ACM Multimedia 2024

Melbourne, Australia

Not All Pairs are Equal: Hierarchical Learning for Average-Precision-Oriented Video Retrieval

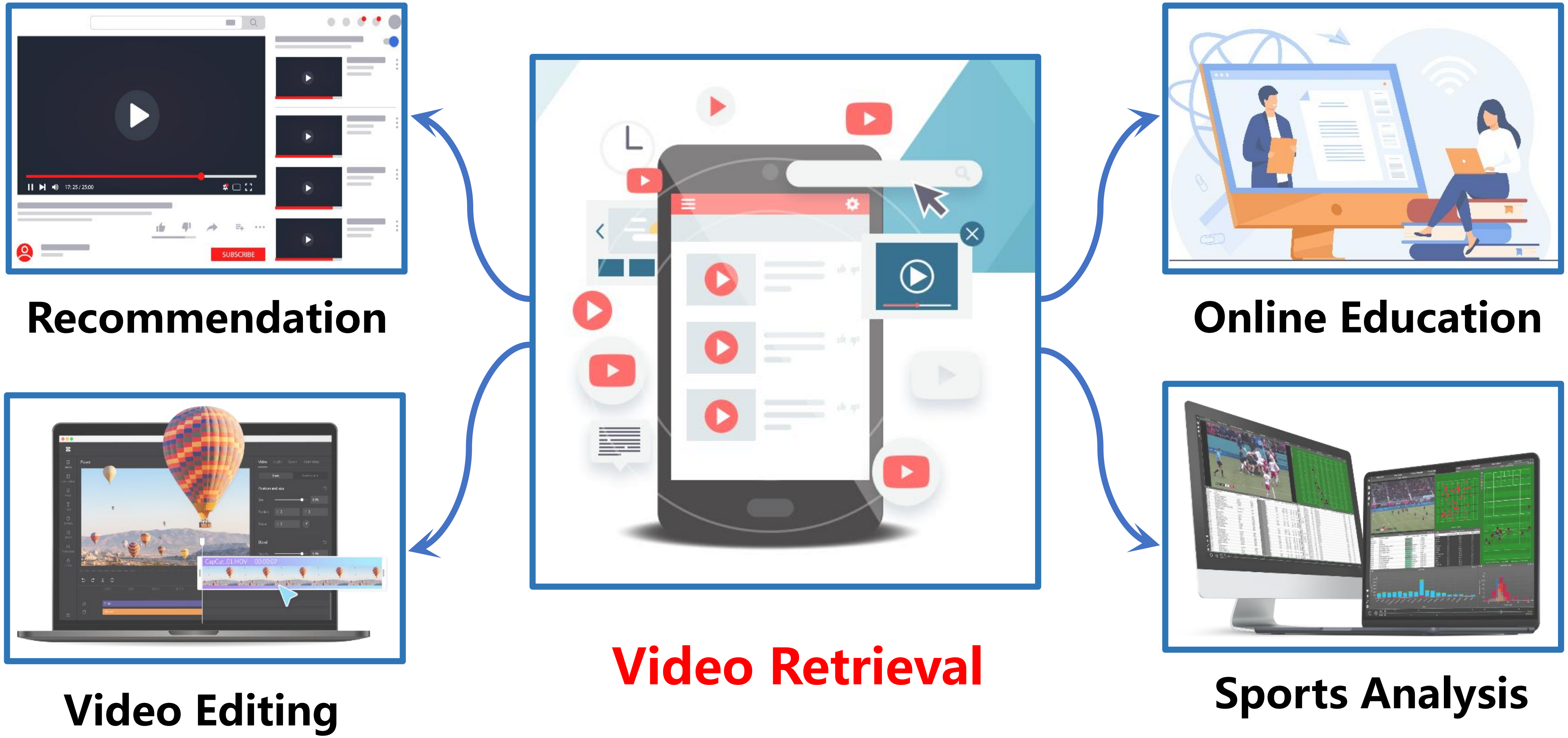
Yang Liu, Qianqian Xu, Peisong Wen,
Siran Dai, Qingming Huang



