Are Binary Annotations Sufficient? Video Moment Retrieval via Hierarchical Uncertainty-based Active Learning

Wei Ji[1] Renjie Liang[1] Zhedong Zheng[1]* Wenqiao Zhang[2] Shengyu Zhang[2]
Juncheng Li[2] Mengze Li[2] Tat-seng Chua[1]
[1]National University of Singapore [2]Zhejiang University
{jiwei,t0924327,zdzheng,dcsts}@nus.edu.sg, {wenqiaozhang,sy_zhang,junchengli,mengzeli}@zju.edu.cn

Abstract

Recent research on video moment retrieval has mostly focused on enhancing the performance of accuracy, efficiency, and robustness, all of which largely rely on the abundance of high-quality annotations. While the precise frame-level annotations are time-consuming and cost-expensive, few attentions have been paid to the labeling process. In this work, we explore a new interactive manner to stimulate the process of human-in-the-loop annotation in video moment retrieval task. The key challenge is to select "ambiguous" frames and videos for binary annotations to facilitate the network training. To be specific, we propose a new hierarchical uncertainty-based modeling that explicitly considers modeling the uncertainty of each frame within the entire video sequence corresponding to the query description, and selecting the frame with the highest uncertainty. Only selected frame will be annotated by the human experts, which can largely reduce the workload. After obtaining a small number of labels provided by the expert, we show that it is sufficient to learn a competitive video moment retrieval model in such a harsh environment. Moreover, we treat the uncertainty score of frames in a video as a whole, and estimate the difficulty of each video, which can further relieve the burden of video selection. In general, our active learning strategy for video moment retrieval works not only at the frame level but also at the sequence level. Experiments on two public datasets validate the effectiveness of our proposed method. Our code is released at https://github.com/renjie-liang/HUAL.

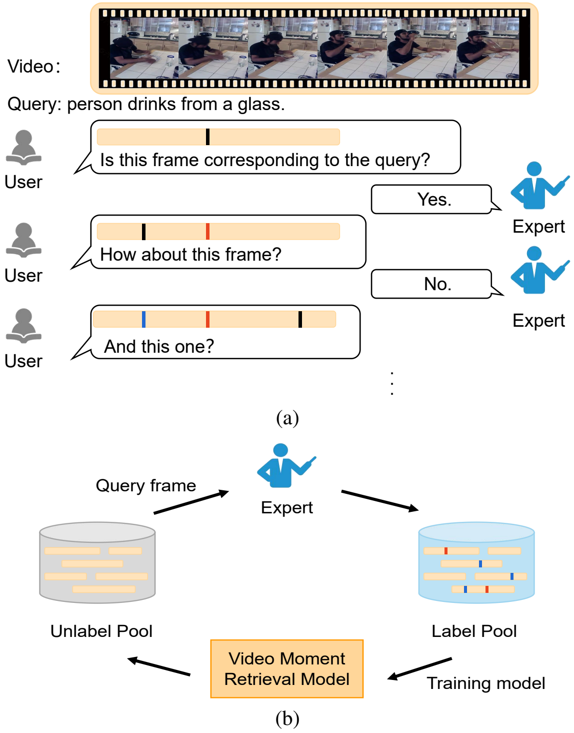


Figure 1. We propose a new interactive method named HUAL which only requires binary annotations to reduce the annotation cost. (a) In each round, user (student) selects a frame with the largest uncertainty, and the expert (teacher) returns the binary label of this frame as feedback. (b) With more labels provided, the VMR model is retrained and the whole process can be treated in a human-in-the-loop manner.

1. Introduction

Video Moment Retrieval (VMR) aims to localize the temporal region of an untrimmed video corresponding to query description, which is a fundamental task in the video understanding area, and can benefit a lot of downstream tasks, such as video question answering [19, 40, 53], dense video captioning [4, 6], video relation detection [14, 34], video dialog [29], *etc*. Recent methods on VMR mainly focus on modeling the cross-modal context in temporal, and have achieved significant performance gains in public datasets, which heavily rely on the well-annotated datasets, such as Charades [9], ActivityNet [35], *etc*.

*Corresponding author.

However, the precise labels on VMR datasets are time-consuming and cost-expensive, and relying on the well-annotated dataset will restrict the generalization ability of the current models. Hence, we revisit the process of annotation and propose a new active learning-based method to relieve the heavy burden of annotations in the VMR task. We have two assumptions: 1) not each frame should be considered equally, as the frame with the higher uncertainty is more valuable than the rest; and 2) not each video can be treated as a hard sample, annotating complex video and query pairs first benefits more than annotating simple ones.

The whole process of our active learning-based method is shown in Figure 1. In each round, for each video in the training set, the user (student) first selects one frame with the highest uncertainty regarding the consistency of the video and query in this video, then the expert (teacher) returns the label of this frame (positive or negative). After that, the user takes the label of a single frame as supervision and trains the model, and the uncertainty score of each video will be updated. In the next round, the user can select another frame with the highest uncertainty. The whole process can be described as: "A student asks a hard question first, then the teacher returns the answer as feedback, the student then digests what have learned, and the process can be repeated." Besides, the amount of videos that need to be annotated can be further compressed. The uncertainty of each video can also be utilized in the sequence level. A certain percentage of videos with high uncertainty can be treated as hard samples with more benefits when annotating.

