JUNE 18-22, 2023
CVPR VANCOUVER, CANADA

UCF | Center for Research in Computer Vision
UNIVERSITY OF CENTRAL FLORIDA

Hybrid Active Learning via Deep Clustering for Video Action Detection
THU-AM-228

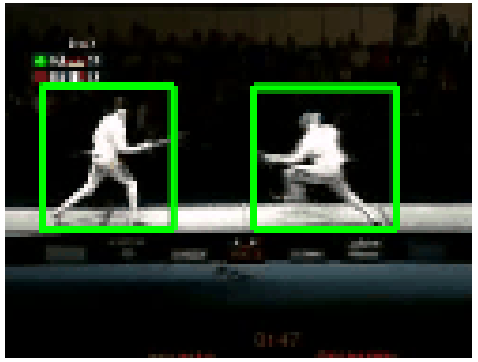
Aayush J Rana, Yogesh S Rawat

Center for Research in Computer Vision (CRCV)

University of Central Florida (UCF)

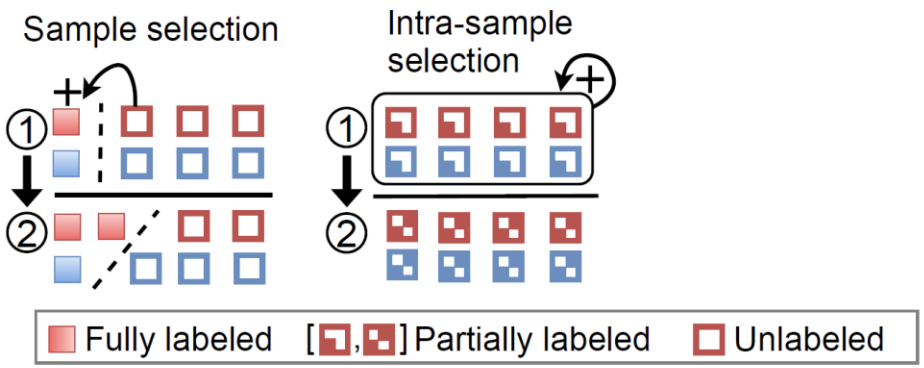
Challenges

- Training requires dense annotation
  - Dense data ∝ large annotation cost
- Unnecessary cost
  - Repetitive nearby frames
  - Unrelated frames annotated
- Comparison across videos
  - Varying length
  - Varying actors
  - Class-wise difficulty
  - No difficulty metric



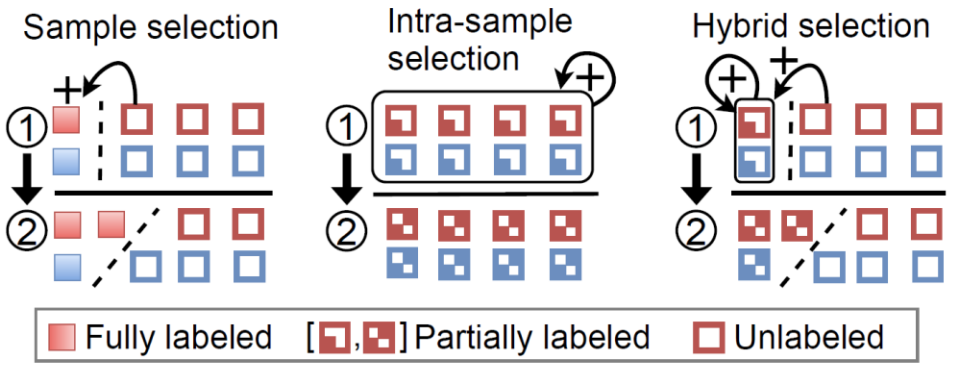
Previous work

- Annotation selection at frame level
- Assumes all videos annotated
  - Partial annotations
  - No metric to compare between videos