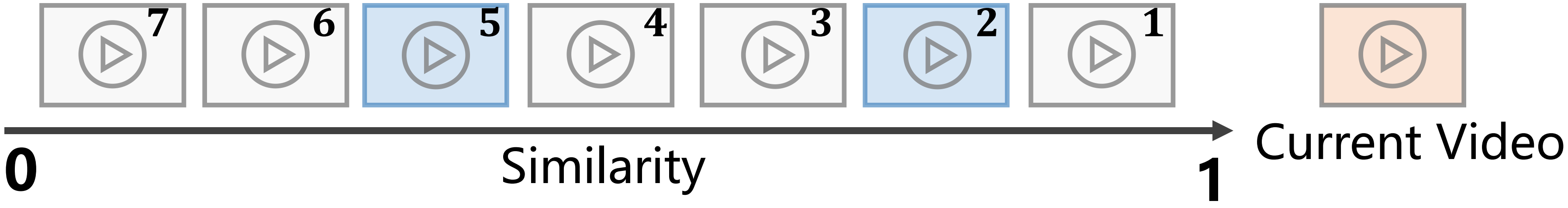


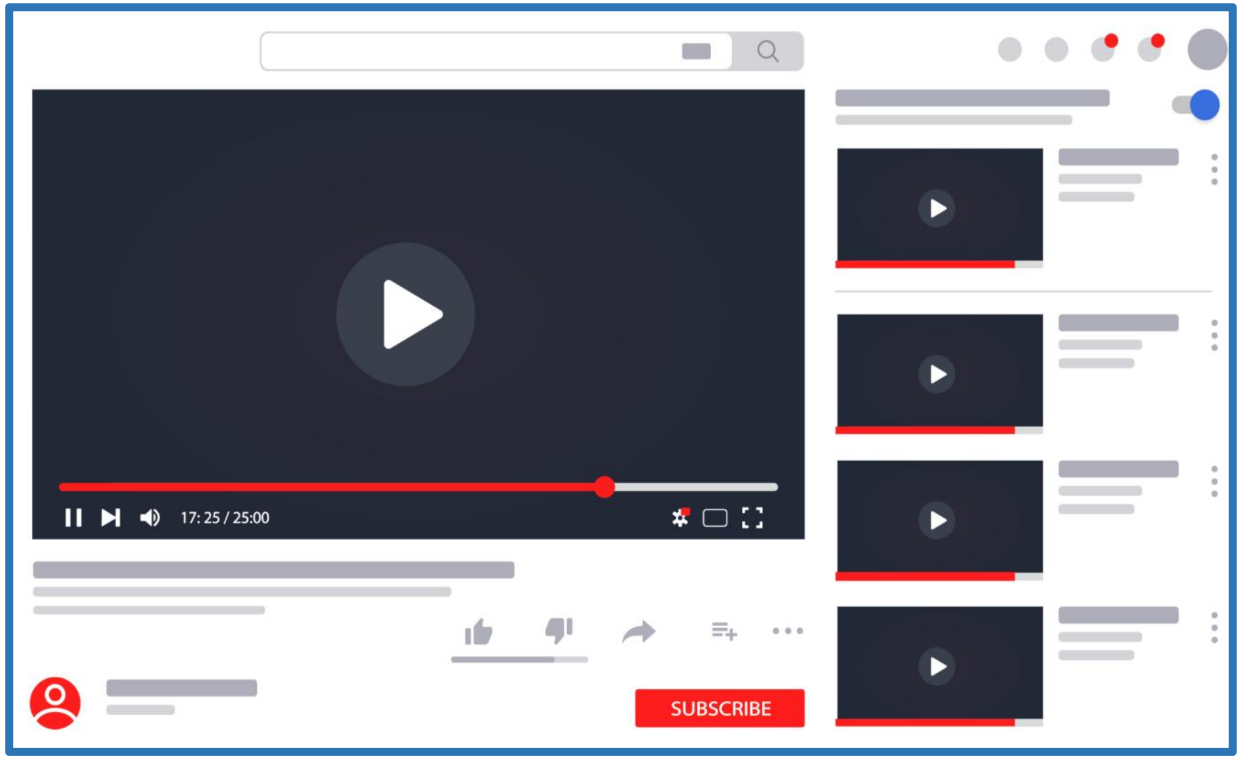
Background

- ☐ **Video retrieval is a fundamental task for multimedia applications**



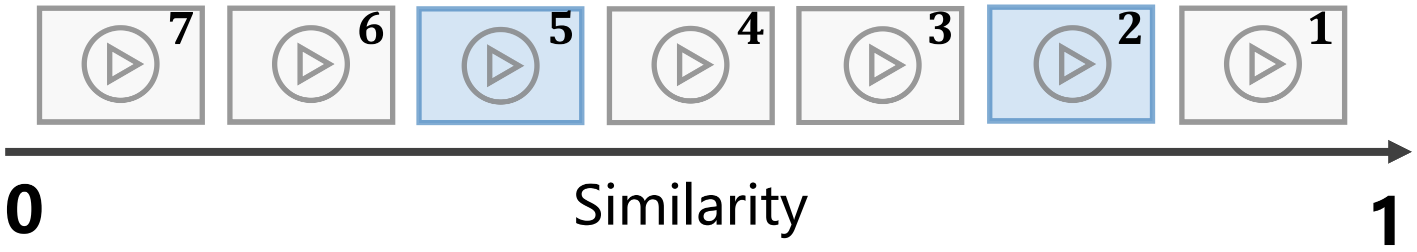
Background





Background

- **Evaluation metric: Average Precision (AP)**
  - AP assigns greater weights on higher-ranked videos
  - $$AP = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{r_i}$$

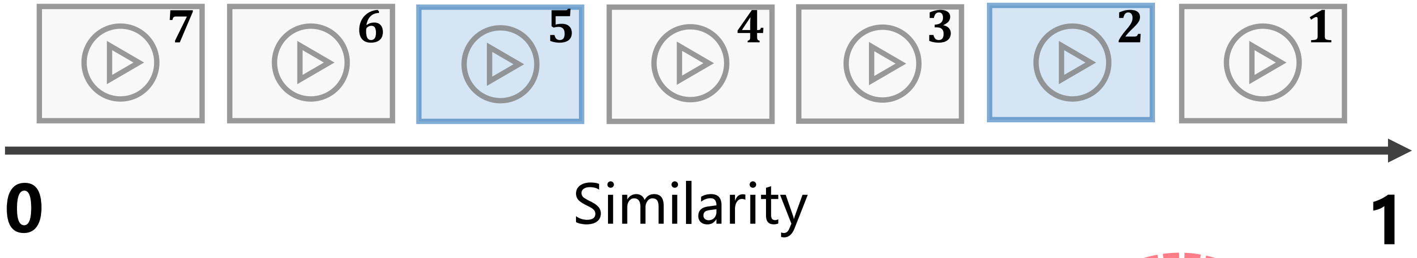


$$AP = \frac{1}{2}(\frac{1}{2} + \frac{2}{5}) = 0.45$$

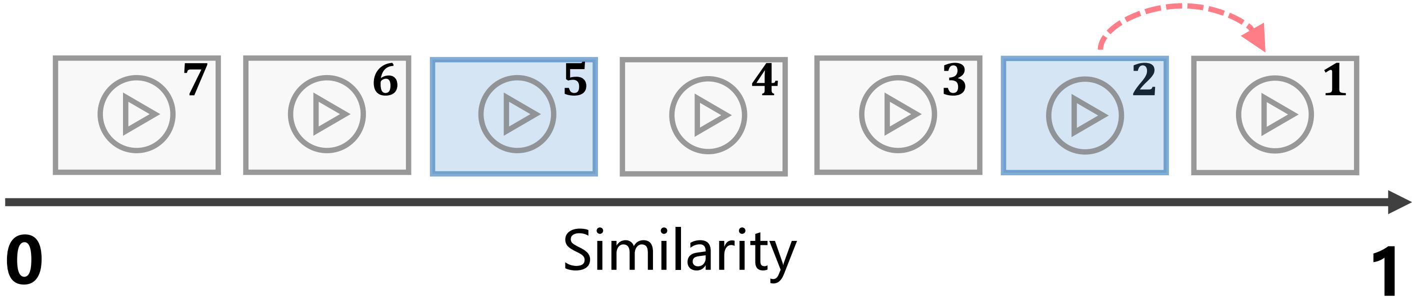
Background

□ Evaluation metric: Average Precision (AP)