Hence, the whole process of interactive annotating can be treated as Human-in-the-Loop. The key techniques rely on the computation of uncertainty and the selection of the video frame or the whole video sequence as hard samples. To be specific, we consider the uncertainty in two aspects: the classification confidence of each frame, and the distance of the current frame from the known labels (*e.g.*, the start and end boundaries of each video are negative samples, and the annotated frames in previous rounds). For the classification confidence of each frame, we choose a weakly-supervised VMR model (such as CPL [50]) to obtain the initial classification result and confidence score of each frame, and the frame with alow confidence score can be treated with high uncertainty. By adding the distance score and confidence score together, the frame with the highest uncertainty is selected. We then utilize a fully-supervised VMR model (such as SeqPAN [45]) to train with the labels provided by the expert in each round. Besides considering the frame-level uncertainty, we also seek the reduction of annotated videos at the sequence level via accumulating the frame-level uncertainty. Hence, the annotation cost can be further reduced with a minor performance drop.

Our main contributions are summarized as follows:

- We propose a new interactive framework named HUAL to reduce the annotation cost, which only requires binary annotations. To verify the feasibility, we stimulate the process of annotation in the video moment retrieval task, which is model-agnostic and can be treated in a Human-in-the-Loop manner.

- Specifically, we consider the hierarchical design, which is frame-level and sequence-level uncertainty estimation to select hard samples and fully take advantages of limited binary annotations by the expert. This annotation method can greatly reduce the annotation cost while achieving comparable performance compared with the fully supervised setting.

- Extensive experimental results on two public datasets indicate that binary annotations are sufficient for video moment retrieval. The proposed method can achieve competitive performance with much fewer annotations, which show the effectiveness of our proposed methods.

2. Related Work

2.1. Video Moment Retrieval

Given the descriptive query, Video Moment Retrieval (VMR) aims to retrieve video segments with consistent semantics of query [21–23, 41, 43], which is also relevant to video grounding [13, 16–18]. Most works focus on achieving satisfying performance in fully-supervised or weakly-supervised setting. In early works, some proposal-based methods [1, 11, 24, 38] treat this task as a ranking problem and follow the propose-and-rank pipeline. By generating proposals in various lengths by sliding window [9] first, these methods calculate the cross-modal semantic similarity to find the best matching proposal for the query. However, densely sampling video moment proposals will leads to large computation costs. Then, some proposal-free methods are proposed [25,44,48]. For example, Zhang *et al.* [47] propose a 2D temporal map to model the temporal relations of different moments with variant length, and the two dimensions indicate the start and end timestamps, respectively. The performance of all these methods mentioned above heavily relied on the well-annotated datasets, while we propose a new interactive labeling method with only binary annotations to achieve satisfying results.

To relieve the heavy burden of annotation, in the weakly-supervised setting, only video-text pairs can be treated as supervision, no temporal information is provided. Existing supervised VMR methods can be categorized into two groups. 1) **Multiple Instance Learning (MIL) based [27, 37]:** They treat video-query pairs as positive samples, and video with other queries and query with other videos as negative samples. Then, they train the model by maxi-

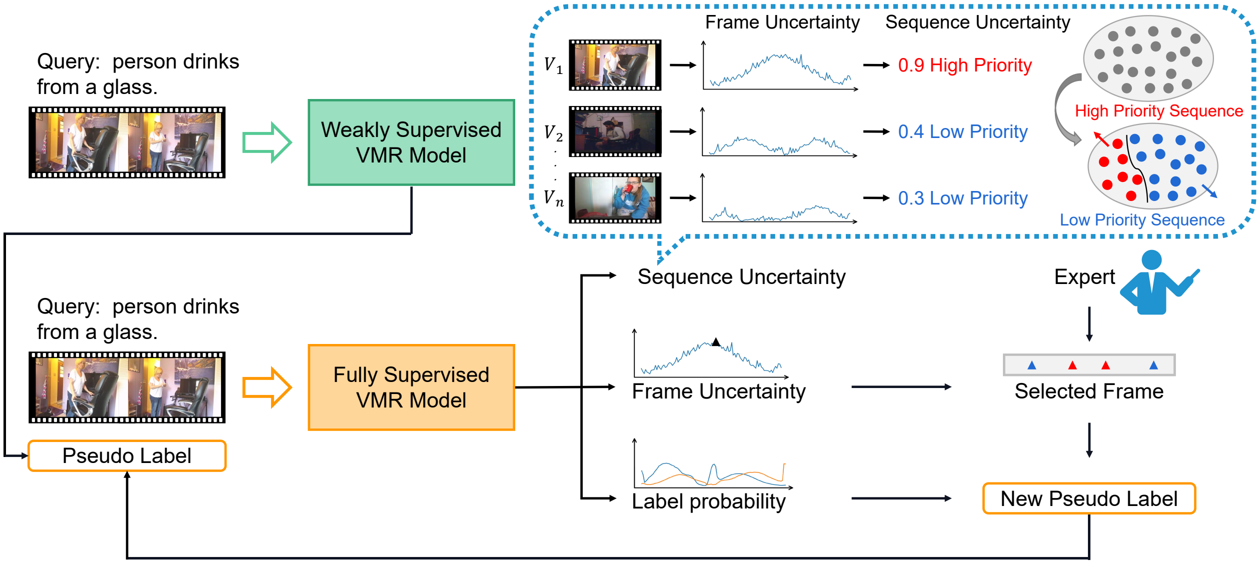


Figure 2. The whole pipeline of our HUAL method. Given video and query pair, we first obtain the pseudo label via a weakly supervised VMR model. Then we calculate the frame-level and sequence-level uncertainty by training a fully supervised VMR, and select the frame and video with highest uncertainty to be annotated by the expert.

