AxIoU: An Axiomatically Justified Measure for Video Moment Retrieval

Supplementary Material

A. Proofs

A.1. Definition of Axioms

We formally define the definition of each axiom and each measure in this supplementary material.

Axiom 1 (Invariance against Top-k Non-Best Moment (**INV-k**).)**.** *For any query $q \in \mathcal{Q}$ and any rank position k ($k > 1$), and for all systems σ and σ' such that σ' differs from σ only for the k-th moment in the ranked lists for q, $\mu(\mathcal{Q}, \sigma) = \mu(\mathcal{Q}, \sigma')$ holds when the k-th moments satisfy the following conditions.*

Condition A.1 (Inequality of relevance scores)**.** The relevance scores of the k-th moment in σ_q and σ'_q satisfy $r(\sigma_q(k)) < r(\sigma'_q(k))$.

Condition A.2 (Non-maximum relevance score of the top-k moment)**.** The k-th moment returned by system σ is less relevant than that returned by system σ'. That is, $r(\sigma'_q(k)) \leq \max_{1 \leq j < k} r(\sigma'_q(j))$.

Axiom 2 (Strict Monotonicity for Top-k Best Moment (**MON-k**).)**.** *For any query $q \in \mathcal{Q}$ and any rank position k, and for all systems σ and σ' such that σ' differs from σ only for the k-th moment in the ranked lists for q, $\mu(\mathcal{Q}, \sigma) < \mu(\mathcal{Q}, \sigma')$ holds whenever the k-th moment satisfies Condition A.1 and the following condition.*

Condition A.3 (Maximum relevance score of the top-k moment)**.** The k-th moment returned by σ' is the most relevant within the top k. That is, $r(\sigma'_q(k)) > \max_{1 \leq k' < k} r(\sigma'_q(k'))$ if $k > 1$.

Note that, Condition A.3 is necessary to avoid the contradiction between **INV-k** and **MON-k**.

A.2. Properties of R@K, θ

Mean R@K, θ, is defined as the ratio of queries for which a system successfully retrieves at least one relevant moment with a sufficient IoU with respect to threshold θ [1].

$$\text{Mean R@}K, \theta(\mathcal{Q}, \sigma) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1} \left\{ \sum_{k=1}^{K} \mathbb{1} \left\{ r(\sigma_q(k)) > \theta \right\} > 0 \right\}. \quad (1)$$

Property 1. *Mean R@K, θ does not satisfy **MON-k** (Axiom 2).*

Proof. For two systems σ and σ' such that σ' differs from σ only for k-th moment in the ranked list for q, the difference of the measurements can be expressed as follows:

$$\begin{aligned} &\text{Mean R@}K, \theta(\mathcal{Q}, \sigma) - \text{Mean R@}K, \theta(\mathcal{Q}, \sigma') \\ &= \frac{1}{|\mathcal{Q}|} \mathbb{1} \left\{ \mathbb{1} \left\{ r(\sigma(q)_k) > \theta \right\} + C > 0 \right\} \\ &- \frac{1}{|\mathcal{Q}|} \mathbb{1} \left\{ \mathbb{1} \left\{ r(\sigma'(q)_k) > \theta \right\} + C > 0 \right\}, \end{aligned} \quad (2)$$

where $C = \sum_{1 \leq j \leq K \wedge j \neq k}^{K} \mathbb{1} \left\{ r(\sigma(q)_j) > \theta \right\}$. Here, when $r(\sigma'_q(k)) \leq \theta$ holds, it also holds that $r(\sigma_q(k)) \leq \theta$ by utilising Condition A.1. Then, the k-th moments do not contribute to the measurements, $\mathbb{1} \left\{ r(\sigma_q(k)) > \theta \right\} = \mathbb{1} \left\{ r(\sigma'_q(k)) > \theta \right\} = 0$. Here, because $\theta \geq r(\sigma'_q(k)) \geq \max_{1 \leq j < k} r(\sigma'_q(j))$ holds by Condition A.3, there is no moment that has a sufficient relevance score in the ranked lists σ_q and σ'_q and $C = 0$ holds. Therefore, combining these and Eq. (2), when $r(\sigma'_q(k)) \leq \theta$, we obtain Mean R@$K, \theta(Q, \sigma) =$ Mean R@$K, \theta(Q, \sigma')$, which proves our proposition. ☐

This problem results from the thresholding of temporal IoUs in the measure. This leads to the information loss of the retrieval effectiveness by binarizing the relevance score of moments and thus to the insensitivity of the measure. Property 1 suggests that the measure may ignore the improvement of systems when utilising a large value of θ.

