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

Wei Ji* Renjie Liang* Zhedong Zheng1* Wenqiao Zhang~ Shengyu Zhang~ Juncheng Li~ Mengze Li~ Tat-seng Chua*

*National University of Singapore ~Zhejiang University

Poster ID: THU-PM-230

Overview

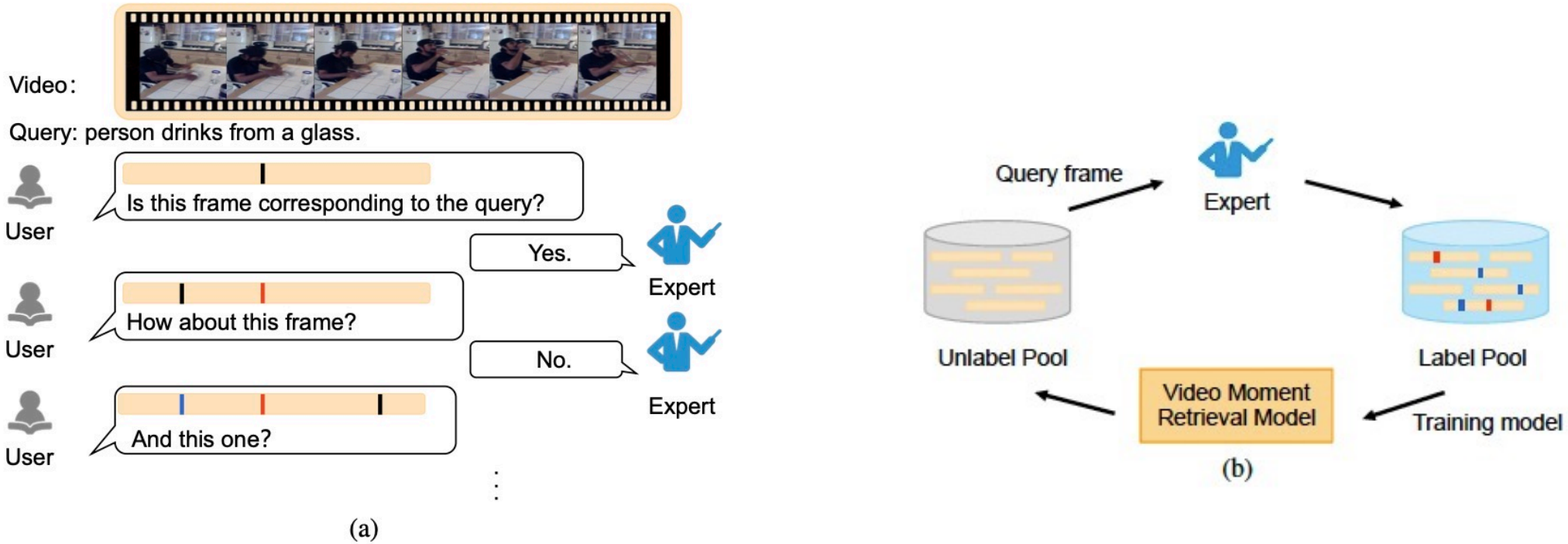


Fig. 1. We propose a new interactive method named HUAL which only requires binary annotations to reduce the annotation cost.

Video Moment Retrieval (VMR)

- **VMR**: Given a query sentence and the video, temporally retrieving a specific moment from the video [1]

.

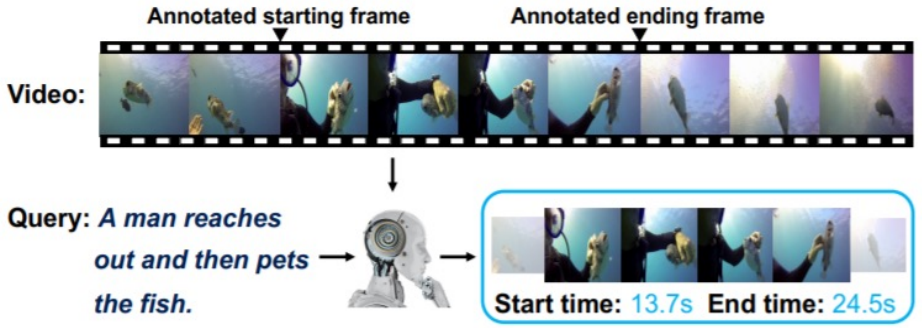


Figure 1. VMR [1]