mizing similarity score between positive samples and minimizing similarity scores between negative samples. 2) **Reconstruction-based method [20, 36, 50]:** They take advantage of the assumption that the video segment that can best reconstruct the text query is close to the ground-truth. Compared with fully supervised methods, there exists a large performance gap although weakly supervised methods are trained without precise temporal information. Hence, we propose to further improve the performance of weakly supervised VMR method with minor annotation cost, such as binary labels.

2.2. Active Learning

Active learning plays an important role in machine learning area, which aims to maximize a model's performance gain while annotating the fewest samples possible [31]. Current methods for active learning can be categorized into three classes: uncertainty-based method [2, 15, 30, 51]; diversity-based method [8, 10, 32], and expect model change [7,33,52]. To be specific, uncertainty-based method uses the probability distribution of prediction, which is simple in form and has low computational complexity. The diversity-based method is the second category selecting diverse samples that expanse the input space maximally and represent the whole distribution of the unlabeled pool. And the last category is based on model performance change, which selects the data points that would cause the greatest change to the current model parameters and encourage optimal model improvement. In the video moment retrieval task, few works pay attention to reducing the annotation burden, the most similar work to ours is ViGA [5], which takes one single frame among the groundtruth as supervision. Different from this work, we sample the frame with highest uncertainty to be annotated without any prior, we don't restrict the sample is positive or negative, which is a more natural and looser setting.

3. Method

We introduce a novel pipeline named Hierarchical Uncertainty-based Active Learning (HUAL) as shown in Fig. 2. In this section, we first provide the problem definition in Sec. 3.1. HUAL includes two parts: frame-level and sequence-level uncertainty estimation, which are introduced in Sec. 3.2 and Sec. 3.4, respectively. We finally introduce the training and testing phrase of the whole model.

3.1. Problem Definition

Given an untrimmed video $V = \{v_t\}_{t=1}^{T}$ and the language query $Q = \{q_j\}_{j=1}^{M}$, where T and M are the number of frames and words, respectively, our goal is to predict the start and end timestamp (τ^s, τ^e) in the video corresponding to query Q, where $(\tau^s, \tau^e) = f(V, Q)$. In this paper, we explore an interactive framework with few selective frames annotated, rather than complete groundtruth of (τ^s, τ^e) in fully supervised setting.

From the perspective of annotation, each frame in the video can be classified as relevant or irrelevant with the query Q, which can be treated as positive or negative samples. The naive and intuitive idea is selecting frames among the video randomly, which is inefficient and heavily relied on the distribution of groundtruth (τ^s, τ^e). For example, if the (τ^s, τ^e) is very short, it will cost numerous rounds before sampling the frame within the groundtruth. Hence, we consider a more effective sampling method to further reduce the annotation cost from two aspects: Frame-level and Sequence-level uncertainty estimation.

3.2. Frame-level Uncertainty

For each video V and query Q pair, we feed them into a weakly-supervised VMR model. Here, we select CPL [50] as example. Then, the model will output the labels of binary classification result v^{class} as initial pseudo labels.

Then, we consider how to effectively select the frame to be annotated by the expert. The selection of a frame is related to two factors: the uncertainty of the selected frame, and the distance to frames with known labels. For example, the start and end boundary of the video sequence can be treated as negative. And the labels returned by the expert in previous rounds are also known. Hence, the uncertainty score of each frame can be calculated as:

$$U_i^{frame} = U_{model}(f_{model}(v_i)) + \alpha * U_{dis}(v_i) \qquad (1)$$

where α is the weighted factor, U_{model} represents the uncertainty score of the fully supervised model f_{model}, such as SeqPAN [45]. Uncertainty denotes the difference of v_i with and without dropout layer. We train the SeqPAN model from scratch with pseudo label v^{class}. Actually, the output of SeqPAN is two curves represent the start and end possibility of each frame. Here we use $f_{model}(v_i)$ to represent the output for convenience. $U_{dis}(v_i)$ represents the distance to known labels, the distance uncertainty will be higher if the selected frame is farther away known labels. If there are m labeled frame with positive/negative values, then there are $m+2$ frames with label, including the start and end boundary of video v. $U_{dis}(v_i)$ is the superposition of $m+1$ Gaussian curves. For m-th Gaussian curve, the peak position is in the middle of $m-1$-th and m-th frame, as shown in Fig. 3.

Then, the frame with the largest uncertainty score in the video will be selected to be annotated by the expert.

3.3. Pseudo Label Generation

After the expert provides the label of the frame (positive or negative), the surrounding frames within a range will also be annotated with the same label.

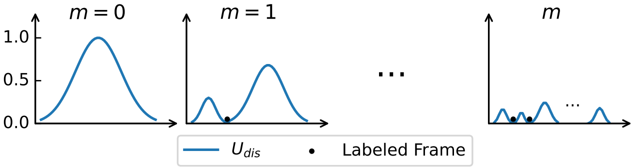


Figure 3. Distance uncertainty of U_{dis}.