Remarkably, R@K, θ obviously does not satisfy **MON-k** even with assuming $r(\sigma_q(k)) > \theta$ in the case of $K > 1$; when $r(\sigma_q(j)) > \theta$ holds for any rank position j ($1 \leq j \leq K \wedge j \neq k$), $C \geq 1$ in Eq. (2) holds, and thus Mean R@$K, \theta(\mathcal{Q}, \sigma) -$ Mean R@$K, \theta(\mathcal{Q}, \sigma') = (1/|\mathcal{Q}|)(1 - 1) = 0$ holds. Therefore, setting a small value of θ, it also leads to information loss.

Property 2. *Mean R@K, θ satisfies **INV-k** (Axiom 1).*

Proof. Because we may assume that Condition A.2 holds, when $r(\sigma'_q(k)) > \theta$, there is at least one moment in a position j that satisfies $r(\sigma'_q(j)) \geq r(\sigma'_q(k)) > \theta$, and thus, $C \geq 1$ holds in Eq. (2). When $r(\sigma'_q(k)) \leq \theta$, the k-th moment does not contribute to the measurement, and $\mathbb{1} \left\{ r(\sigma'_q(k)) > \theta \right\} = 0$ holds. Therefore, by utilising Condition A.1, $r(\sigma_q(k)) \leq r(\sigma'_q(k)) \leq \theta$, we have,

$$\begin{aligned} &\text{Mean R@}K, \theta(\mathcal{Q}, \sigma) - \text{Mean R@}K, \theta(\mathcal{Q}, \sigma') \\ &= \frac{1}{|\mathcal{Q}|} \mathbb{1} \left\{ C > 0 \right\} - \frac{1}{|\mathcal{Q}|} \mathbb{1} \left\{ C > 0 \right\} = 0, \end{aligned} \quad (3)$$

regardless of $r(\sigma'(q)_k) > \theta$ or $r(\sigma'(q)_k) \leq \theta$. Thus, we obtain Mean R@$K, \theta(Q, \sigma)$ = Mean R@$K, \theta(Q, \sigma')$, which proves our proposition. □

This result suggests that the thresholding and indicator function in R@K, θ play a vital role in ensuring invariance against the redundant moments in the lower rank positions. Although these mechanisms are indispensable as the invariance is required under the problem settings of VMR, they are the main causes of information loss (See Property 1).

A.3. Properties of AP Measures

Using the average precision (AP) measure is one approach to consider the rank of relevant moments [4]. AP and Mean AP (a.k.a. mAP) can be expressed as follows:

$$\text{AP@}K, \theta(q, \sigma) := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}\{r(\sigma_q(j) > \theta)\}. \quad (4)$$

$$\text{Mean AP@}K, \theta(\mathcal{Q}, \sigma) := \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}\{r(\sigma_q(j) > \theta)\}. \quad (5)$$

As the AP measure is for binary relevance grades, it also requires a thresholding process for IoU values.

Property 3. *Mean AP@K, θ does not satisfy **INV-k** (Axiom 1).*

Proof. For two systems σ and σ' such that σ' differs from σ only for k'-th moment in the ranked list for q, the difference of the measurements can be expressed as follows:

$$\begin{aligned} &\text{Mean AP@}K, \theta(\mathcal{Q}, \sigma') - \text{Mean AP@}K, \theta(\mathcal{Q}, \sigma) \\ &= \frac{1}{|\mathcal{Q}|K} \sum_{k=1}^{K} \frac{1}{k} \sum_{j=1}^{k} \left(\mathbb{1}\{r(\sigma'_q(j) > \theta)\} - \mathbb{1}\{r(\sigma_q(j) > \theta)\}\right) \\ &= \frac{1}{|\mathcal{Q}|K} \sum_{k=1}^{K} \frac{1}{k} \left(\mathbb{1}\{r(\sigma'_q(k') > \theta)\} - \mathbb{1}\{r(\sigma_q(k') > \theta)\}\right). \end{aligned} \quad (6)$$