- **Applications**:
  - **Video QA**: Return the localized moment as the visual evidence to support the answer [3]
  - **City Video Surveillance**: Temporally finding a crime suspect in video surveillance system with a textual query.
  - ….

"the system also comes with an auto function it automatically controls the temperature air distribution and air flow to reach and maintain a comfort level based on the temperature you selected"

[1] Runhao Zeng et al. Dense Regression Network for Video Grounding. CVPR 2020
[2] Luo, Hongyin, et al. "Integrating Video Retrieval and Moment Detection in a Unified Corpus for Video Question Answering." INTERSPEECH. 2019.
[3] Junyeong Kim et al. Modality Shifting Attention Network for Multi-Modal Video Question Answering. CVPR 2020

Video Moment Retrieval (VMR)

- **Motivation:**
  - Precise labels on VMR datasets are time-consuming and cost-expensive;
  - Relying on the well- annotated dataset will restrict the generalization ability of the current models.
- **Assumption:**
  - not each frame should be considered equally, as the frame with the higher uncertainty is more valuable than the rest;
  - not each video can be treated as a hard sample, annotating complex video and query pairs first benefits more than annotating simple ones.

Proposed Method

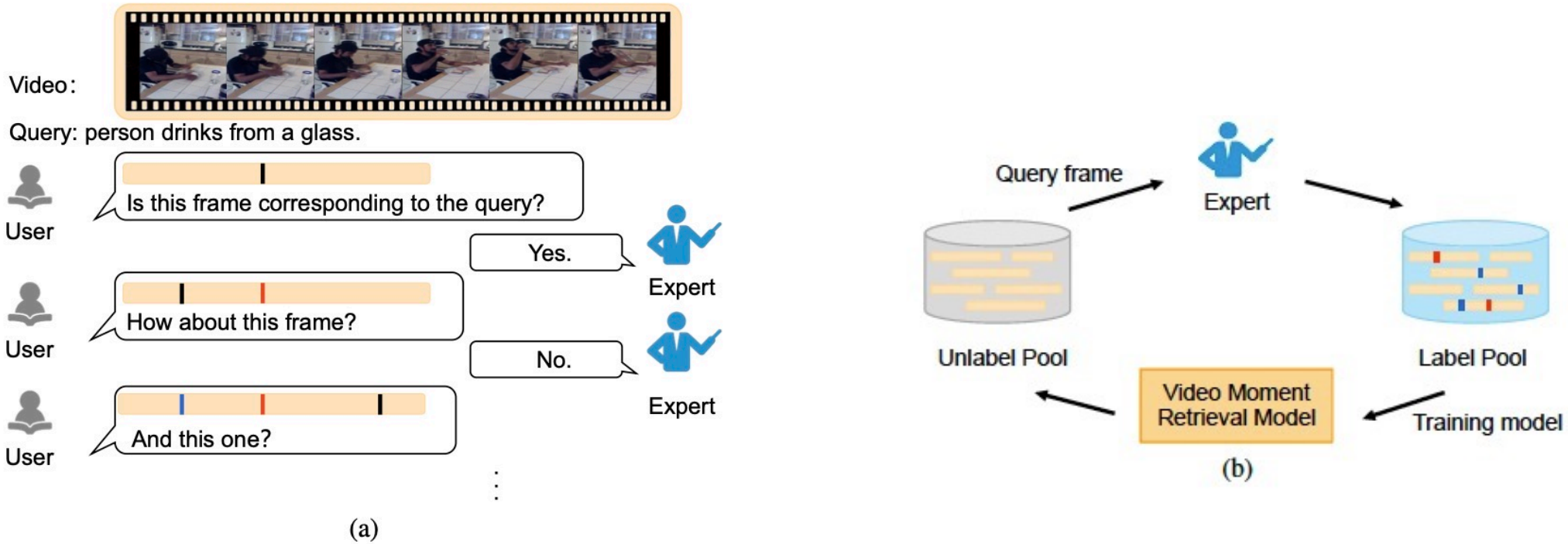


Fig. 1. We propose a new interactive method named HUAL which only requires binary annotations to reduce the annotation cost.

