivityNet contains 19,209 YouTube videos. Each video is associated with the captions and their temporal locations. The number of the test queries is 17,031.

Retrieval Systems for Evaluation In our experiment, we utilise multiple VMR systems to evaluate the measures; for example, we create two rankings of the systems based on two measures, and then compute the similarity of the rankings (i.e. Kendall's τ-b [1]) as agreement between the two measures. To examine each measure in a realistic setting, we employ real VMR systems trained on each dataset. Throughout this paper, we used three conventional methods, **Action-Aware Blind** (Blind) [26], **SCDM** [39] and **2DTAN** [43]. In addition, we include the variants of 2DTAN, *i.e*., (1) **2DTAN nonms**, a variant without Non-maximum suppression (NMS) [24], (2) **2DTAN rand**, a variant with randomisation of video frames proposed by Otani *et al*. [26], and (3) **2DTAN rand+nonms**, a variant without NMS and with randomisation.

Figure 5 compares the effectiveness of the above six systems according to different measures on Charades-STA (left) and ActivityNet (right). In each graph, the systems have been sorted by Mean R@$5, 0.5$. For Charades-STA, the R@$10, 0.3$ score of Blind (green solid line), which is a video-agnostic baseline, is almost one; as the Charades-STA dataset is a relatively easy dataset, R@$10, 0.3$ is a too relaxed measure even for Blind. This result suggests that a inappropriate choice of K and θ leads to uninformative evaluation results.

6.2. Agreement between Measures

Figure 4 shows the agreement between each pair of measures among R@K,θ ($K = 1, 5, 10, \theta = 0.3, 0.5, 0.7$) and the AxIoU@$K$ ($K = 1, 5, 10$) on Charades-STA and ActivityNet datasets, respectively. To assess the agreement between two measures, we first rank the six systems using each measure. We then compute Kendall's τ-b [1], which considers the ties in a ranking. Hereafter, we shall refer to τ-b simply as τ. A high τ value means that the rankings according to the two measures are similar [4].

In the Charades-STA dataset, AxIoU@10 agrees well with all instances of R@K,θ ($0.36 \leq \tau \leq 0.87$). The values of AxIoU@K with different values of K agree reasonably well with one another ($0.36 \leq \tau \leq 0.73$). By contrast, different instances of R@K,θ can conflict with one another; R@$5, 0.7$ agrees well with R@$1,\theta$ ($\tau = 0.64$) whereas R@$10,\theta$ does not agree with R@$1,\theta$ instances ($-0.21 \leq \tau \leq 0.21$). Probably, the main reason for this result is that R@K,θ is rank-insensitive. On the other hand, AxIoU@K, which satisfies **MON-k**, aligns well with itself for different values of K. Although R@$5, 0.5$ and R@$5, 0.7$, which are popular instances of R@K,θ, agree relatively well with the other R@K,θ instances ($0.21 \leq \tau \leq 0.73$ for R@$5, 0.5$ and $0.2 \leq \tau \leq 0.64$ for R@$5, 0.7$), the agreement between the two measures is $\tau = 0.60$ despite the small difference in the setting of θ; remarkably, AxIoU@10 agrees with R@$5, 0.5$ and R@$5, 0.7$ with