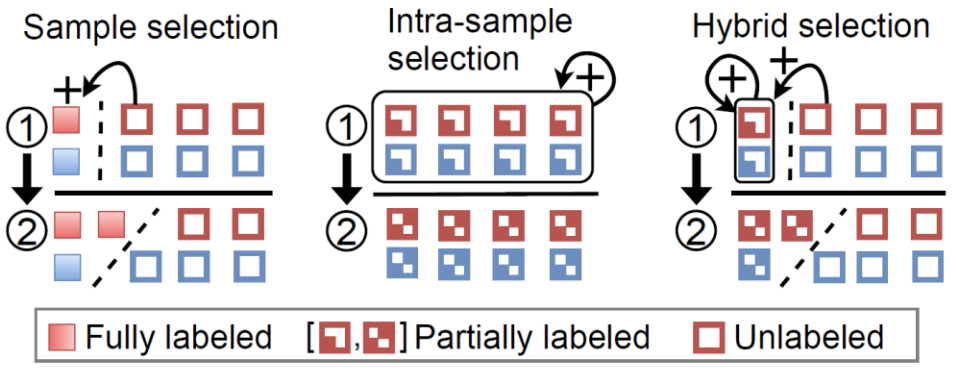
Motivation

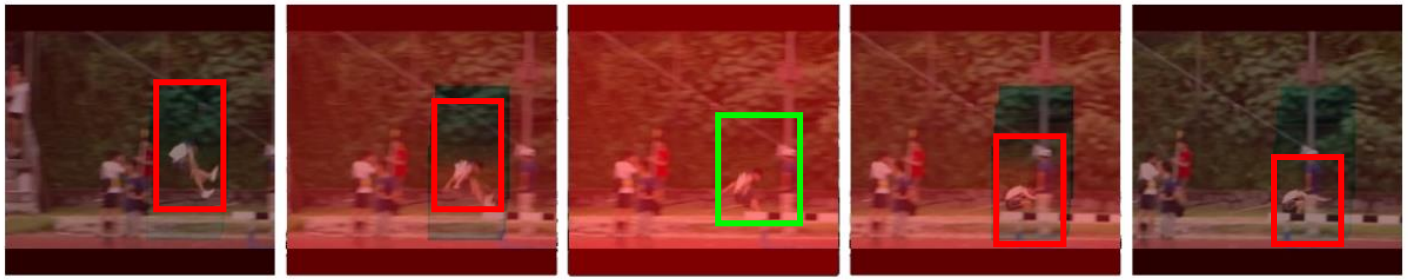
- Reduce annotation cost
  - Video level selection
  - Frame level selection
  - Remove redundant videos
- Enable video comparison using
  - Informativeness
  - Diversity
- Improve sparse training
  - Improve pseudo-label usage