To derive the second equality, we assume that the top-$(k'-1)$ ranked lists of σ_q and σ'_q are identical, and the partial ranked lists from the $(k'+1)$-th position are also identical. When $r(\sigma'_q(k')) > \theta \geq r(\sigma_q(k'))$ holds, we have the following: $\mathbb{1}\{r(\sigma'_q(k') > \theta)\} - \mathbb{1}\{r(\sigma_q(k') > \theta)\} = 1 - 0 = 1$. Therefore, we can obtain the following:

$$\begin{aligned} &\text{Mean AP@}K, \theta(\mathcal{Q}, \sigma') - \text{Mean AP@}K, \theta(\mathcal{Q}, \sigma) \\ &= \frac{1}{|\mathcal{Q}|K} \sum_{k=1}^{K} \frac{1}{k} > 0 \\ &\iff \text{Mean AP@}K, \theta(\mathcal{Q}, \sigma') > \text{Mean AP@}K, \theta(\mathcal{Q}, \sigma). \end{aligned}$$

□

AP cannot handle the redundant moments in a ranked list because each top-K ranked relevant moment contributes to the measurement as an equally relevant one; in other words, AP is concerned with the number of the relevant moments in a ranked list. It suggests that a system without NMS can unfairly take an advantage in the evaluation based on AP.

Property 4. *Mean AP@K, θ does not satisfy **MON-k** (Axiom 1).*

Proof. In Eq. (6), when $\theta \geq r(\sigma'_q(k')) > r(\sigma_q(k'))$ and Condition A.3 hold, $\mathbb{1}\{r(\sigma'_q(k') > \theta)\} - \mathbb{1}\{r(\sigma_q(k') > \theta)\} = 0 - 0 = 0$. Therefore, we obtain the following:

$$\begin{aligned} &\text{Mean AP@}K, \theta(\mathcal{Q}, \sigma') - \text{Mean AP@}K, \theta(\mathcal{Q}, \sigma) \\ &= 0 \\ &\iff \text{Mean AP@}K, \theta(\mathcal{Q}, \sigma') = \text{Mean AP@}K, \theta(\mathcal{Q}, \sigma). \end{aligned}$$

□

Although AP is rank-sensitive, it has the threshold θ as in R@K, θ and can ignore the improvement of IoU values of relevant moments.

A.4. Properties of DCG-type Measures

The naïve approach to remove thresholding parameter θ while considering the rank positions of relevant moments is to utilise the measures for multiple relevance grades, such as normalised discounted cumulative gain (nDCG) [2] because an IoU value can be considered as a continuous relevance score. A DCG-type measure can be expressed as follows:

$$\text{DCG@}K(q, \sigma) := \sum_{k=1}^{K} \frac{g(r(\sigma(q)_i))}{d(k)}. \quad (7)$$

$$\text{Mean DCG@}K(\mathcal{Q}, \sigma) := \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{k=1}^{K} \frac{g(r(\sigma(q)_i))}{d(k)}, \quad (8)$$

where $g(r) \geq 0$ and $d(k) > 0$ denotes the gain and discounting functions that are strictly monotonically increasing with respect to the relevance score r and rank position k, respectively. However, any DCG-type measure obviously does not satisfy **INV-k** (Axiom 1).

Property 5. *Mean DCG@K does not satisfy **INV-k** (Axiom 1).*

Proof. For any query $q \in \mathcal{Q}$ and any rank position k, and for all systems σ and σ' such that σ' differs from σ for only the k-th moment in the ranked lists for q, we can obtain the following equation.

$$
\begin{aligned}
&\text{Mean DCG@}K(\mathcal{Q},\sigma') - \text{Mean DCG@}K(\mathcal{Q},\sigma)\\
&= \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\sum_{j=1}^{K}\frac{g(r(\sigma'_q(j)))}{d(j)} - \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\sum_{j=1}^{K}\frac{g(r(\sigma_q(j)))}{d(j)}\\
&= \frac{1}{|\mathcal{Q}|}\frac{g(r(\sigma'_q(k))) - g(r(\sigma_q(k)))}{d(k)}. \qquad (9)
\end{aligned}
$$

Based on Condition A.1, $r(\sigma_q(k)) < r(\sigma'_q(k))$ and the strict monotonicity of the gain function $g(\cdot)$, we have the following: Mean DCG@$K(\sigma')$ − Mean DCG@$K(\sigma) > 0$, which completes the proof. □

Because DCG@K is defined as the sum of element-wise discounted gain values of moments in a ranked list, it cannot handle the redundant moments in a ranked list appropriately. Moreover, it is difficult to utilise DCG-type measures with normalisation (nDCG) for VMR as the definition of the ideal list is not trivial owing to **INV-k**; thus, DCG@K is under-normalised and can take a large value for a single query.