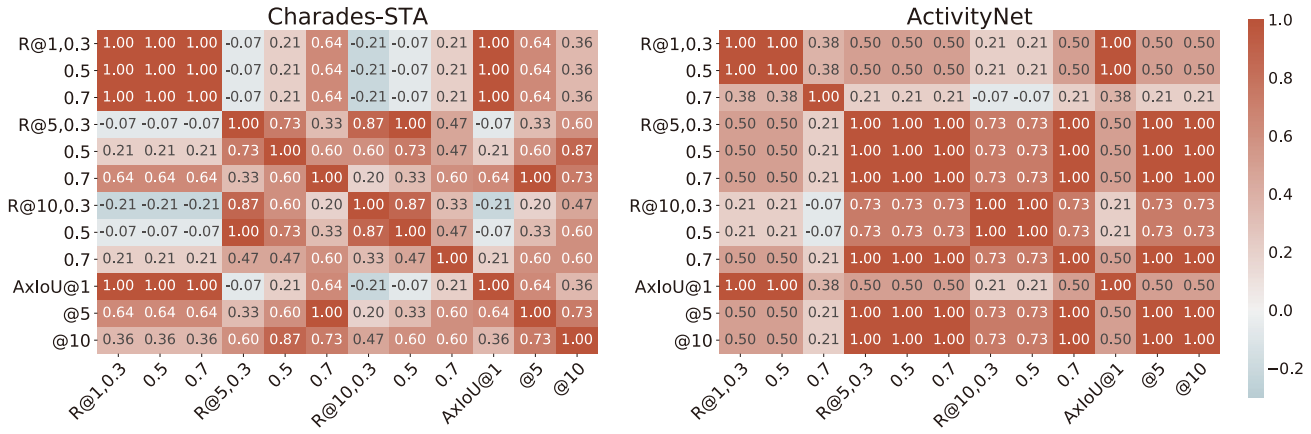

Figure 4. Agreement between two measures on Charades-STA (left) and ActivityNet (right).

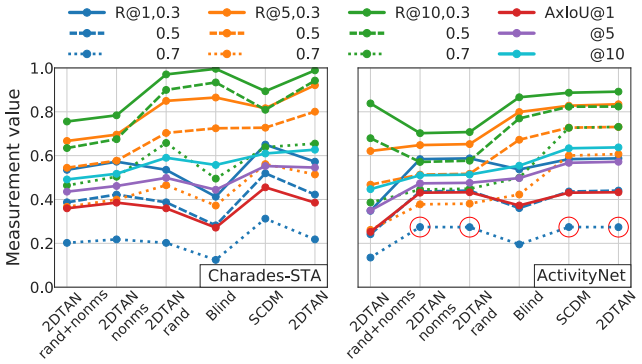

Figure 5. Effectiveness of each system on Charades-STA and ActivityNet datasets according to R@K, θ and AxIoU@K.

$\tau = 0.87$ and $\tau = 0.73$, respectively.

Because the ActivityNet dataset has a much larger number of test queries than that of the Charades-STA dataset, most of the measures agree well with each other. Nevertheless, R@$1, 0.7$, which is a widely adopted instance of R@K, θ, does not agree with the other instances ($-0.07 \leq \tau \leq 0.38$). It is worth mentioning that R@$1, 0.7$ is highly demanding (*i.e.* requiring systems to return a highly relevant moment at rank 1). Thus, R@$1, 0.7$ lacks the sensitivity to distinguish between the systems. As shown in Fig. 5 (red circles in the right side), the scores by R@$1, 0.7$ are low for all six systems, and the scores for four out of six systems are all 0.274 (1,019/3,720). AxIoU@10 achieves strong agreement ($\tau \geq 0.5$) with all instances of R@K, θ except R@$1, 0.7$.

6.3. Stability of Measures against the Choice of Evaluation Data

In this section, we investigate the stability of the measures, *i.e.*, the consistency of evaluation results based on a measure on different test datasets [4]. The stability of an effective measure against different test datasets is one of the essential properties: If a measure is unstable, the conclusion drawn for a certain test dataset may not generalise well. We evaluate each measure based on Kendall's τ-b between the system rankings based on two different subsets of a dataset as the *self-agreement* of the measure. To examine the stability with respect to the choice and size of test data we investigate the self-agreement on conjoint query set pairs with different sizes.

Figure 6 visualises the effect of reducing the size of the query subsets on self-agreement. We experimented with 5,000 trials for each query subset size. The horizontal axes represent the size of each query subset. The top graphs show the means of the self-agreement τ's; the bottom graphs show the variances.

For Charades-STA in the first to third columns, it can be observed that, for each K, the R@K, θ instances with $\theta = 0.3, 0.5$ substantially underperform the other in terms of mean and variance of τ; on the other hand, the R@$K, 0.7$ (red dotted line) instances are consistently stable. The AxIoU@K instances outperform most of the R@K, θ instances with the same value of K whereas it performs relatively poorly with small query sets for $K = 5$. The most robust batch of measures for this dataset are R@$1, 0.7$, R@$5, 0.7$, R@$10, 0.7$, AxIoU@1 and AxIoU@10. Also for ActivityNet in the forth to fifth columns, the AxIoU@K instances outperform most of the R@K, θ; the R@$K, 0.7$ instances also perform well.