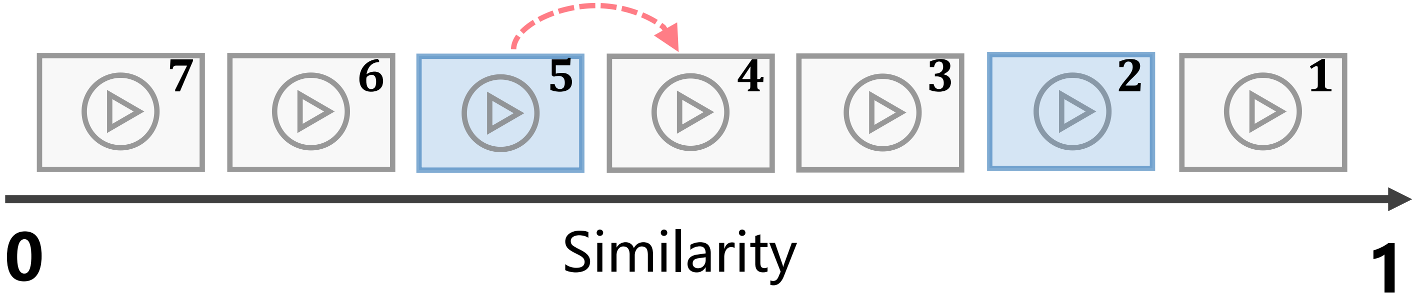
- ➢ AP focuses on the top-ranked pairs



$$AP = \frac{1}{2}\left(\frac{1}{2} + \frac{2}{5}\right) = 0.45$$



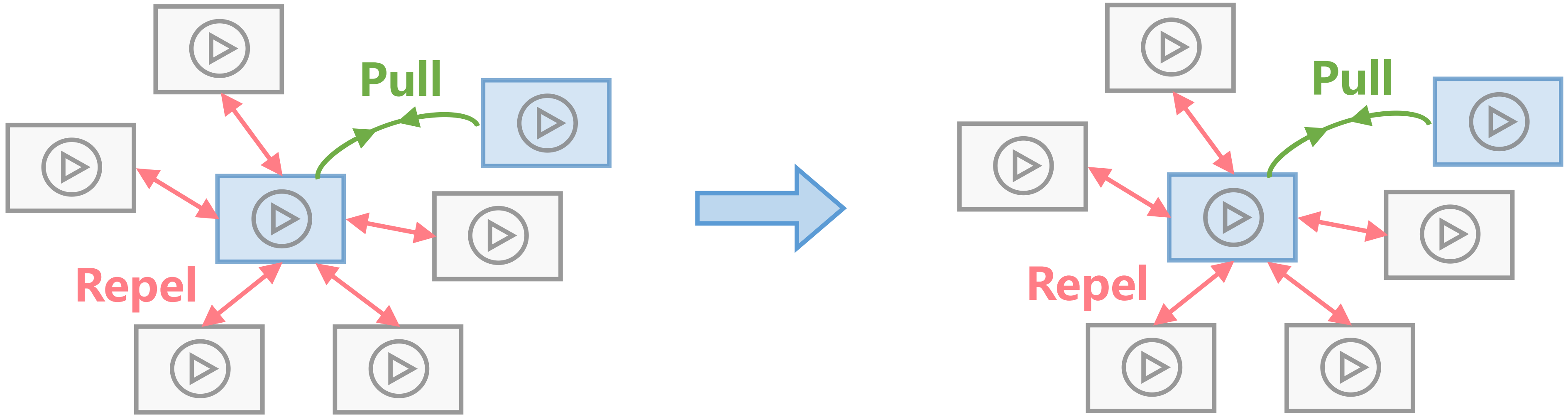
$$AP = \frac{1}{2}\left(\frac{1}{1} + \frac{2}{5}\right) = 0.70 \ (0.25 \uparrow)$$



$$AP = \frac{1}{2}\left(\frac{1}{2} + \frac{2}{4}\right) = 0.50 \ (0.05 \uparrow)$$

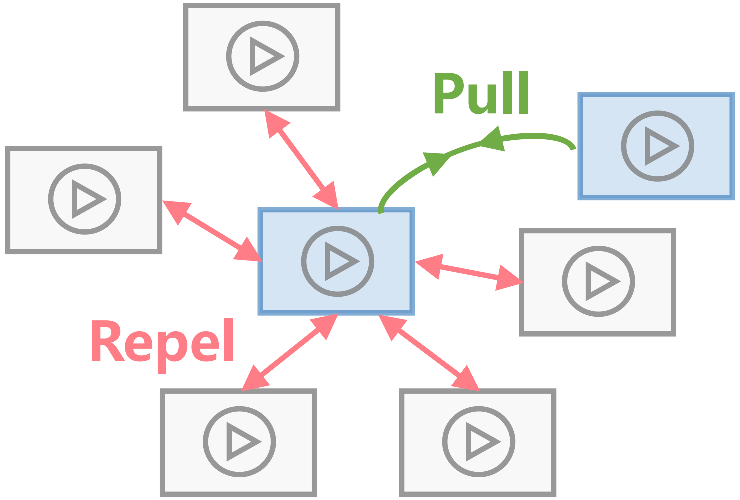
Background

- **Training objective of previous video retrieval methods**
  - **Pair-wise objective:**
    - ➢ Assigns equal weight to all video pairs



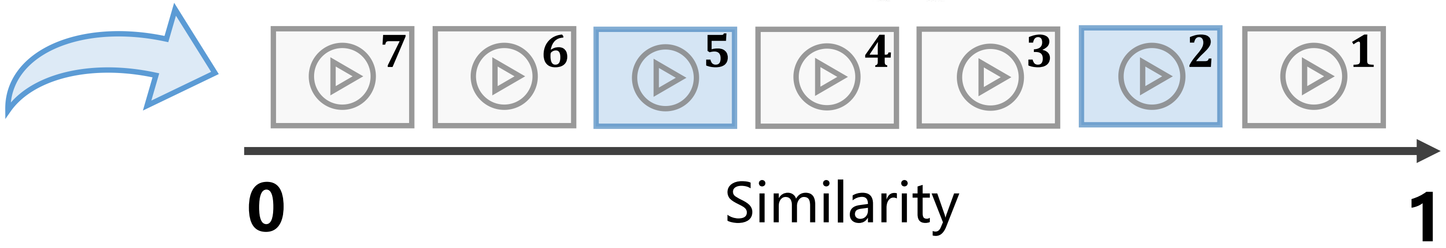
Background

Are Pairs All Equal?



Mismatch!

$$AP = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{r_i}$$

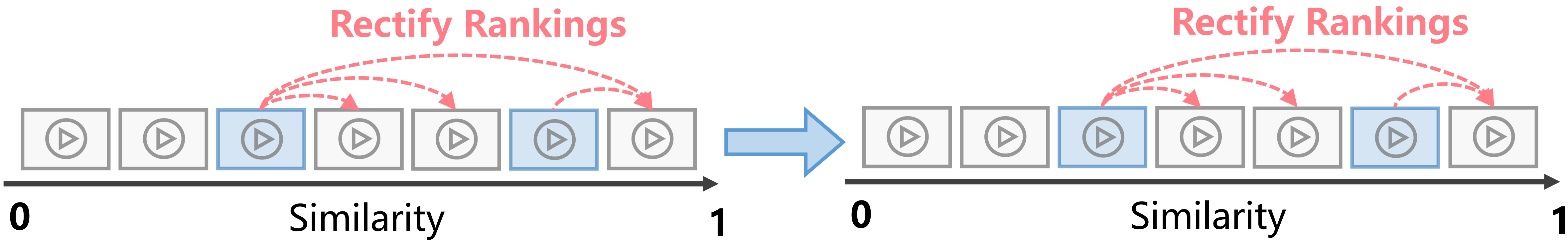


Pair-wise Training Objective

Ranking-based Evaluation Metric

Background

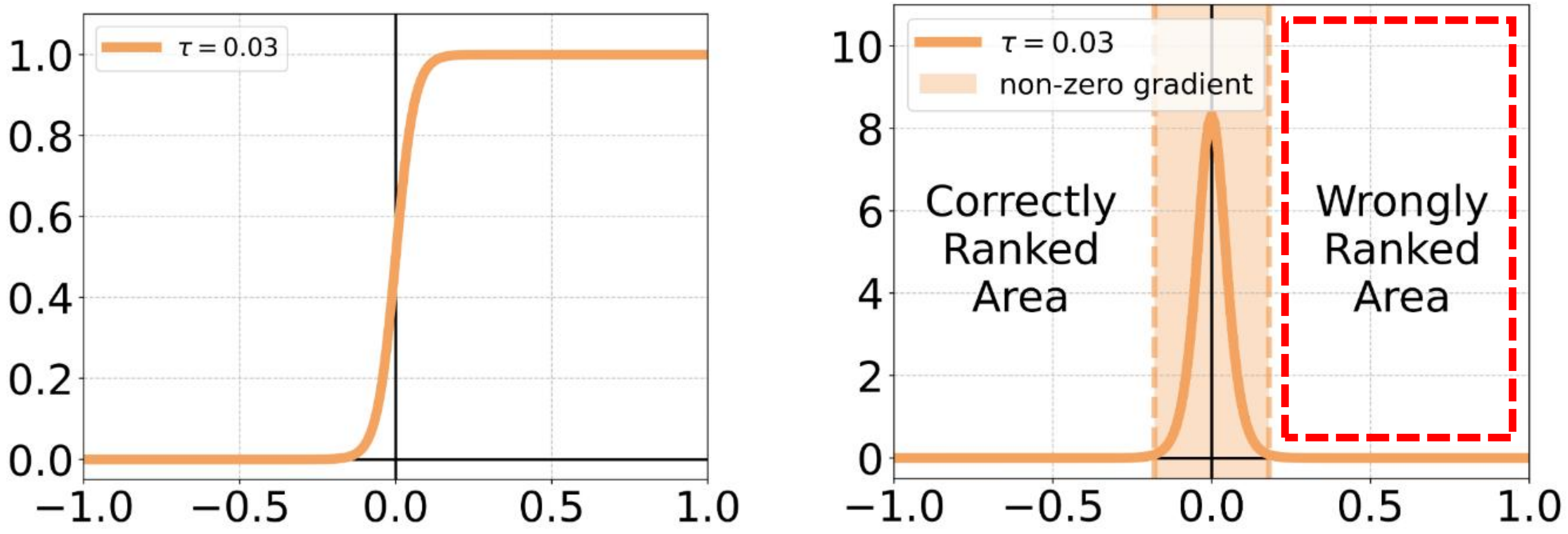
- **New training objective for video retrieval**
  - **AP-based objective**
    - ➢ Rectify the wrongly ranked positive-negative pairs in the list