Property 6. *Mean DCG@K satisfies **MON-k** (Axiom 1).*

Proof. In Eq. (9), by the strict monotonicity of the gain function $g(\cdot)$ and non-negativity of the discount function $d(\cdot)$, $g(r(\sigma'_q(k))) - g(r(\sigma_q(k))) > 0$, and thus, Mean DCG@$K(\sigma')$ − Mean DCG@$K(\sigma) > 0$ always holds, which completes the proof. □

DCG@K is thresholding-free and rank-sensitive, and thereby satisfies **MON-k**; however, due to these properties, it does not satisfies **INV-k**.

By considering the properties of R@K,θ and DCG@K, it is challenging for conventional information retrieval measures to satisfy **INV-k** and **MON-k** simultaneously.

A.5. Properties of AxIoU Measure

In this section, we demonstrate that AxIoU is thresholding-free and rank-sensitive while satisfying both **INV-k** and **MON-k**. AxIoU can be expressed as follows:

$$\text{AxIoU@}K(q,\sigma) \coloneqq \frac{1}{K}\sum_{k=1}^{K}\max_{1\le j\le k} r(\sigma_q(j)), \qquad (10)$$

$$\text{Mean AxIoU@}K(\mathcal{Q},\sigma) \coloneqq \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{K}\sum_{k=1}^{K}\max_{1\le j\le k} r(\sigma_q(j)). \qquad (11)$$

Property 7. *Mean AxIoU@K satisfies **INV-k** (Axiom 1).*

Proof. For two systems σ and σ' such that σ' differs from σ only for the k'-th moment in the ranked list for q,

$$
\begin{aligned}
&\text{Mean AxIoU@}K(\mathcal{Q},\sigma') - \text{Mean AxIoU@}K(\mathcal{Q},\sigma)\\
&= \frac{1}{|\mathcal{Q}|K}\sum_{k=k'}^{K}\left(\max_{1\le j\le k} r(\sigma'_q(j)) - \max_{1\le j\le k} r(\sigma_q(j))\right) \qquad (12)
\end{aligned}
$$

By utilising $r(\sigma'_q(k')) \le \max_{1\le j<k'} r(j)$ and $r(\sigma_q(k')) < r(\sigma'_q(k'))$, it holds that $r(\sigma_q(k')) \le \max_{1\le j<k'} r(\sigma_q(j))$ because the top-$(k'-1)$ lists of σ and σ' are identical. Therefore, $\max_{1\le j\le k'} r(\sigma'_q(j)) = \max_{1\le j\le k'} r(\sigma_q(j))$ holds. In addition, because the partial ranked lists of σ and σ' from the $(k'+1)$-th position are identical, we have $\max_{1\le j\le k} r(\sigma'_q(j)) = \max_{1\le j\le k} r(\sigma_q(j))$ for any position $k (1 \le k \le K)$. Therefore, we have the following:

$$
\begin{aligned}
&\sum_{k=k'}^{K}\left(\max_{1\le j\le k} r(\sigma'_q(j)) - \max_{1\le j\le k} r(\sigma_q(j))\right) = 0\\
&\iff \text{Mean AxIoU@}K(\mathcal{Q},\sigma') = \text{Mean AxIoU@}K(\mathcal{Q},\sigma).
\end{aligned}
$$

□

AxIoU determines the contribution of the relevant moments in a ranked list by comparing the relevance of these moments. Therefore, it can handle the redundant moments without any binarisation and thresholding processes.