Proposed Method

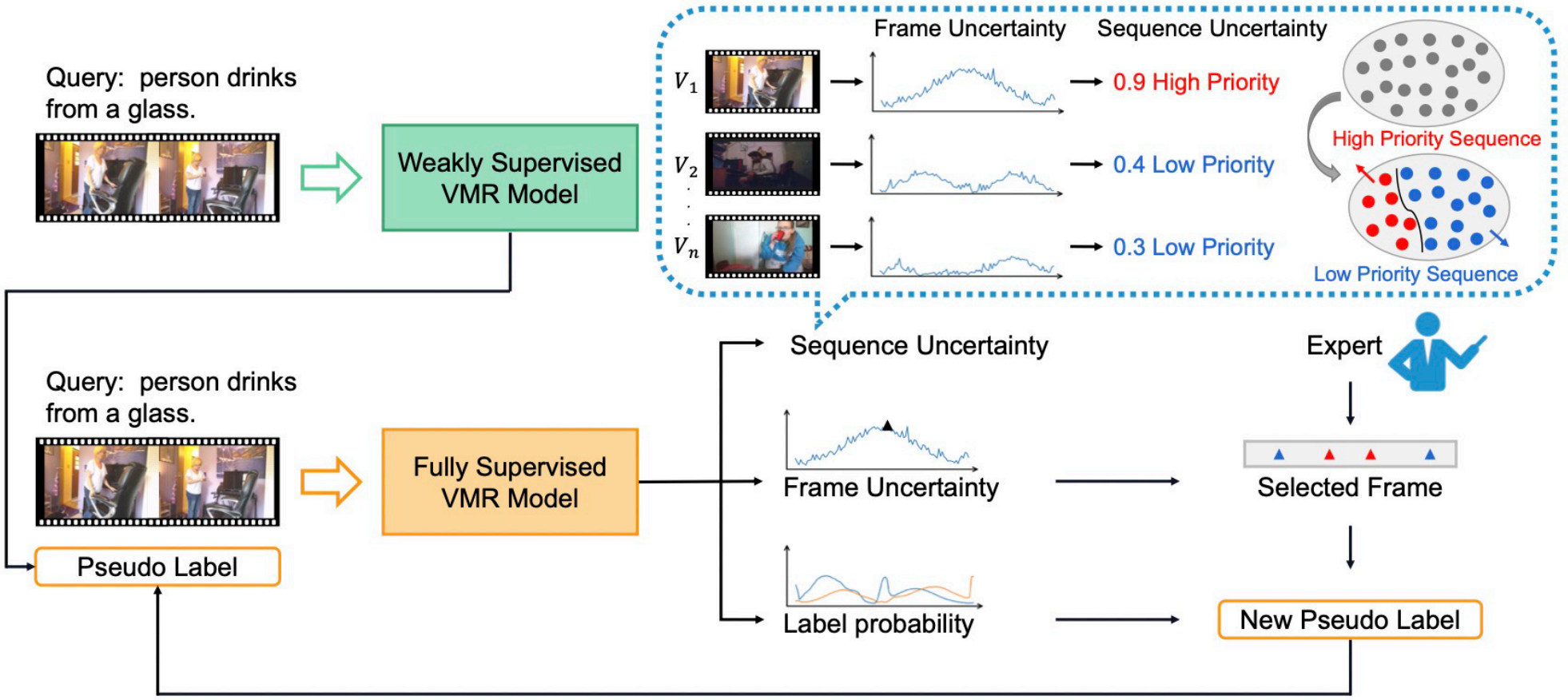


Fig. 2. The whole pipeline of our HUAL method.