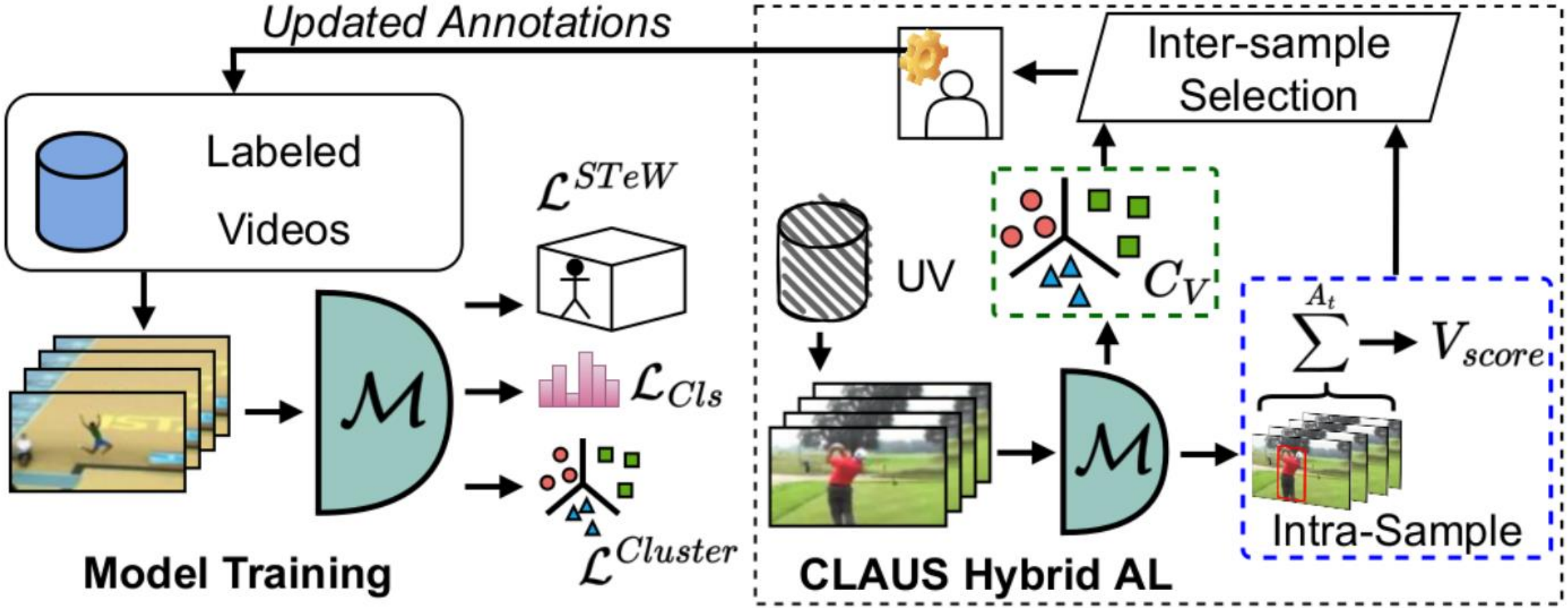
Contributions

- Hybrid selection (*CLAUS*)
  - AL based strategy
  - Video + frame selection
  - Uncertainty based video ranking
  - Clustering based video selection
- Improved pseudo-label loss (*STeW*)
  - Pixel-level weight
  - BG/FG consistency