Background

◻ Challenge 1: Gradient vanishing issue in AP loss

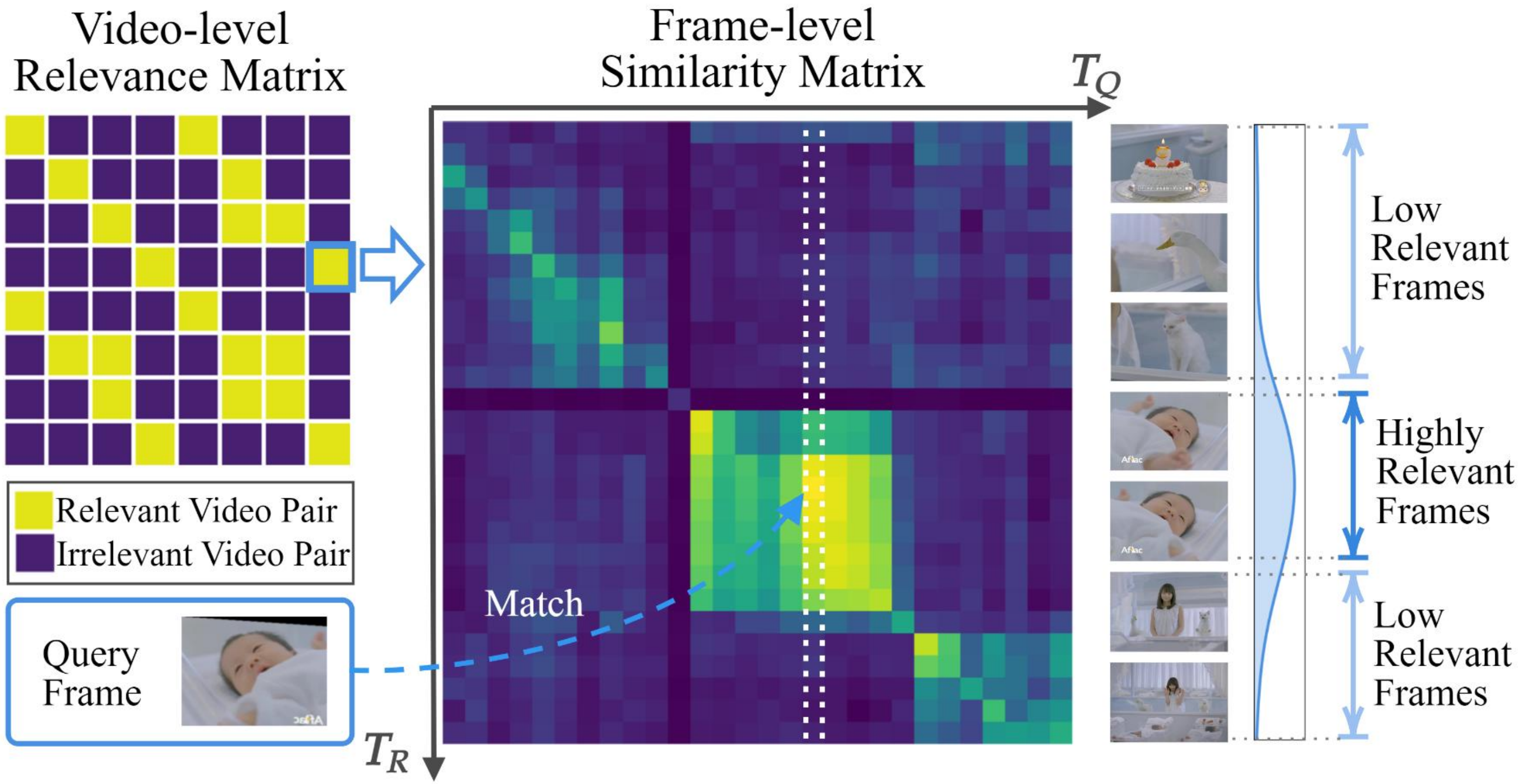
- Seriously mis-ranked pairs fall into gradient vanishing area
- Additional matching complexity of videos intensified this issue



(a) $\mathcal{G}(x;\tau)$ in Smooth-AP. **(b) Derivative of $\mathcal{G}(x;\tau)$.**

Background

- **Challenge 2: Noisy frame-level matching leads to a biased AP estimation**
  - Two relevant videos might not share consistent relevance across all frame pairs
  - This ambiguity harms the rankings of relevant videos when optimizing AP



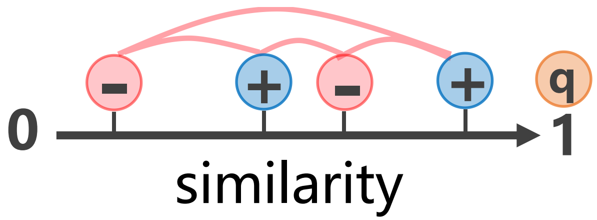
Method

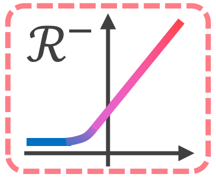
□ For challenge 1: Reformulate AP to design our gradient-enhanced loss

$$\widehat{AP}^{\downarrow}_{k}(f) = \frac{1}{|\boldsymbol{S}^{k+}|} \sum_{s_{ki} \in \boldsymbol{S}^{k+}} h\left(\frac{\sum_{s_{kj} \in \boldsymbol{S}^{k-}} \mathcal{R}^{-}(d^{k}_{ji}; \delta)}{1 + \rho \sum_{s_{kj} \in \boldsymbol{S}^{k+}} \mathcal{R}^{+}(d^{k}_{ji})} \right)$$

Positive-negative Pairs

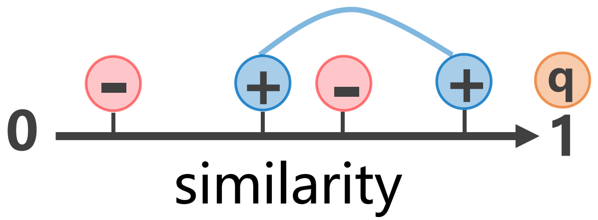
- Rankings should be **corrected**
- Require proper gradients for optimization
- $\mathcal{R}^{-}(x;\delta) = \begin{cases} \frac{2}{\delta}x + 1, & \text{if } x \geq 0 \\ \frac{1}{\delta^2}x^2 + \frac{2}{\delta}x + 1, & \text{if } -\delta \leq x < 0 \\ 0, & \text{if } x < -\delta \end{cases}$

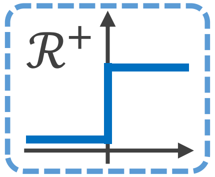




Positive-positive Pairs

- Serve as **weights**
- Not needed for optimization
- $\mathcal{R}^{+}(x) = \mathcal{H}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}$