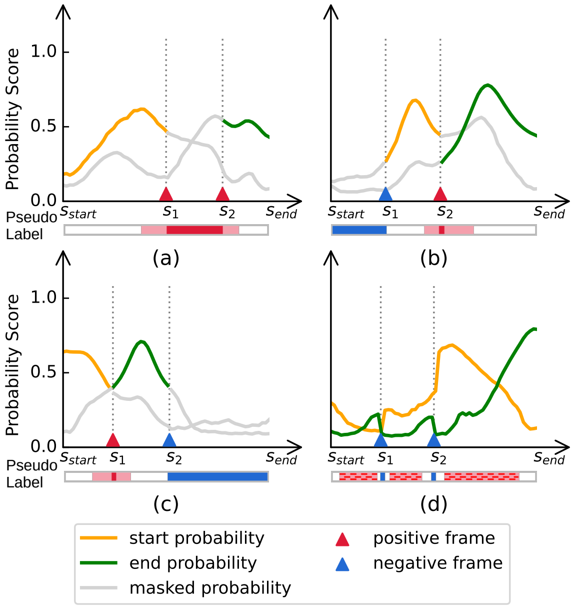


Figure 4. Since there is a prior that only one continuous range is corresponding to the query. After several rounds of annotation with positive/negative frames, there are some rules to generate informative labels. In the probability score curve, yellow/green curve represents start/end timestamp. Red/blue triangle represents positive/negative sample, respectively. In the pseudo label, light red region represents possible positive frames in video, dark red means absolute positive, blue means negative frames. Red dotted line represents the positive regions tcan possibly occur in three intervals.

Some rules can be utilized to eliminate the uncertainty section, based on the prior of only one continuous interval corresponding to the query. For example:

1) If there are two positive samples s_1 and s_2 as shown in Fig 4 (a), then the region in pseudo label between s_1 and s_2 must be positive, and the frames after s_1 can not be start frames, which are masked with zero (grey curve). Accordingly, frames before s_2 can not be end frames, which are also masked with zero.

2) If s_1 is negative and s_2 is positive as shown in Fig 4 (b), then the corresponding region must be after s_1 and includes s_2, so the start frame is within (s_1, s_2), end frame is within (s_2, s_{end}). In the pseudo label, frames within (s_{start}, s_1) are negative, frames surround s_2 are positive in Gaussian distribution. Fig 4 (c) is the mirror version of Fig 4 (b) but in different order of sampled frame.

3) If the two annotated frame s_1 and s_2 are both negative as shown in Fig 4 (d), then the positive frames are possible to locate among (s_{start}, s_1), (s_1, s_2), (s_2, s_{end}). The only useful information is the negative frames surrounding s_1 and s_2.

Based on the rules above, the label of each frame L^{frame} is updated. For the r round, the pseudo label is calculated as:

$$L_r^{frame} = L_{r-1}^{frame} + \beta * P_{model}(v) + \gamma * P_{dis}(v) \quad (2)$$

where β and γ are the weighted parameters, $P_{model}(v)$ is the probability score of whole video sequence v, which is the output of SeqPAN, $P_{dis}(v)$ is the probability of distance between samples, the computation of $P_{dis}(v)$ is similar as $U_{dis}(v_i)$, the only difference is we add an offset in the $P_{dis}(v)$.

3.4. Sequence-level Uncertainty

Moreover, we think in the whole training set, not all videos should be treated equally. Some video and query pairs with simple semantics (such as the description of common objects and actions) can be treated as simple samples, and annotating hard samples is more valuable than simple ones.

Hence, considering the whole training dataset $M_{train} = \{V\}_{i=1}^{N}$, we calculate the uncertainty score U^{seq} of the whole video sequence:

$$U^{seq} = \sum_{i=1}^{n} U_i^{frame} \quad (3)$$

which is the accumulation of frame-level uncertainty. And it is reasonable that the video with more frames with uncertainty should be annotated first.

After calculating the uncertainty score of all videos, we sort them in descending order according to their uncertainty score and choose $K\%$ top-ranked from N videos to annotated and update the labels.

Training. In the training stage, the output of SeqPAN [45] model is the probability distributions of start/end boundaries $P_{s/e}$. The training objective is:

$$\mathcal{L}_{\text{loc}} = \frac{1}{2} \times \left[f_{\text{CE}}(P_s, Y_s) + f_{\text{CE}}(P_e, Y_e) \right] \quad (4)$$

where f_{CE} is the cross-entropy function, $Y_{s/e}$ is the one-hot labels for start/end (i^s/i^e) boundaries.

Since we only have pseudo labels to train the SeqPAN, which is not precise as ground truth, we propose soft label to replace hard label: For the start/end frame, we use a Gaussian distribution to model the labels of surrounding frames, where the peak position of Gaussian is the start/end frame.

The overall training loss of SeqPAN is to minimize the combined loss of $\mathcal{L}_{\text{loc}}$ and supervision to intermediate features during the training process. Considering the uncertainty regularization [51], the rectified loss is:

$$\mathcal{L}_{\text{u}} = \frac{\mathcal{L}_{\text{loc}}}{exp(\sigma)} + \sigma, \quad (5)$$

where σ is the variance of the prediction $P_{model}(V)$, and $\sigma \geq 0$. If the prediction is consistent, the σ is close to zero, and the loss will converge to the original $\mathcal{L}_{\text{loc}}$. If the prediction fluctuates, which indicates the pseudo label may contain noise, we decrease the punishment on the such label.