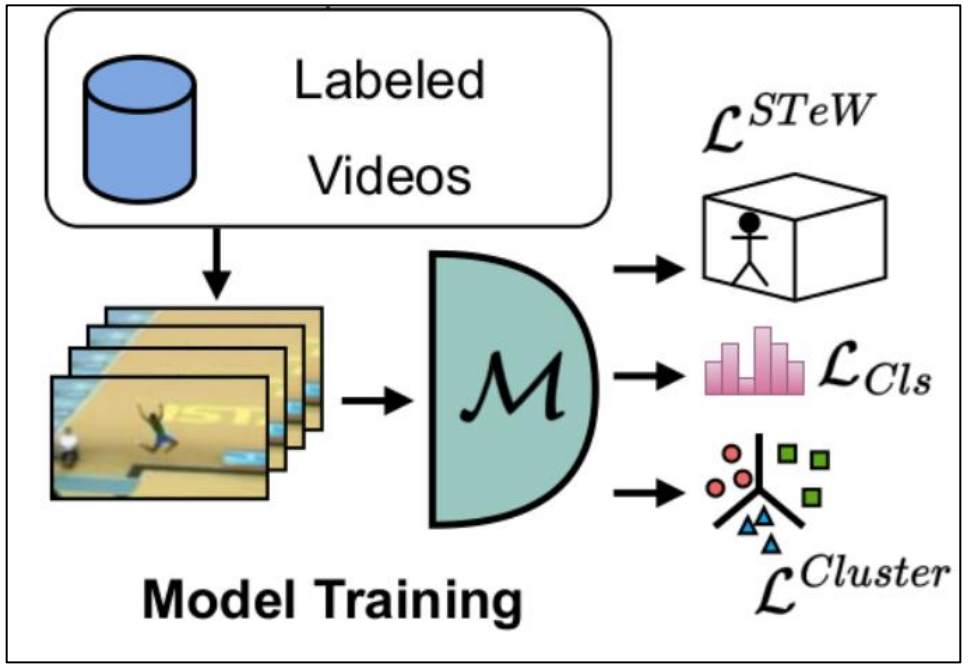
Proposed approach



Model Training Objectives

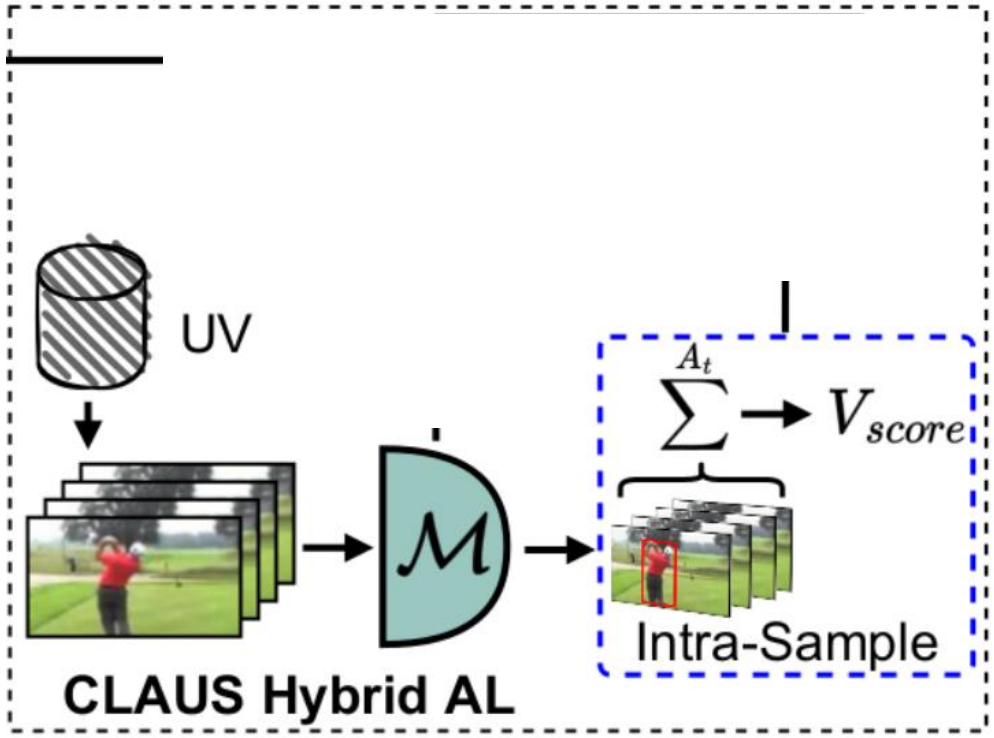
- Classification loss
- Localization loss
  - Spatio-Temporally Weighted loss (STeW)
  - Uses pixel-level consistency as weight
- Cluster loss
  - K arbitrary clusters
  - Adjust centers using video features

$$\min_{\theta} \mathcal{L} = \mathcal{L}^{Cluster} + \mathcal{L}_l^{STeW} + \mathcal{L}^{Cls}$$



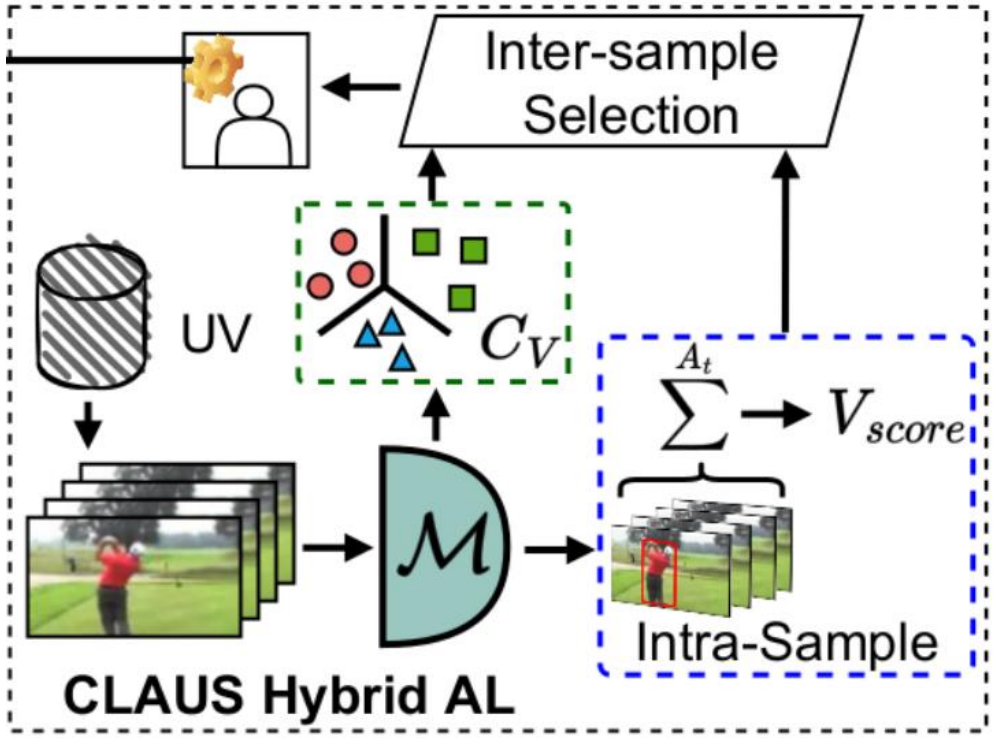
Intra-sample selection

- Frame-level selection
- Uncertainty based score
- Distance based redundancy reduction
- Top *t* frames used for video score