Method

□ For challenge 1: Decompose AP to design our gradient-enhanced AP loss

Previous AP

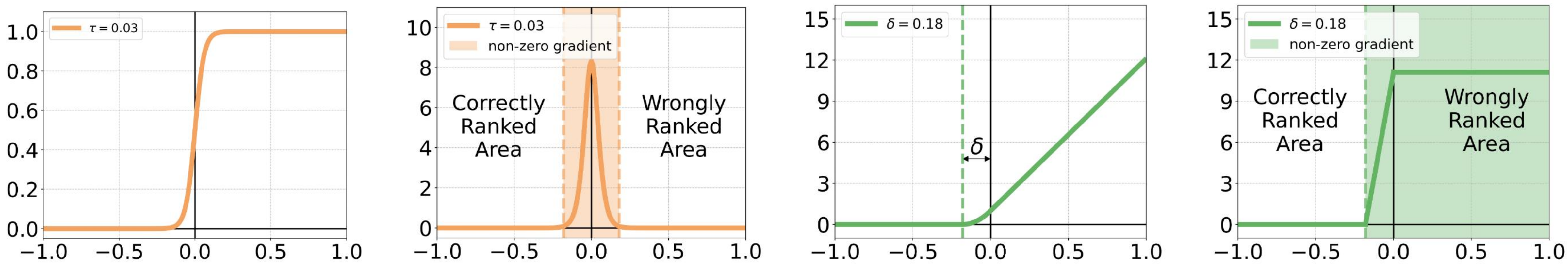
$$\frac{\sum_{s_{kj}\in \boldsymbol{S}^{k-}} \mathcal{G}(d^k_{ji};\tau)}{1+\sum_{s_{kj}\in \boldsymbol{S}^{k+}} \mathcal{G}(d^k_{ji};\tau)}$$

Gradient Vanishing

QuadLinear-AP (Ours)

$$\frac{\sum_{s_{kj}\in \boldsymbol{S}^{k-}} \mathcal{R}^-(d^k_{ji};\delta)}{1+\rho\sum_{s_{kj}\in \boldsymbol{S}^{k+}} \mathcal{R}^+(d^k_{ji})}$$

Favorable properties



(a) $\mathcal{G}(x;\tau)$ in Smooth-AP. **(b) Derivative of $\mathcal{G}(x;\tau)$.** **(c) $\mathcal{R}^-(x;\delta)$ in QuadLinear-AP.** **(d) Derivative of $\mathcal{R}^-(x;\delta)$.**

Method

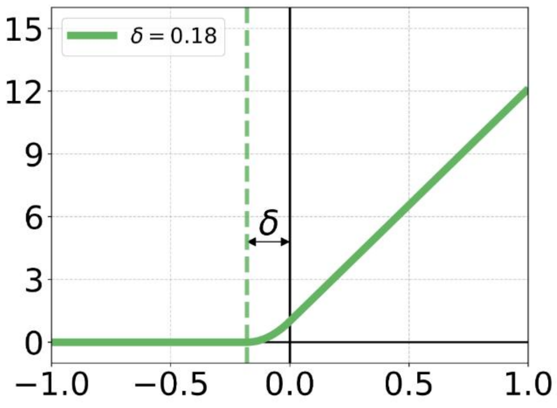
ACM Multimedia 2024
Melbourne, Australia

□ For challenge 1: Decompose AP to design our gradient-enhanced AP loss

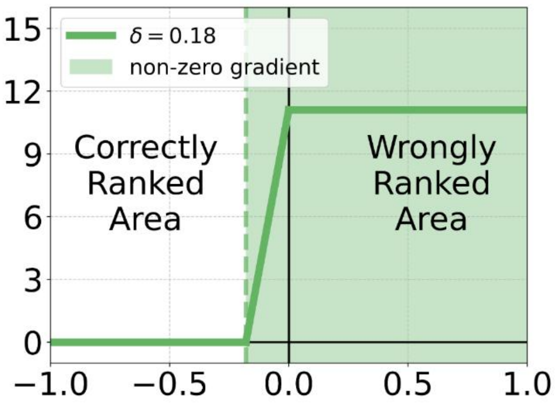
QuadLinear-AP (Ours)

$$\frac{\sum_{s_{kj}\in S^{k-}} \mathcal{R}^-(d^k_{ji};\delta)}{1+\rho\sum_{s_{kj}\in S^{k+}} \mathcal{R}^+(d^k_{ji})}$$

Favorable properties



(c) $\mathcal{R}^-(x;\delta)$ in QuadLinear-AP.



(d) Derivative of $\mathcal{R}^-(x;\delta)$.

Favorable properties for optimization

- ➢ Suitable gradients for low AP area
- ➢ Differentiable AP optimization
- ➢ Convex, smooth and continuous
- ➢ Upper bound of Heaviside function
- ➢ Monotonically increasing (non-strictly)

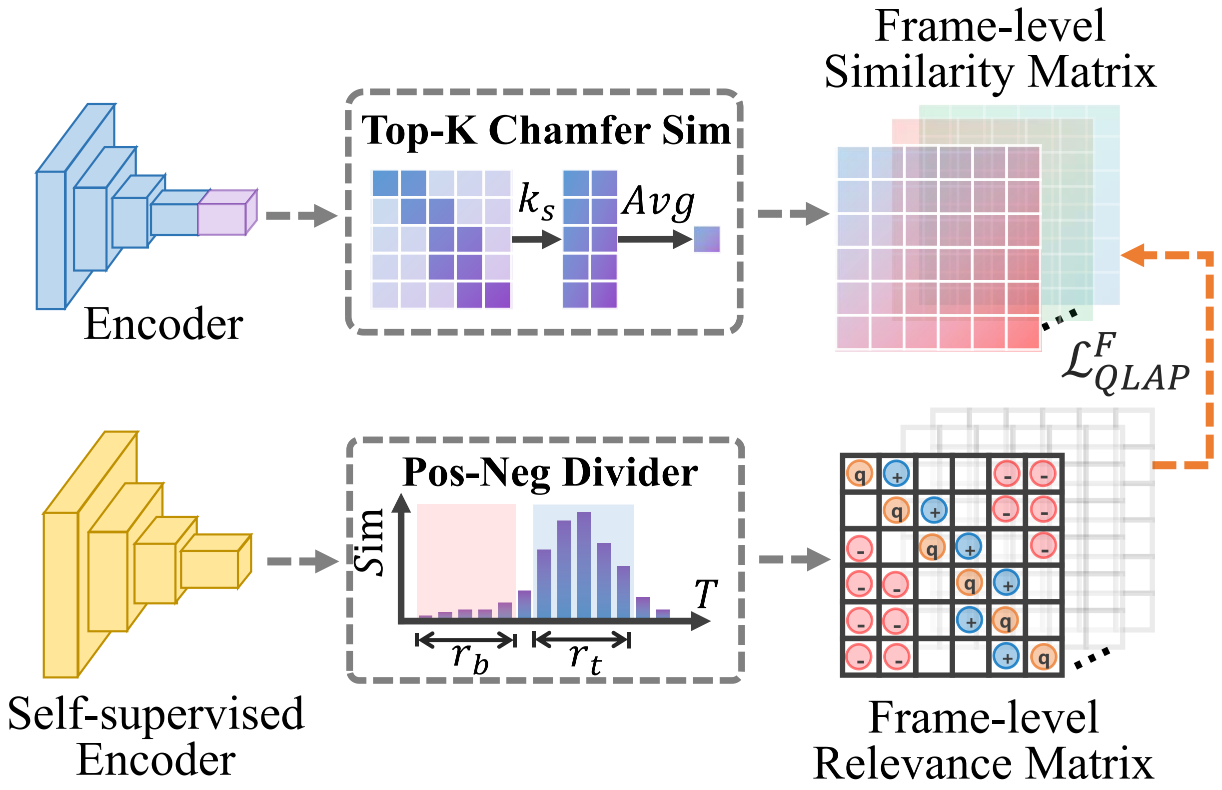
Method

□ For challenge 2: Label purification to form noise-robust frame matching

① Similarity aggregation
- ➢ Top-K Chamfer Similarity measure
- ➢ Fine-grained similarity aggregation