Inference. When testing, the predicted start and end boundaries of the given video-query pair (V, Q) are generated by maximizing the joint probability as:

$$\begin{aligned} (\hat{i}^s, \hat{i}^e) &= \arg\max_{\hat{a}^s, \hat{a}^e} P_s(\hat{a}^s) \times P_e(\hat{a}^e) \\ \text{s.t.: } & 0 \leq \hat{i}^s \leq \hat{i}^e \leq N-1 \end{aligned} \quad (6)$$

where $\hat{i}^s$ and $\hat{i}^e$ are the best start and end boundaries of the predicted moment for the given video-query pair. And the predicted start/end time is computed by $\hat{t}^{s(e)} = \hat{i}^{s(e)}/(N-1) \times \mathcal{T}$, where $\mathcal{T}$ is the duration of the given video.

4. Experiment

4.1. Datasets

To evaluate the performance of our proposed, we conduct experiments on two challenging video moment retrieval datasets: **Charades-STA** [9] is composed of daily indoor activities videos, which is based on Charades dataset [35]. This dataset contains 6672 videos, 16,128 annotations, and 11,767 moments. The average length of each video is 30 seconds. $12,408$ and $3,720$ moment annotations are labeled for training and testing, respectively; **ActivityNet Caption** [3] is originally constructed for dense video captioning, which contains about 20k YouTube videos with an average length of 120 seconds.

4.2. Evaluation Metrics

Following existing video moment retrieval works, we evaluate the performance in two main metrics: **mIoU:** "mIoU" is the average predicted Intersection over Union over all testing samples; **Recall:** We adopt "R@n, IoU $= \mu$" as the evaluation metrics, following [9]. The "R@n, IoU $= \mu$" represents the percentage of language queries having at least one result whose IoU between top-n predictions with ground-truth is larger than μ. In our experiments, we reported the results of $n = 1$ and $\mu \in \{0.3, 0.5, 0.7\}$.

4.3. Implementation Details

For language query Q, we use the 300-D GloVe [28] vectors to initialize each lowercase word, and these word embeddings are fixed during training. For video V, we downsample frames and extracted RGB visual features using the 3D ConvNet which was pre-trained on the Kinetics dataset. We set the dimension of all the hidden layers in the model as 128, the kernel size of the convolutional layer as 7, and

Supervision	Method	Charades-STA				ActivityNet Captions			
		R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Full Supervision	CTRL [9]	-	23.63	8.89	-	-	-	-	-
	QSPN [42]	54.7	35.6	15.8	-	45.3	27.7	13.6	-
	2D-TAN [47]	-	39.7	23.31	-	59.45	44.51	26.54	-
	VSLNet [46]	73.84	60.86	41.34	53.92	61.65	45.50	28.37	45.11
	SeqPAN [45]	70.46	54.19	35.22	50.02	63.16	43.22	26.16	43.19
Weak Supervision	TGA [27]	32.14	19.94	8.84	-	-	-	-	-
	SCN [20]	42.96	23.58	9.97	-	47.23	29.22	-	-
	BAR [39]	44.97	27.04	12.23	-	49.03	30.73	-	-
	RTBPN [49]	60.04	32.36	13.24	-	49.77	29.63	-	-
	VLANet [26]	45.24	31.83	14.17	-	-	-	-	-
	MARN [36]	48.55	31.94	14.81	-	47.01	29.95	-	-
	LoGAN [37]	51.67	34.68	14.54	-	-	-	-	-
	CRM [12]	53.66	34.76	16.37	-	55.26	32.19	-	-
	CPL [50]	66.40	49.24	22.39	-	55.73	31.37	-	-
Single Frame	ViGA [5]	71.21	45.05	20.27	44.57	59.61	35.79	16.96	40.12
Active Learning	Random	44.17	14.65	3.58	30.57	50.11	23.47	11.91	35.07
	HUAL (Baseline)	66.91	45.48	22.5	43.76	51.58	31.5	16.12	36.78
	HUAL (50%, 2)	69.89	50.78	26.69	46.63	56.62	32.94	15.31	38.11
	HUAL (50%, 5)	**70.40**	**52.69**	**28.9**	**48.11**	**59.95**	**38.09**	**19.64**	**40.86**

Table 1. Performance comparison with the state-of-the-art methods under different supervision settings. With binary annotations, our HUAL can achieve comparable performance with some fully supervised method. HUAL ($K = 50\%, r = 2$) means selecting 50% videos in the training and annotated two frames for each video in 2 rounds. Best performance are noted in **bold**.

the head size of multi-head attention as 8. For all datasets, models are trained for 50 epochs. The batch size is set to 64. Dropout and early stopping strategies are adopted to prevent overfitting. The whole framework is trained by Adam optimizer with an initial learning rate 0.0002. For ActivityNet Captions, We set $\alpha = 4$, $\beta = 2$, and a initial value $\gamma = 2$ that gradually decreases with iteration. For Charades-STA, $\alpha = 4$, $\beta = 0.8$ were used, and γ was gradually decreased from 4 in the following experiments. All experiments are conducted on 1 Nvidia RTX A5000 GPU with 24GB memory. More details can be found in Supplementary Material. We will make our code open-source for reproducing all experiments.