Inter-sample selection

- Video-level selection
- Video score from intra-sample
- Cluster assignment per video
- Top *V* videos per cluster selected
  - Frames from intra-sample



Datasets

- UCF-101
  - 3207 videos
  - **24** action classes
  - Spatio-temporal bounding box annotation
- J-HMDB
  - 928 videos
  - **21** action classes
  - Spatio-temporal pixel-wise annotation




Comparing with baselines

Method	$\mathcal{A}\%$	UCF-101-24 v-mAP	UCF-101-24 f-mAP	J-HMDB-21 v-mAP	J-HMDB-21 f-mAP
Random	1%	52.6	54.1	36.6	42.1
Equi.	1%	53.3	55	38.1	43.5
Entropy [1] †	1%	52.2	53.5	40.7	49.0
Uncertainty [14] †	1%	44.0	46.7	46.0	47.9
Our	1%	61.8	61.6	58.6	61.9
Random	5%	67.5	67.3	69.3	70.1
Equi.	5%	67.2	67.0	70.0	70.4
Entropy [1] †	5%	71.3	70.2	70.7	70.8
Uncertainty [14] †	5%	69.7	68.2	69.0	69.3
Our	5%	72.2	72.1	71.3	72.7

Comparing with prior weakly-supervised

Method	$\mathcal{A}$%	f-mAP@	v-mAP@			
		0.5	0.1	0.2	0.3	0.5
Mettes et al. [40]	V	-	-	37.4	-	-
Escorcia et al. [12]	V	-	-	45.5	-	-
Zhang et al. [67]	V	30.4	62.1	45.5	-	17.3
Arnab et al. [3]	V	-	-	61.7	-	35.0
Mettes et al. [39]	P	-	-	41.8	-	-
Cheron et al. [9]	P	-	-	70.6	-	38.6
Weinz. et al. [64]	1.1%	-	-	57.1	-	46.3
Weinz. et al. [64]	2.8%	63.8	-	57.3	-	46.9
MixMatch [5]	S-20%	20.2	-	60.2	-	13.8
Pseudo-label [32]	S-20%	64.9	-	93.0	-	65.6
Co-SSD(CC) [24]	S-20%	65.3	-	93.7	-	67.5
Kumar et al. [31]	S-20%	69.9	-	95.7	-	72.1
Ours	1%	61.6	98.1	95.9	88.9	61.8
Ours	5%	72.1	98.1	96.1	91.2	72.2

UCF-101-24

Method	$\mathcal{A}$%	f-mAP@	v-mAP@			
		0.5	0.1	0.2	0.3	0.5
Zhang et al. [67]	V	65.9	81.5	77.3	-	50.8
Weinz. et al. [64]	6%	50.7	-	-	-	58.5
Weinz. et al. [64]	15%	56.5	-	-	-	64.0
MixMatch [5]	S-30%	7.5	-	46.2	-	5.8
Pseudo-label [32]	S-30%	57.4	-	90.1	-	57.4
Co-SSD(CC) [24]	S-30%	60.7	-	94.3	-	58.5
Kumar et al. [31]	S-30%	64.4	-	95.4	-	63.5
Ours	1%	61.9	99.0	96.8	91.5	58.6
Ours	5%	72.7	99.1	97.3	94.8	71.3