② Pseudo label generation
- ➢ Self-supervised encoder
- ➢ Positive-negative divider

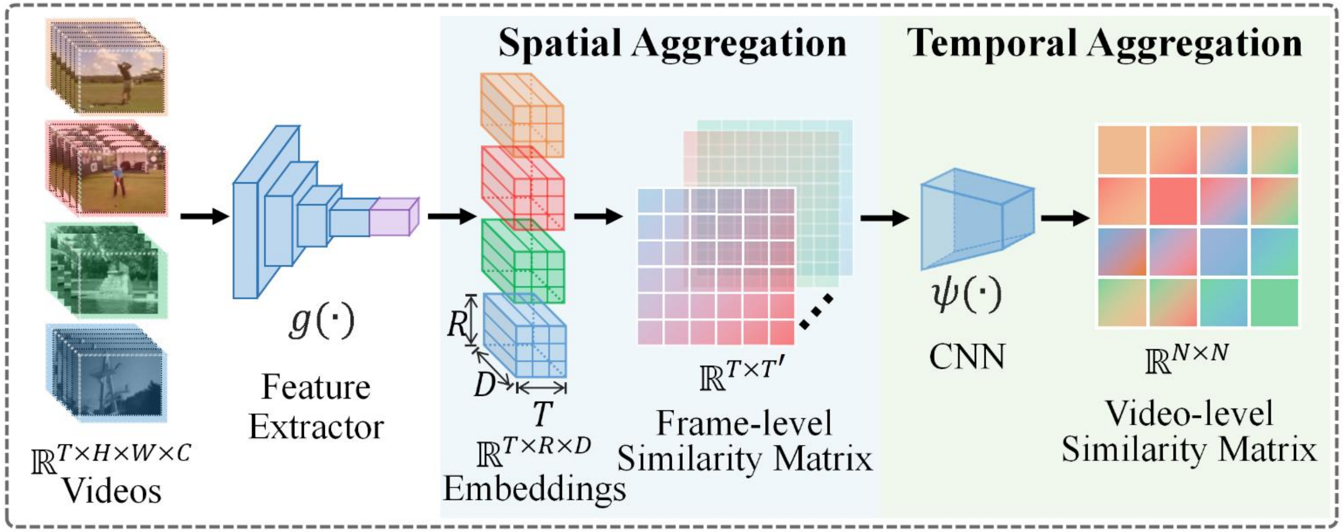
③ Noise-robust AP loss
- ➢ QuadLinear-AP Adaption
- ➢ Reduce weights of noisy pairs



Framework

☐ Bottom-up video similarity measure

- Patch: $S(\boldsymbol{V}, \boldsymbol{V}') = cosine(g(\boldsymbol{V}), g(\boldsymbol{V}'))$
- Frame: $m_s(\boldsymbol{V}, \boldsymbol{V}')_{x,y} = \frac{1}{RK} \sum_{i=1}^{R} \sum_{j=1}^{K} S(\boldsymbol{V}, \boldsymbol{V}')_{x,i,[j],y}$
- Video: $f(\boldsymbol{V}, \boldsymbol{V}') = \frac{1}{TK} \sum_{i=1}^{T} \sum_{j=1}^{K} \psi\left(m_s(\boldsymbol{V}, \boldsymbol{V}')\right)_{i,[j]}$



Algorithm 1 Hierarchical AP Optimization

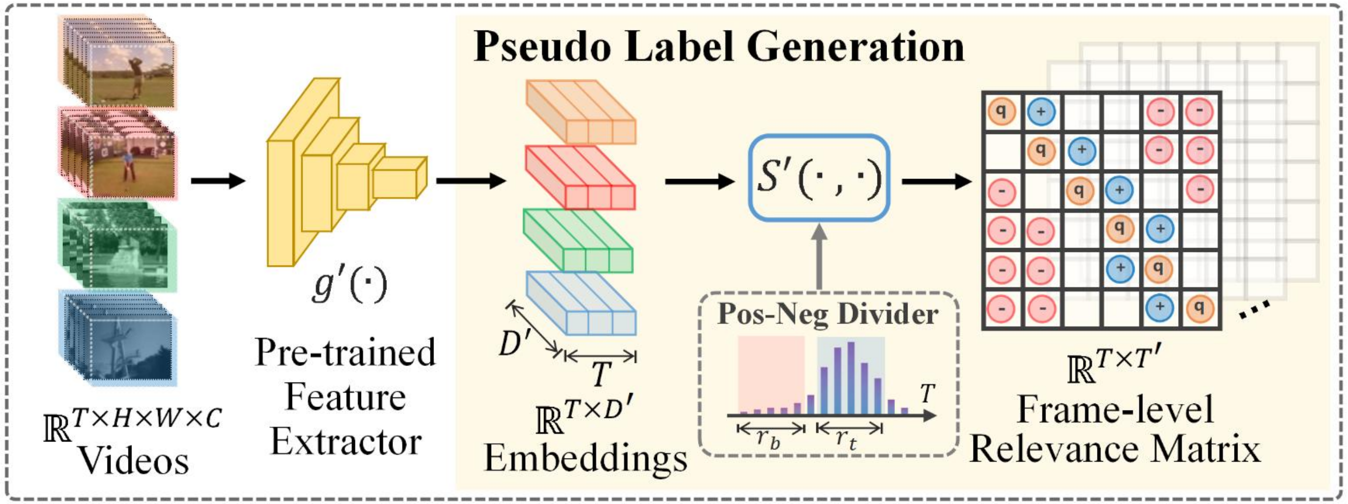
Input: Iteration L, learning rate $\{\eta_l\}_{l=1}^{L}$.
Output: Model parameters $\boldsymbol{\Theta}_{L+1}$.

```
Initialize model parameters Θ_1.
for l = 1 to L do
    Sample a batch of videos {V_i}_{i=1}^N.
    Extract video embeddings g(V_i).
    Calculate similarity matrix by f.
    Extract video embeddings g'(V_i).
    Generate pseudo label matrix Ŷ.
    Compute L^V_QLAP and L^F_QLAP.
    Calculate the total loss L.
    Update: Θ_{l+1} = Θ_l − η_l ∇L.
end for
```

Framework

☐ Pseudo-label generation

- Patch: $S'(\boldsymbol{V}, \boldsymbol{V}') = cosine(g'(\boldsymbol{V}), g'(\boldsymbol{V}')) \in \mathbb{R}^{T \times T'}$
- Frame: $\hat{Y}_{x,y} = \begin{cases} 1, \text{ if } S'(\boldsymbol{V}, \boldsymbol{V}')_{x,y} \geq S'(\boldsymbol{V}, \boldsymbol{V}')_{x,[r_t \times T']} \\ 0, \text{ if } S'(\boldsymbol{V}, \boldsymbol{V}')_{x,y} \leq S'(\boldsymbol{V}, \boldsymbol{V}')_{x,[(1-r_b) \times T']} \end{cases}$