6.4. Stability against Label Ambiguity

In this section, we evaluate the measures in terms of the stability to label ambiguity (i.e. disagreement between human annotations). We generate a testing sample based on a simple noise model by following steps; (1) we consider each annotation in an original testing dataset as a low-noise sample and denote one by $(s^*, e^*) \in \mathbb{R}^2$ where s^* and e^* are the start and end points of a temporal boundary; (2) we draw

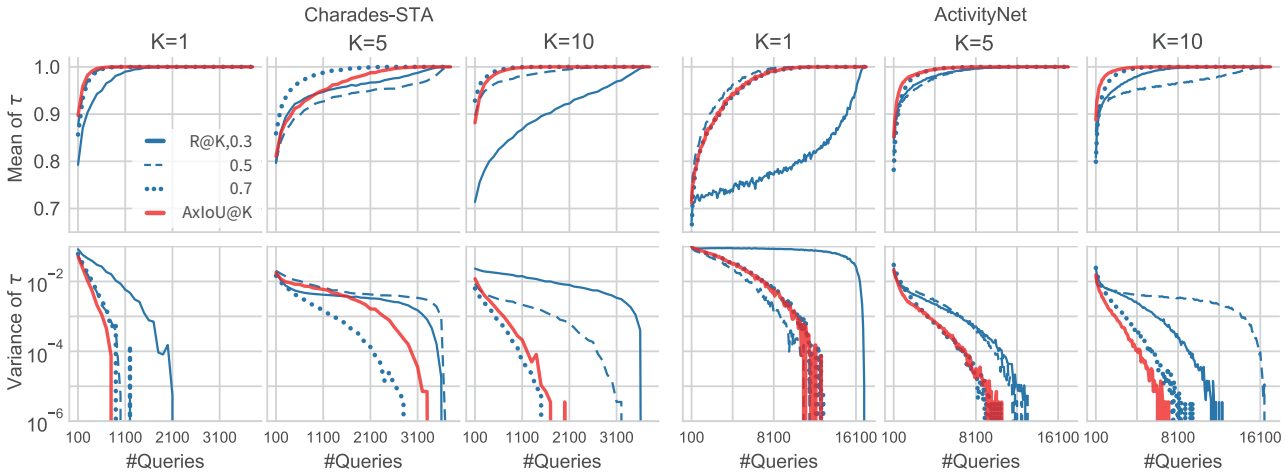

Figure 6. Effect of reducing the size of query subsets on means and variances of self-agreement for Charades-STA and ActivityNet.

a start point s by a normal distribution with mean s^* and variance β^2; (3) we then draw a length l by an exponential distribution with mean $e^* - s^*$; and (4) we obtain the drawn sample $(s, s+l)$ as a noisy one. For each testing sample in Charades-STA and ActivityNet, we independently draw five samples from the noise model and then create a final testing annotation by taking medians of s and $s+l$. Here, it should be noted that the variance parameter β^2 can be considered as the quality (i.e. noise level) of five raters who annotate temporal boundaries to one sample. We generate datasets with different noise levels by varying β^2 in $\{1, 2, 3, 4\}$. The IoU between the mean of the median IoU values between the original and a drawn annotation for each noise level $\{1, 2, 3, 4\}$ is respectively 0.906, 0.870, 0.835 and 0.802 for Charades-STA, and 0.846, 0.778, 0.712 and 0.650 for ActivityNet; note that, the noise levels are in a realistic range as the previously reported IoU agreement between human annotations is around 0.725 in Charades-STA [35] and 0.641 in ActivityNet [2]. We generate independent 100 testing datasets for each dataset and each noise level. To evaluate the effect of label ambiguity for each measure, we compute the root mean squared error (RMSE) between the measurements based on the original dataset and 100 noisy datasets for each of six systems used in the above experiments.