4.4. Comparison with SOTA methods

Table 1 summarizes the experimental results on Charades-STA and ActivityNet Captions dataset. We mainly compare our HUAL with the following SOTA methods. **Fully-supervised method:** CTRL [9], QSPN [42], 2D-TAN [47], VSLNet [46], SeqPAN [45]; **Weakly-supervised method:** TGA [27], SCN [20], BAR [39], RTBPN [49], VLANet [26], MARN [36], LoGAN [37], CRM [12], CPL [50]; **Single Frame-supervised method:** ViGA [5].

From the results, we observe that our HUAL method can effectively improve the performance of baseline networks in all metrics and benchmarks. Here HUAL (baseline) means there is no supervision provided, the pseudo label is the output of weak-supervised model (CPL). Random means randomly selecting one frame from each video to annotated. HUAL ($K = 50\%$, $r = 2$) represents selecting 50% video sequences to annotate with 2 rounds. With more annotations, there exists steady performance gains in our HUAL model, and even comparable with fully-supervised methods.

For Charades-STA dataset, we can see that HUAL works well in even stricter metrics, such as R@0.7. Compared with ViGA [5], HUAL (50%, 2) achieves a significant 3.82% absolute improvement in R@0.7, which demonstrates the effectiveness of the proposed model. By comparing the performance of ViGA and Random, we can discover that positive samples have huge benefits to performance improvement, which is 14% in mIoU on Charades.

We further compare the results on ActivityNet Captions dataset. Note that the ground-truth video segments in ActivityNet Captions have a longer averaged duration with various lengths. So it needs more rounds to annotate frames with distinction. Since the ground truth may cover more than 80% duration of the video, or smaller than 5% length of the video, annotating with several rounds will meet the condition that all samples are positive or negative. Then, our HUAL model needs more rounds to annotate than in the Charades.

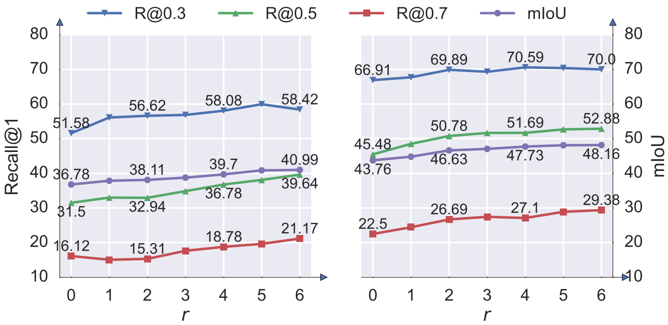
4.5. Quantitative Analysis

We further perform qualitative analysis of our method so as to enable a better understanding of its strength. The qualitative results of HUAL on Charades-STA dataset are demonstrated in Figure 6. With more rounds of annotation, our HUAL method continues refining the prediction of video and language query pairs.

4.6. Ablation Studies

We mainly conduct the ablation studies and make comparisons with other SOTA methods on the Charades dataset.

Performance in Different Rounds. We first compare the performance of HUAL in different rounds. As shown in Fig. 5, in each round with one more frame annotated, the performance gain is about 0.9% on average in the metric of mIoU. With more annotated frames as supervision, the performance of HUAL is increasing steadily, which proves that our HUAL model can effectively utilize the new-annotated labels. To make a trade-off between accuracy and annotation cost, we set $r = 5$ with the best performance. On the Charades dataset, HUAL (50%, 2) can suppress ViGA with the same annotation cost but in looser restriction. With more rounds of annotations, the performance of HUAL can keep increasing. However, as shown in Table 1, HUAL needs more rounds of annotation to reach a comparable performance with ViGA on the ActivityNet than Charades dataset. This is because of the unbalanced distribution of ground truth with various lengths in AcitivityNet dataset. It will take more rounds to sample frames with both positive and negative frames, and the positive frame is more beneficial to improve the performance of HUAL.



(a) Charades-STA (b) ActivityNet Captions

Figure 5. Performance comparison (%) of HUAL with different rounds on Charades and ActivityNet Captions datasets. With more rounds of annotation provides, our HUAL can achieve steady performance gain in all metrics.

Table 2. Performance comparison (%) of HUAL with different components on Charades dataset. Each components can improve the performance.

Probability			R@0.3	R@0.5	R@0.7	mIoU
$P_{distance}$	P_{model}	L_{r-1}^{frame}				
✓			56.56	33.58	13.25	36.82
✓	✓		66.8	46.64	21.37	43.5
✓	✓	✓	70.40	52.69	28.90	48.11

Different Components in Frame-level Uncertainty. As shown in Table 2, we analyse the effectiveness of different components in frame-level uncertainty introduced in Eq. 2. If we only consider the probability of distance, HUAL can only reach the performance of 36.82% in mIoU. With the probability score of the model considered, the performance of HUAL can be improved to 43.5% in mIoU. By further considering the labels in the last round, HUAL can achieve the best performance of 48.11% in mIoU. Hence, each component is proved effective in the performance.