Algorithm 1 Hierarchical AP Optimization

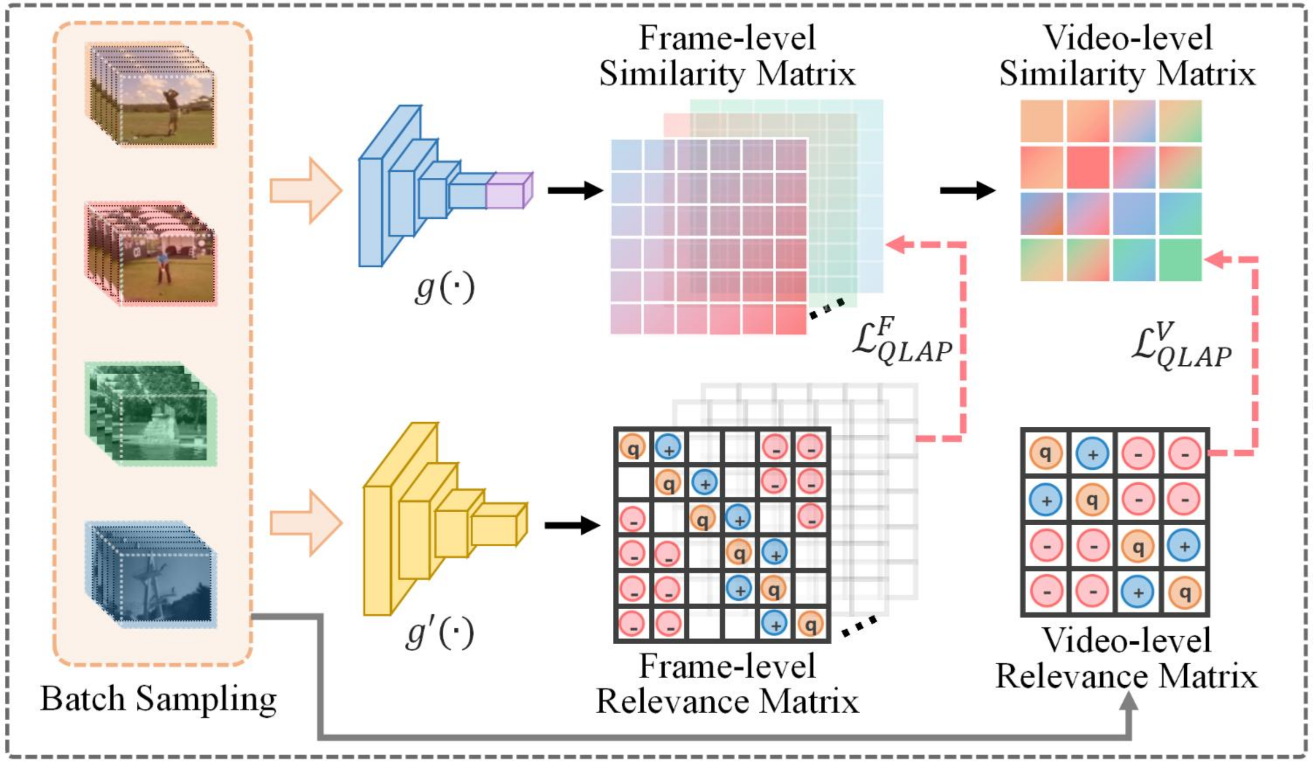
Input: Iteration L, learning rate $\{\eta_l\}_{l=1}^{L}$.
Output: Model parameters $\boldsymbol{\Theta}_{L+1}$.

```
Initialize model parameters Θ_1.
for l = 1 to L do
    Sample a batch of videos {V_i}_{i=1}^N.
    Extract video embeddings g(V_i).
    Calculate similarity matrix by f.
    Extract video embeddings g'(V_i).
    Generate pseudo label matrix Ŷ.
    Compute L^V_QLAP and L^F_QLAP.
    Calculate the total loss L.
    Update: Θ_{l+1} = Θ_l − η_l ∇L.
end for
```

Framework

□ Hierarchical AP optimization

➢ Total loss: $\mathcal{L} = \underbrace{\lambda_f \mathcal{L}^F_{QLAP}}_{frame-level} + \underbrace{\lambda_v \mathcal{L}^V_{QLAP} + \mathcal{L}_{base}}_{video-level}$