Figure 7 shows the effect of the label noise on the evaluation based on the measures. The x- and y-axes indicate the noise level and Mean RMSE for each measure. In all datasets and all K, the AxIoU instances show lower errors than the R@$K, 0.7$ instances but higher errors than R@$K, 0.3$ instances in a wide range of noise levels. In particular, the R@$K, 0.7$ instances shows severely high errors. This is because R@K, θ with large IoU threshold requires exactly localised moments thereby drastically changing evaluation results even with small perturbation in a ground truth. Therefore, the use of a large θ assumes the low-noise condition of human annotations, which is difficult to ensure [2, 15, 26]. On the other hand, the instances of R@K, θ with $\theta = 0.3, 0.5$ show comparable or lower errors than the AxIoU instances because these R@K, θ instances ignore localisation quality.

6.5. Summary: Agreement and Stability

We demonstrated the undesirable properties of R@K, θ from various aspects; (1) the R@$K, 0.3$ and R@$1, \theta$ instances (i.e. non-demanding measures) often disagree with other R@K, θ instances (Section 6.2); (2) the R@$K, 0.3$ and R@$K, 0.5$ instances are unstable to the change of the size of a dataset (Section 6.3); and (3) the R@$K, 0.7$ instances are unstable to label ambiguity and potentially noisy (Section 6.4). By contrast, our AxIoU measure reconciles the agreement and stability while reducing the hyper-parameter θ, which is difficult to tune. Moreover, it should be noted that the cut-off parameter K for AxIoU@K is easier to handle than that of R@K, θ as it considers the ranking quality of top-K ranked lists; the agreement between the AxIoU instances (Section 6.2) is also an evidence for this.

6.6. Model Selection

As discussed in Section 6.3, the stability with respect to the choice of test queries is vital for avoiding inconsistent evaluation on different dataset splits. This is true also in the process of model selection; when we select the best model based on a validation split and evaluate it on a test split, the measurement on the validation split should be consistent with that on the test split.

This section investigates the effectiveness of AxIoU as the criterion for model selection. To this end, we first created 640 variants of the 2DTAN system (See Section 6.1) by varying its hyper-parameters such as the learning rate and the threshold of NMS. Then, based on each instance of

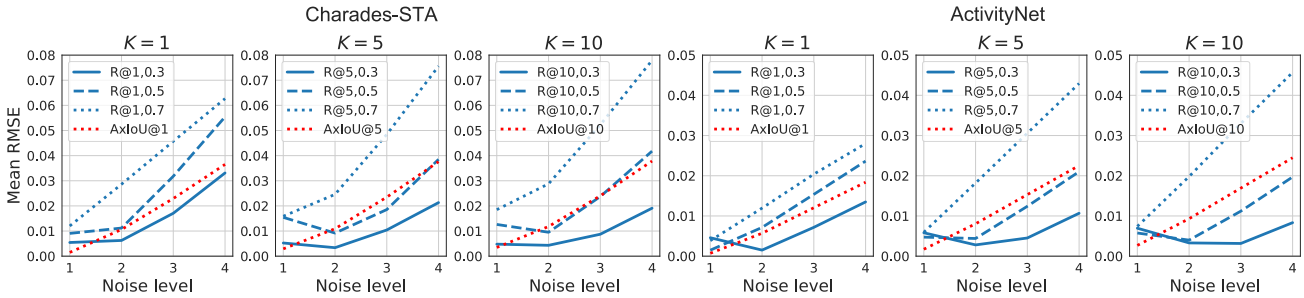

Figure 7. Effect of the ambiguity of annotated temporal regions.

R@K, θ as well as AxIoU (12 measures in total), we select the best model using the validation split. Finally, we evaluate the above 12 models using R@K, θ on the test split. As the R@$10, 0.3$ scores saturate easily (see also Figure 5), we omitted it from the test measures for visibility of figures; thus, we utilise 8 test measures in total. For each of the 8 test measures, we compute the Z-scores of the 12 models so that the average of the 12 scores equals zero.