J-HMDB-21

Action Detection Results

Soccer Juggling

Salsa Dancing

Floor Gymnastics

Horse Riding

Long Jumping

Red: GT
Blue: Our detection

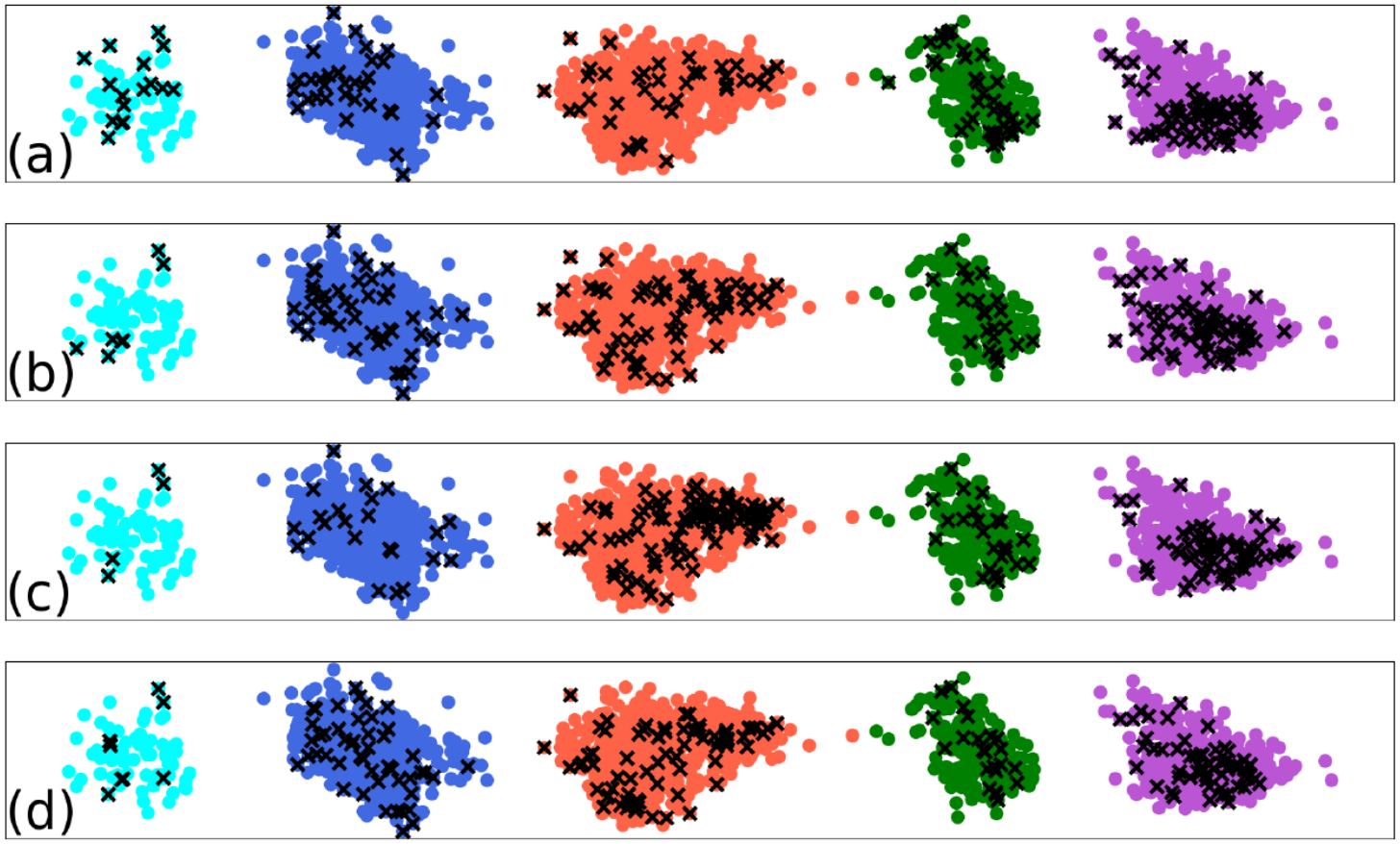
Cluster representation

a: CLAUS