Selection of Sequence-level Uncertainty. Table 3 shows the ablation studies of different settings in the sequence-level uncertainty on Charades dataset. We first conduct experiments on the setting of different proportions K of selected queries. When $K = 50\%$, the HUAL can achieve 48.11% in mIoU.

With more videos annotated, the performance of HUAL will be improved with no doubt, but with a higher label cost. To achieve a balance between accuracy and label cost, we select $K = 50\%$ as the best choice.

Quality of Generated Pseudo Label. We also analysis the quality of generated pseudo label in each round. As can be shown in Fig. 7, with more rounds of annotations, the pseudo label is more similar to ground truth, and the predicated result is more accurate, which verifies the effectiveness of our HUAL in sample selection based on frame-level and sequence-level uncertainty and pseudo label generation with rules.

Annotation Cost. To relieve the annotation burden, we stimulate the process of labeling with binary annotations. Compared with annotating with precise start and end timestamps, our HUAL only needs the binary label of the selected frame, which needs less time than annotation in fully super-

Table 3. Performance comparison (%) of HUAL with different selection K on Charades dataset.

Components (K)	R@0.3	R@0.5	R@0.7	mIoU
HUAL (10%, 5)	67.77	49.11	24.95	45.46
HUAL (30%, 5)	68.44	50.38	26.51	46.53
HUAL (50%, 5)	70.40	52.69	28.90	48.11
HUAL (70%, 5)	71.16	53.09	28.82	48.84
HUAL (100%, 5)	70.91	56.13	32.69	49.70

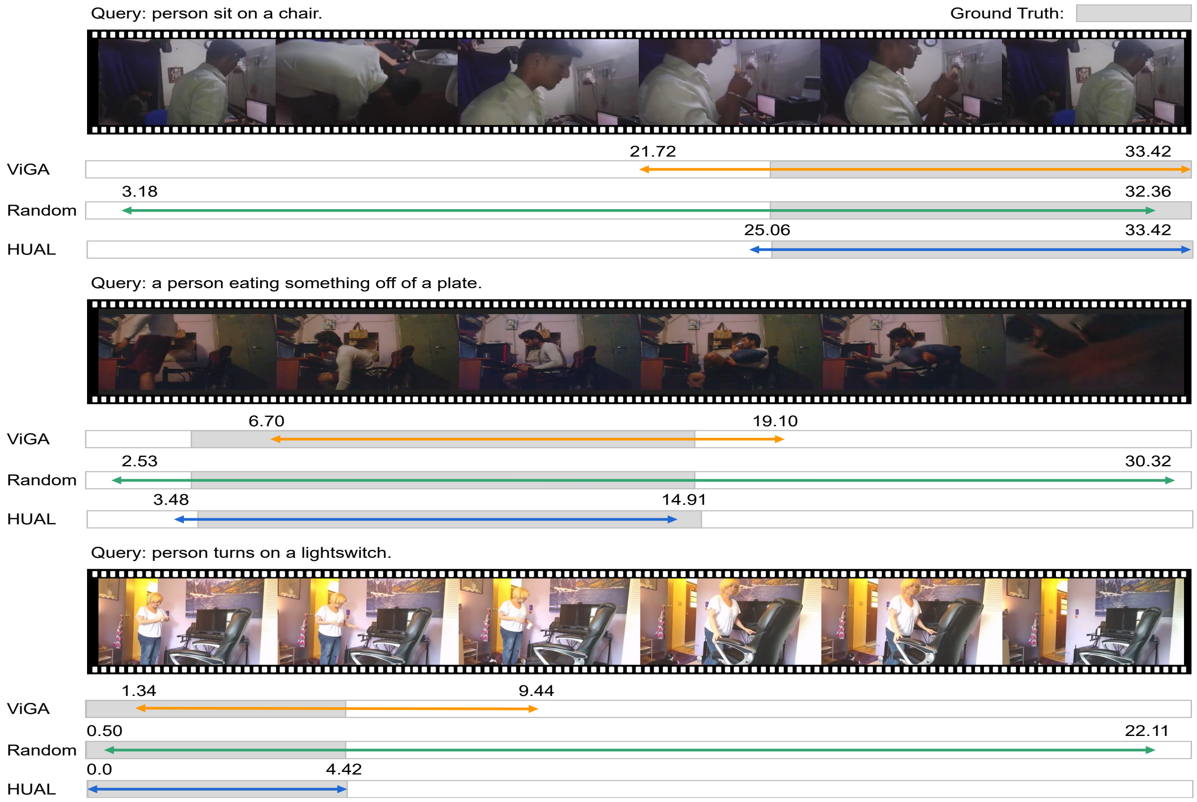
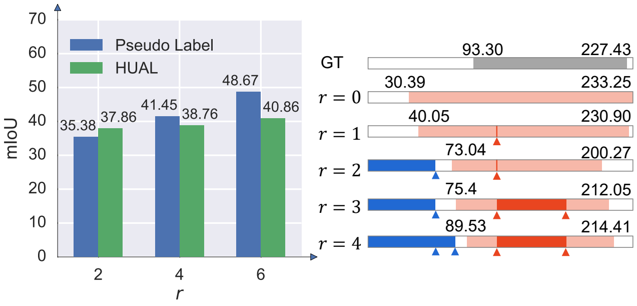


Figure 6. Qualitative results of HUAL on Charades-STA dataset. Compared with other methods, our HUAL method can locate the temporal region more accurately via several binary annotations.