Figures 8 (a)–(d) show the results for Charades-STA. The x-axis in each figure shows the test measure; each of the 12 lines represents a validation measure; the y-axis shows the Z-scores of "all" the 12 models for each test measure. Since each line represents a single model selected by a particular validation measure, if the line is straight and horizontal, it would imply that the validation measure is useful for effective model selection. R@$10, 0.3$ (blue line in (c)) and R@$5, 0.5$ (orange line in (b)) perform poorly as validation measures: when the models selected according to these measures are evaluated with R@$1, 0.7$ on the test data, these systems are actually the worst among the 12 systems by far. Similarly, R@$10, 0.7$ (green in (c)), R@$1, 0.3$ (blue in (a)), and AxIoU@1 (blue in (d)) perform relatively poorly: for example, when the model selected according to AxIoU@1 is evaluated with R@$10, 0.5$ on the test data, this system is one of the worst performers among the twelve. On the other hand, it can be observed that AxIoU@5, AxIoU@10 and some other R@K, θ instances such as R@$10, 0.5$ (orange in (c)) perform well: that is, the models selected based on these measures generally perform well regardless of what the test measure is. Only AxIoU@10 (green in (d)) could select a system that is above average in terms of all the test measures.

The above result is consistent to the insights obtained in Sections 6.2-6.4. However, the disagreement and instability of each R@K, θ instance is severe for model selection because we must use a single evaluation measure to determine the best model on a validation split. As we cannot know the best setting of K and θ in the validation phase, AxIoU, which is an expectation of R@K, θ (Section 4.2), is a reasonable measure for model selection.

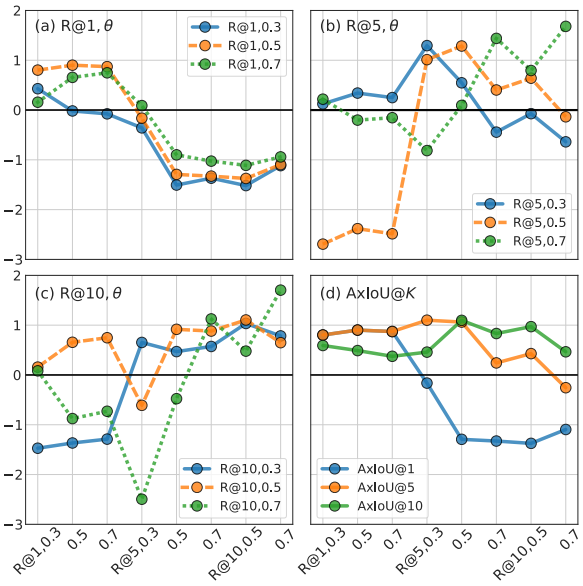

Figure 8. The effect of the validation measure for model selection on effectiveness on the test split.

7. Conclusion

In this paper, we proposed an evaluation measure, AxIoU, for video moment retrieval. AxIoU can offer consistent evaluation compared to R@K, θ without the threshold parameter θ in R@K, θ, which is the main cause of the insensitivity of R@K, θ. We analytically examined the properties of AxIoU through an axiomatic approach and empirically showed that AxIoU@10 can provide stable evaluation while maintaining the similarity to R@K, θ instances. We also demonstrated that AxIoU@10 is a reliable measure for model selection, even if the final test measures are R@K, θ instances. As future work, we will explore a more sophisticated distribution for abandonment position k, $P_A(k)$ [6].

8. Acknowledgement

This work was partly supported by JST CREST Grant No. JPMJCR20D3, FOREST Grant No. JPMJFR216O, and Academy of Finland project number 324346.

References

[1] Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010. 5

[2] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *Eur. Conf. Comput. Vis.*, pages 256–272, 2018. 1, 2, 7

[3] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *Int. ACM SIGIR Conf. on Research and Devet. in Inform. Retrieval*, pages 625–634, 2018. 2

[4] Chris Buckley and Ellen M Voorhees. Retrieval evaluation with incomplete information. In *Int. ACM SIGIR Conf. on Research and Devet. in Inform. Retrieval*, pages 25–32, 2004. 5, 6

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 961–970, 2015. 2, 5

[6] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *ACM Conf. Inform. and Knowledge Management*, pages 621–630, 2009. 2, 3, 8