Property 8. *Mean AxIoU@K satisfies **MON-k** (Axiom 2).*

Proof. For two systems σ and σ' such that σ' differs from σ only for k'-th moment in the ranked list for q, the evaluation measures can be expressed as follows:

$$
\begin{aligned}
&\text{Mean AxIoU@}K(\mathcal{Q},\sigma') - \text{Mean AxIoU@}K(\mathcal{Q},\sigma)\\
&= \frac{1}{|\mathcal{Q}|K}\sum_{k=k'}^{K}\left(\max_{1\le j\le k} r(\sigma'_q(j)) - \max_{1\le j\le k} r(\sigma_q(j))\right)\\
&= \frac{1}{|\mathcal{Q}|K}\Bigg(r(\sigma'_q(k')) - \max_{1\le j\le k'} r(\sigma_q(j))\\
&\quad + \sum_{k=k'+1}^{K}\left(\max_{1\le j\le k} r(\sigma'_q(j)) - \max_{1\le j\le k} r(\sigma_q(j))\right)\Bigg) \qquad (13)
\end{aligned}
$$

In the second equality, we utilised $r(\sigma'_q(k')) = \max_{1\le j<k'} r(\sigma'_q(j))$ to derive the first term in the right hand side. By utilising $r(\sigma'_q(k')) > \max_{1\le j<k'} r(\sigma'_q(j))$ and $r(\sigma'_q(k')) > r(\sigma_q(k'))$, $r(\sigma'_q(k')) - \max_{1\le j<k'} r(\sigma_q(j)) > 0$ holds in the right hand side of the second equality. For the second term, because we may assume that the partial ranked lists of σ and σ' from the $(k'+1)$-th position are identical, $\max_{1\le j\le k} r(\sigma'_q(j)) \ge \max_{1\le j\le k} r(\sigma_q(j))$ holds for any position k ($k' < k \leq K$). Thus, the following inequality holds.

$$r(\sigma'_q(k')) - \max_{1\leq j\leq k'} r(\sigma_q(j)) > 0$$
$$\wedge \sum_{k=k'+1}^{K} \left(\max_{1\leq j\leq k} r(\sigma'_q(j)) - \max_{1\leq j\leq k} r(\sigma_q(j)) \right) \geq 0$$
$$\iff \text{Mean AxIoU@}K(\mathcal{Q}, \sigma') > \text{Mean AxIoU@}K(Q, \sigma),$$

which completes the proof. □

As a summary, AxIoU reflects the rank positions of the relevant moments in a ranked list as in AP and considers IoU values as in DCG, while it can handle the redundant moments as in R@K, θ.

B. Analysis of Number of Tied Results

To demonstrate the behaviours of R@K, θ and AxIoU@K, we investigate the number of queries for which the 6 systems have the exactly same score for each VMR measure; we define the ratio of such queries in all test queries as *all-tied query ratio* of a measure. Figure 1 shows the all-tied query ratio of each measure on Charades-STA and ActivityNet, respectively. From Figure 1, the R@K, θ instances with relaxed or demanding settings, such as R@$10, 0.3$, R@$1, 0.7$, show higher all-tied query ratios than the other instances. It indicates that these measures cannot distil any information from the evaluation results based on a large number of queries. R@$5, 0.7$ performs well in both Charades-STA and ActivityNet. On the other hand, the AxIoU@K instances show substantially lower all-tied query ratios for $K = 1, 5, 10$. It is remarkable that, with a larger K, AxIoU@K performs well whereas R@K, θ with $\theta = 0.3, 0.5$ becomes worse. Probably, it is because AxIoU@K can leverage the information of the lower positions in ranked lists owing to its rank-sensitivity, whereas R@K, θ, which is a set retrieval measure, becomes insensitive when with a large K and requires a large θ to detect the difference of systems. This suggests that the setting of θ is rather difficult when K is large such as in the TVR dataset [3]; an extremely large θ may be required although it can make difficult queries uninformative.

C. On the Stability to Label Ambiguity

To show the stability to label ambiguity of the measures, we showed the behaviour of the measures through numerical experiments (Section 6.4). In this section, we discuss the effect of the IoU thresholding on the estimation stability.