(a)

(b)

Figure 7. (a) Comparison between the pseudo label and HUAL results. (b) Visualization of the pseudo label in different rounds. With more rounds of annotation, the pseudo label is more similar to the ground truth in fully supervised setting.

vised method or ViGA. For the simple case like Charades, annotating 0.5 frames on average will achieve comparable performance as ViGA, as shown in Table 3 of (HUAL (10%, 5)). For complex cases like ActivityNet, our method needs more rounds of annotation, which is still a relatively small expense. And positive sample can bring more benefits to the performance improvement compared with negative one.

5. Conclusion

In this paper, we propose a new interactive manner to stimulate the process of annotation in video moment retrieval task, which can be treated as human-in-the-loop. To be specific, we model the uncertainty of each frame among the whole video sequence corresponding to query description, and select the frame with the highest uncertainty to be annotated by expert. After obtaining the label provided by the expert, the model can be trained with sparse but correct labels. Moreover, we treat the uncertainty score of frames in a video as a whole, and estimate the uncertainty of each video, which can further relieve the burden of video annotations. In general, our active learning strategy for video moment retrieval works not only at the frame level but also at the sequence level. Experiments on two public datasets validate the effectiveness of our proposed method. In the future, we consider further exploring the positive and negative frames with contrastive learning for better feature learning and lower label costs, and extend this pipeline to other relevant tasks.

6. Acknowledgement

This research is supported by the Sea-NExT Joint Lab.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 2

[2] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *CVPR*, pages 9368–9377, 2018. 3

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 5

[4] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *CVPR*, pages 8425–8435, 2021. 1

[5] Ran Cui, Tianwen Qian, Pai Peng, Elena Daskalaki, Jingjing Chen, Xiaowei Guo, Huyang Sun, and Yu-Gang Jiang. Video moment retrieval from text queries via single frame annotation. *SIGIR*, 2022. 3, 6

[6] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *CVPR*, pages 234–243, 2021. 1

[7] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *ECCV*, pages 562–577. Springer, 2014. 3

[8] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, pages 1183–1192. PMLR, 2017. 3

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 1, 2, 5, 6

[10] Yuhong Guo. Active instance sampling via matrix partition. *NIPS*, 23, 2010. 3

[11] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. *ACL*, 2018. 2

[12] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, pages 7199–7208, 2021. 6

[13] Wei Ji, Long Chen, Yinwei Wei, Yiming Wu, and Tat-Seng Chua. Mrtnet: Multi-resolution temporal network for video sentence grounding. *arXiv preprint arXiv:2212.13163*, 2022. 2

[14] Wei Ji, Yicong Li, Meng Wei, Xindi Shang, Junbin Xiao, Tongwei Ren, and Tat-Seng Chua. Vidvrd 2021: The third grand challenge on video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4779–4783, 2021. 1

[15] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, pages 2372–2379. IEEE, 2009. 3

[16] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Wei Ji, Zhou Zhao, Shengyu Zhang, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[17] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, et al. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8707–8717, 2022. 2

[18] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3801–3810, 2022. 2

[19] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, pages 2928–2937, 2022. 1

[20] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, volume 34, pages 11539–11546, 2020. 3, 6

[21] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. *EMNLP*, 2021. 2

[22] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, pages 11235–11244, 2021. 2

[23] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM Multimedia*, pages 4070–4078, 2020. 2

[24] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, pages 15–24, 2018. 2

[25] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP*, pages 5147–5156, 2019. 2

[26] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *ECCV*, pages 156–171, 2020. 6

[27] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 2, 6

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 5

[29] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Video dialog as conversation about objects living in space-time. In *ECCV*, pages 710–726. Springer, 2022. 1

[30] Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active

learning for image classification. In *ICIP*, pages 3934–3938. IEEE, 2017. 3

[31] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 3

[32] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *ICLR*, 2018. 3

[33] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *NeurIPS*, 20, 2007. 3

[34] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM international conference on Multimedia*, pages 3654–3663, 2021. 1

[35] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 1, 5

[36] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020. 3, 6

[37] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *WACV*, pages 2083–2092, 2021. 2, 6

[38] Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, and Yi Yang. Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE Transactions on Multimedia*, 2022. 2

[39] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *ACM Multimedia*, pages 1283–1291, 2020. 6

[40] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. AAAI, 2022. 1

[41] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *AAAI*, volume 35, pages 2986–2994, 2021. 2

[42] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, volume 33, pages 9062–9069, 2019. 6

[43] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. *SIGIR*, 2021. 2

[44] 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

[45] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Parallel attention network with sequence matching for video grounding. *ACL Findings*, 2021. 2, 4, 5, 6

[46] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, pages 6543–6554, 2020. 6

[47] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, pages 12870–12877, 2020. 2, 6

[48] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, pages 655–664, 2019. 2

[49] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM Multimedia*, pages 4098–4106, 2020. 6

[50] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, pages 15555–15564, 2022. 2, 3, 4, 6

[51] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021. 3, 5

[52] Zhedong Zheng and Yi Yang. Adaptive boosting for domain adaptation: Toward robust predictions in scene segmentation. *IEEE Transactions on Image Processing*, 31:5371–5382, 2022. 3

[53] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *EMNLP*, 2022. 1