[7] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019. 2

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015. 4

[9] Hui Fang, Tao Tao, and Chengxiang Zhai. Diagnostic evaluation of information retrieval models. *ACM Trans. Inform. Syst.*, 29(2):1–42, 2011. 2

[10] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5267–5275, 2017. 1, 2, 3, 5

[11] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Int. Conf. Comput. Vis.*, pages 1523–1532, 2021. 2

[12] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. ExCL: Extractive clip localization using natural language descriptions. In *Conf. the North American Ch. the Assoc. Comput. Linguistics: Human Language Tech.*, pages 1984–1990, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[13] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Winter Conf. Apl. Comput. Vis.*, 2020. 2

[14] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, volume 33, pages 8393–8400, 2019. 2

[15] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Int. Conf. Comput. Vis.*, 2017. 1, 2, 7

[16] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4555–4564, 2016. 2

[17] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inform. Syst.*, (4), 2002. 2, 4

[18] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inform. retrieval*, 16(2):138–178, 2013. 2

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, 2017. 2, 5

[20] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. TVR: A large-scale dataset for video-subtitle moment retrieval. In *Eur. Conf. Comput. Vis.*, 2020. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 2

[22] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *Int. ACM SIGIR Conf. on Research and Devet. in Inform. Retrieval*, pages 15–24, 2018. 1, 2

[23] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inform. Syst.*, 27(1):1–27, 2008. 2

[24] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *Int. Conf. Pattern Recog.*, volume 3, pages 850–855. IEEE, 2006. 5

[25] Kemal Oksuz, Baris Cam, Emre Akbas, and Sinan Kalkan. Localization recall precision (lrp): A new performance metric for object detection. In *Eur. Conf. Comput. Vis.*, 2018. 2

[26] Mayu Otani, Yuta Nakahima, Rahtu Esa, and Heikkilä Janne. Uncovering hidden challenges in query-based video moment retrieval. In *Brit. Mach. Vis. Conf.*, 2020. 1, 2, 5, 7

[27] Rafael Padilla, Wesley L. Passos, Thadeu L. B. Dias, Sergio L. Netto, and Eduardo A. B. da Silva. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3), 2021. 2

[28] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Trans. the Assoc. for Comput. Linguistics*, 1:25–36, 2013. 2

[29] Stephen Robertson. A new interpretation of average precision. In *Int. ACM SIGIR Conf. on Research and Devet. in Inform. Retrieval*, pages 689–690, 2008. 2, 3

[30] Tetsuya Sakai. Metrics, statistics, tests. In *PROMISE winter school*, pages 116–163. Springer, 2013. 2, 3

[31] Tetsuya Sakai and Stephen Robertson. Modelling a user population for designing information retrieval metrics. In *NTCIR Workshop*, 2008. 2, 3

[32] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation

measures line up? In *Int. ACM SIGIR Conf. on Research and Devet. in Inform. Retrieval*, pages 555–562, 2010. 2

[33] Fabrizio Sebastiani. An axiomatically derived measure for the evaluation of classification algorithms. In *Int. Conf. The Theory of Inform. Retrieval*, pages 11–20, 2015. 2

[34] Fabrizio Sebastiani. Evaluation measures for quantification: An axiomatic approach. *Inform. Retrieval J.*, 23(3):255–288, 2020. 2

[35] Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *Int. Conf. Comput. Vis.*, pages 2137–2146, 2017. 1, 2, 7

[36] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Eur. Conf. Comput. Vis.*, pages 510–526. Springer, 2016. 5

[37] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7026–7035, 2021. 2

[38] Aming Wu and Yahong Han. Multi-modal circulant fusion for video-to-language and backward. In *Int. Joint Conf. on Artificial Intelligence*, page 1029–1035, 2018. 1, 2

[39] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Adv. Neural Inform. Process. Syst.*, pages 536–546, 2019. 1, 2, 5

[40] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, volume 33, pages 9159–9166, 2019. 2

[41] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10287–10296, 2020. 2

[42] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1247–1257, 2019. 1, 2

[43] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks formoment localization with natural language. In *AAAI*, 2020. 5