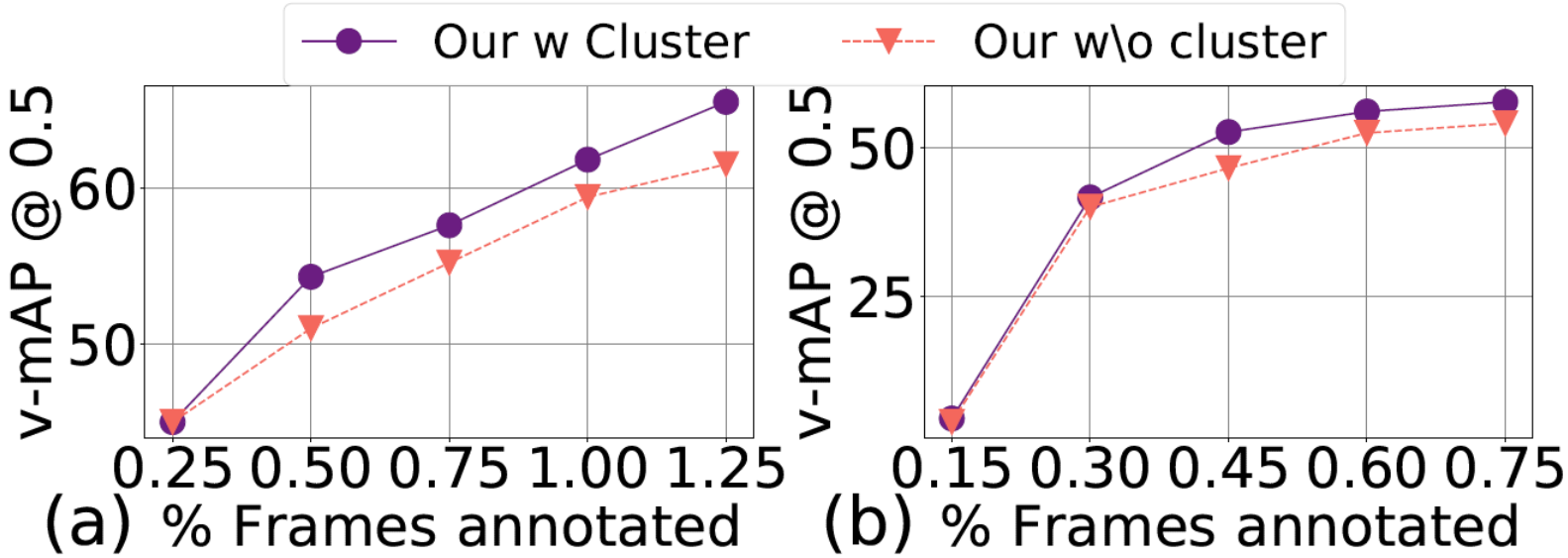
b: Entropy

c: Uncertainty

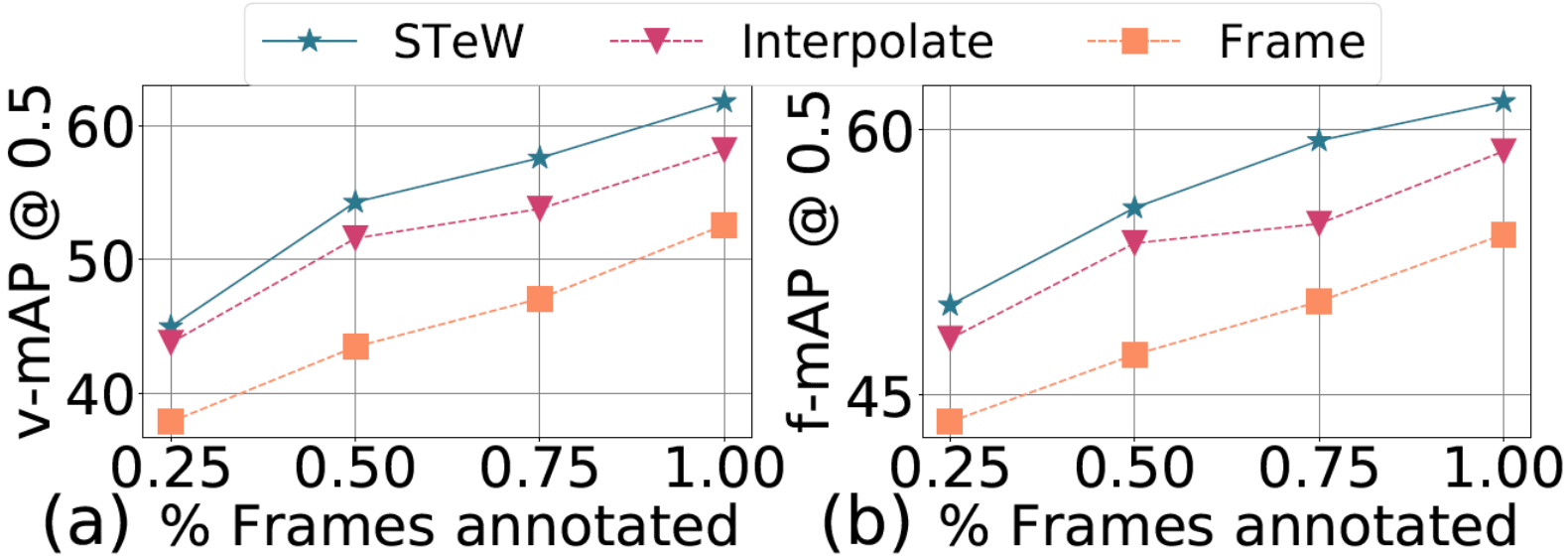
d: Random



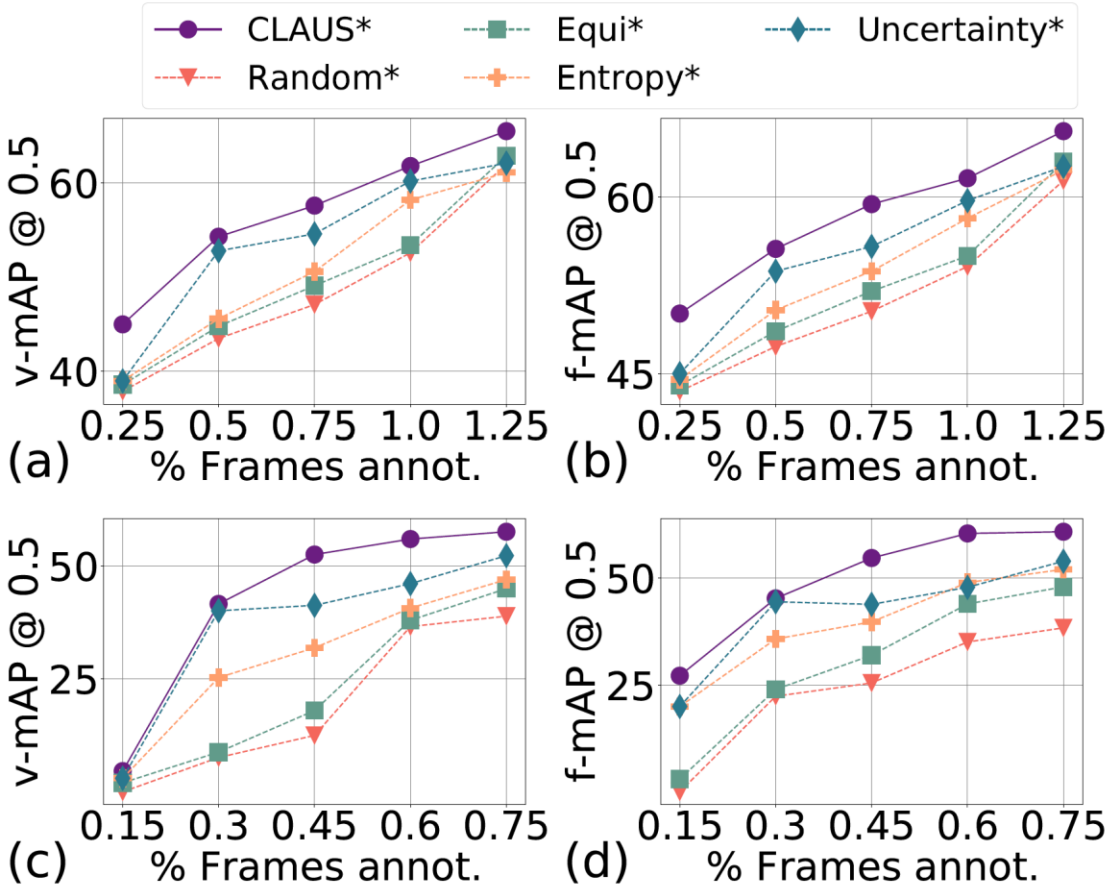
Cluster effectiveness



Loss effectiveness



Selection method analysis



Summary

- Hybrid selection improves performance
  - Clustering-aware selection strategy
  - Reduces similar video
  - Enables inter-sample comparison
- *STeW* loss improves sparse label training

Thank You

Project Link: https://tinyurl.com/hybridclaus