Proposed Method

- **Frame-level Uncertainty:**

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

- **Pseudo Label Generation :**

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

- **Sequence-level Uncertainty:**

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

Proposed Method

- **Training:**

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

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:

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

where σ is the variance of the prediction $P_{model}(v)$, and σ ≥ 0.

- **Inference:**

$$(\hat{i}^s, \hat{i}^e) = \arg\max_{\hat{a}^s, \hat{a}^e} P_s(\hat{a}^s) \times P_e(\hat{a}^e) \tag{6}$$

$$\text{s.t.: } 0 \leq \hat{i}^s \leq \hat{i}^e \leq N - 1$$

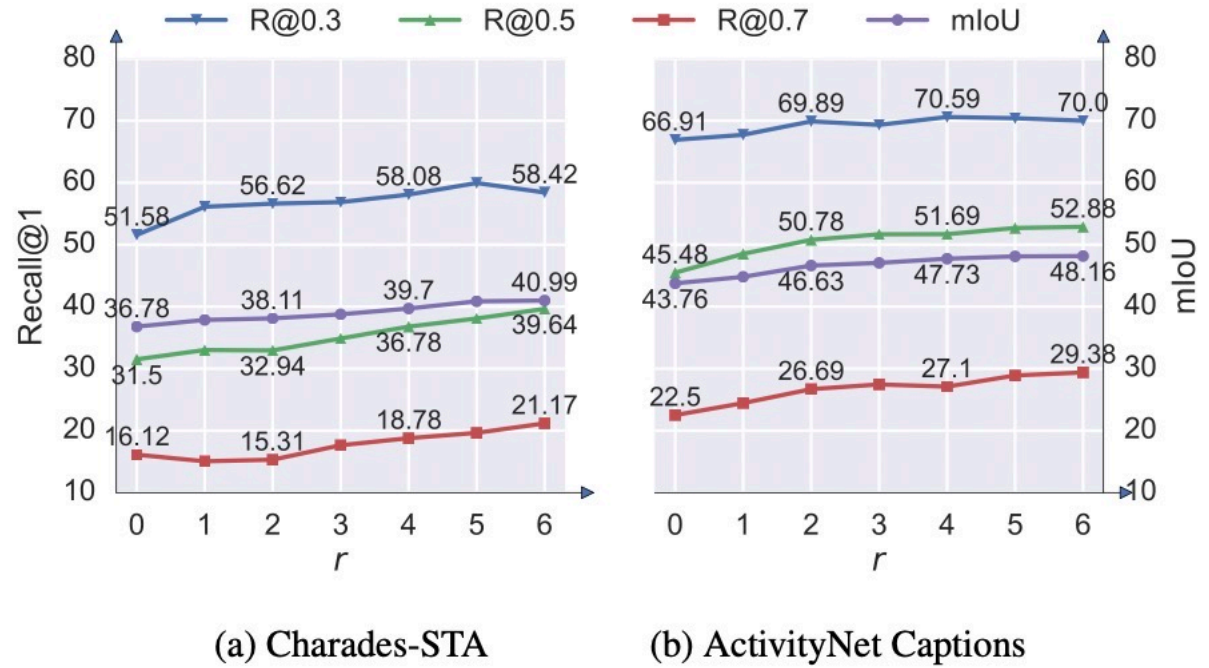
Experiments

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**

Tab. 1. Performance comparison with the state-of-the-art methods under different supervision settings.

Experiments

- **Performance in Different Rounds:**



(a) Charades-STA (b) ActivityNet Captions

Fig. 3. 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.

Experiments

- **Different Components in Frame-level Uncertainty :**

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

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

- **Selection of Sequence-level Uncertainty:**

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

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

Contribution

- 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.

- 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.

Thank You!
Q & A