We here show the case of $K = 1$ for a simple example. Let r be the IoU value of the top-1 moment for the true unobservable ground truth and $\hat{r}$ be that for the noisy ground truth. Under a Gaussian noise model $\hat{r} = r + \epsilon$, where

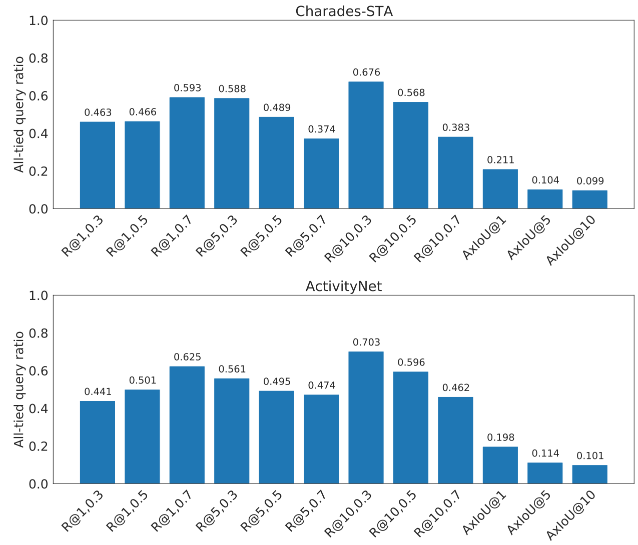

Figure 1. All-tied query ratio of each measure.

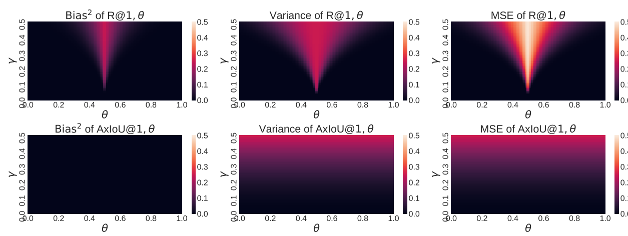

Figure 2. Effect of θ and γ on estimation errors.

$\epsilon \sim N(0, \gamma^2)$, noisy IoU $\hat{r}$ also obeys a normal distribution $\hat{r} \sim N(r, \gamma^2)$. The expected difference between the true and observed AxIoU@1 (*i.e.* bias) is obtained as $\mathbb{E}_{\hat{r}}[r - \hat{r}] = r - \mathbb{E}_{\hat{r}}[\hat{r}]$. Hence, AxIoU@1 is unbiased (*i.e.* $\mathbb{E}_{\hat{r}}[r - \hat{r}] = 0$) because $\hat{r}$ is an unbiased estimator of r (*i.e.* $\mathbb{E}_{\hat{r}}[\hat{r}] = r$). The variance of AxIoU@1 is exactly that of $\hat{r}$ (*i.e.* $\mathbb{V}[\hat{r}] = \gamma^2$).

On the other hand, the bias of R@$1, \theta$ can be obtained as

$$\mathbb{E}_{\hat{r}}[\mathbb{1}\{\hat{r} \geq \theta\} - \mathbb{1}\{r \geq \theta\}] = \mathbb{E}_{\hat{r}}[\mathbb{1}\{\hat{r} \geq \theta\}] - \mathbb{1}\{r \geq \theta\}$$
$$= P(\hat{r} \geq \theta) - \mathbb{1}\{r \geq \theta\}.$$

If $\theta \leq r$ holds, because the true R@$1, \theta$ is one, the bias is then $P(\hat{r} \geq \theta) - 1 = -P(\hat{r} < \theta)$. If $\theta > r$ holds, the bias is $P(\hat{r} \geq \theta)$. Therefore, R@$1, \theta$ is statistically biased; that is, it has the error even in the expectation. Because $\mathbb{1}\{\hat{r} \geq \theta\}$ obeys the Bernoulli distribution $\mathrm{Bern}(P(\hat{r} \geq \theta))$, The variance of R@$1, \theta$ is $P(\hat{r} \geq \theta)P(\hat{r} < \theta)$, which depends on θ and γ. Figure 2 shows the theoretical (squared) bias, variance and mean squared error (MSE) of AxIoU@1 and R@$1, \theta$ for different θ and γ under $r = 0.5$. We can observe that both AxIoU@1 and R@$1, \theta$ have large estimation errors (*i.e.* MSE) when noise level γ is large. In addition to this, R@$1, \theta$ suffers from a severe error even with small γ, particularly when θ is close to $r = 0.5$. This is an undesirable property because we often need to discriminate competitive VMR methods and thus to use θ around the boundary, which leads to estimation errors under label noise.

References

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

[3] 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. 4

[4] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008. 2