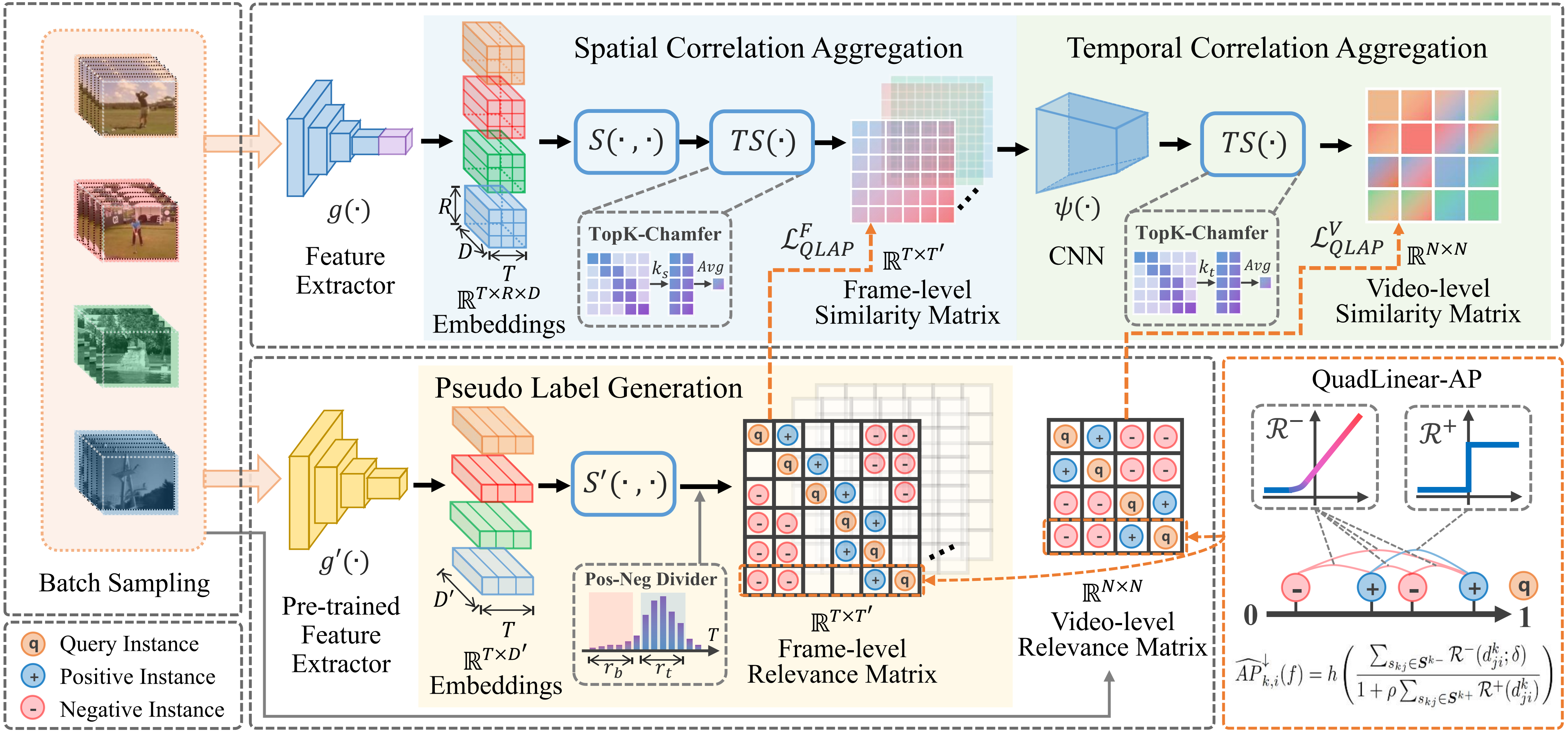


Algorithm 1 Hierarchical AP Optimization

Input: Iteration L, learning rate $\{\eta_l\}_{l=1}^{L}$.
Output: Model parameters $\mathbf{\Theta}_{L+1}$.
1: Initialize model parameters $\mathbf{\Theta}_1$.
2: **for** $l = 1$ to L **do**
3: Sample a batch of videos $\{\boldsymbol{V}_i\}_{i=1}^{N}$.
4: Extract video embeddings $g(\boldsymbol{V}_i)$.
5: Calculate similarity matrix by f.
6: Extract video embeddings $g'(\boldsymbol{V}_i)$.
7: Generate pseudo label matrix $\hat{Y}$.
8: Compute $\mathcal{L}^V_{QLAP}$ and $\mathcal{L}^F_{QLAP}$.
9: Calculate the total loss $\mathcal{L}$.
10: Update: $\mathbf{\Theta}_{l+1} = \mathbf{\Theta}_l - \eta_l \nabla \mathcal{L}$.
11: **end for**

Framework

Hierarchical learning for Average-Precision-oriented Video Retrieval (HAP-VR)



Experiments

ACM Multimedia 2024 Melbourne, Australia

- Our HAP-VR improves mAP and μ AP effectively on multiple benchmarks

Method	Label	Trainset	Retrieval (mAP)					Detection (μAP)				
			EVVE	SVD	FIVR-200K			EVVE	SVD	FIVR-200K		
					DSVR	CSVR	ISVR			DSVD	CSVD	ISVD
DML† [32]	✓	VCDB ($\mathcal{C}\&\mathcal{D}$)	61.10	85.00	52.80	51.40	44.00	75.50	/	39.00	36.50	30.00
TMK† [46]	✓	internal	61.80	86.30	52.40	50.70	42.50	/	/	/	/	/
TCA [53]	✓	VCDB ($\mathcal{C}\&\mathcal{D}$)	63.08	89.82	86.81	82.31	69.61	76.90	56.93	69.09	62.28	49.24
ViSiL† [30]	✓	VCDB ($\mathcal{C}\&\mathcal{D}$)	65.80	88.10	89.90	85.40	72.30	79.10	/	75.80	69.00	53.00
DnS (S_a) [34]	✓	DnS-100K	65.33	90.20	92.09	87.54	74.08	74.56	72.24	79.66	69.51	54.20
DnS (S_b) [34]	✓	DnS-100K	64.41	89.12	90.89	86.28	72.87	75.80	66.53	78.05	68.52	53.48
LAMV† [2]	✗	YFCC100M	62.00	88.00	61.90	58.70	47.90	80.60	/	55.40	50.00	38.80
VRL† [24]	✗	internal	/	/	90.00	85.80	70.90	/	/	/	/	/
ViSiL$_f$† [30]	✗	ImageNet	62.70	/	89.00	84.80	72.10	74.60	/	66.90	59.50	45.90
S^{2}VS [33]	✗	VCDB ($\mathcal{D}$)	67.17	88.40	92.53	87.73	74.51	80.72	65.04	86.12	77.41	63.26
HAP-VR (Ours)	✗	VCDB ($\mathcal{D}$)	**69.15**	89.00	**92.83**	**88.21**	**74.72**	**82.88**	67.87	**88.41**	**79.85**	**64.79**

Experiments

ACM Multimedia 2024 Melbourne, Australia

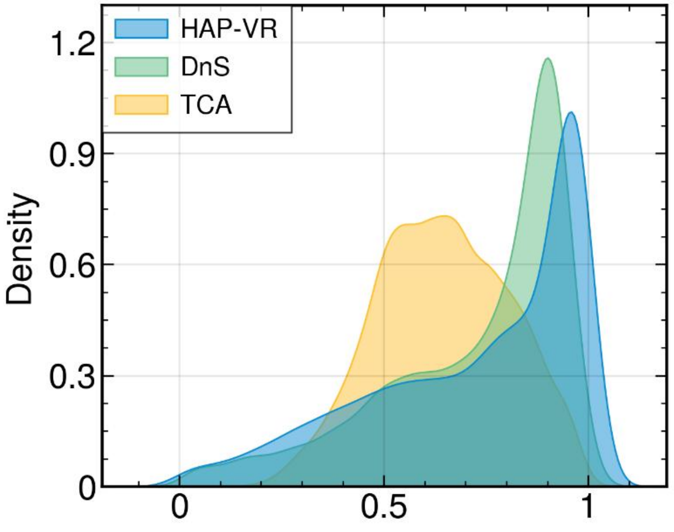
- Our QuadLinear-AP outperforms previous pair-wise and AP-based losses

Losses	Retrieval (mAP)			Detection (μAP)		
	DSVR	CSVR	ISVR	DSVD	CSVD	ISVD
MAE	89.07	88.03	80.86	78.08	75.69	65.26
MSE	89.22	88.26	80.80	78.66	76.07	65.44
Contrastive [18]	88.67	88.09	80.97	75.12	74.23	67.41
Triplet [50]	88.11	87.77	**81.21**	72.94	73.18	69.23
Circle [55]	87.53	86.11	78.77	73.26	71.15	59.33
FastAP [6]	89.30	88.42	81.16	78.83	77.51	**69.95**
DIR [47]	89.65	**88.57**	80.64	78.50	76.22	65.42
BlackBox [45]	**89.70**	88.55	80.53	**80.07**	77.37	66.00
Smooth-AP [4]	89.36	88.33	80.73	79.85	**77.75**	68.42
QuadLinear-AP (**Ours**)	**90.80**	**89.68**	**81.31**	**82.92**	**80.03**	**71.45**

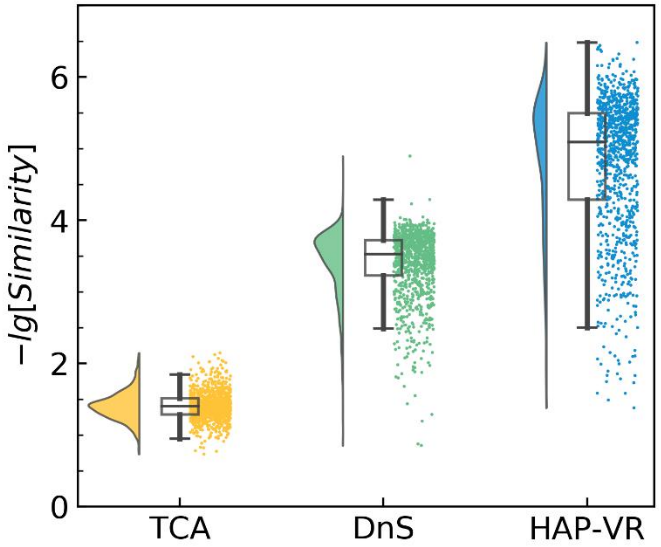
Experiments

◻ Visualization for comparison and ablation study of our HAP-VR

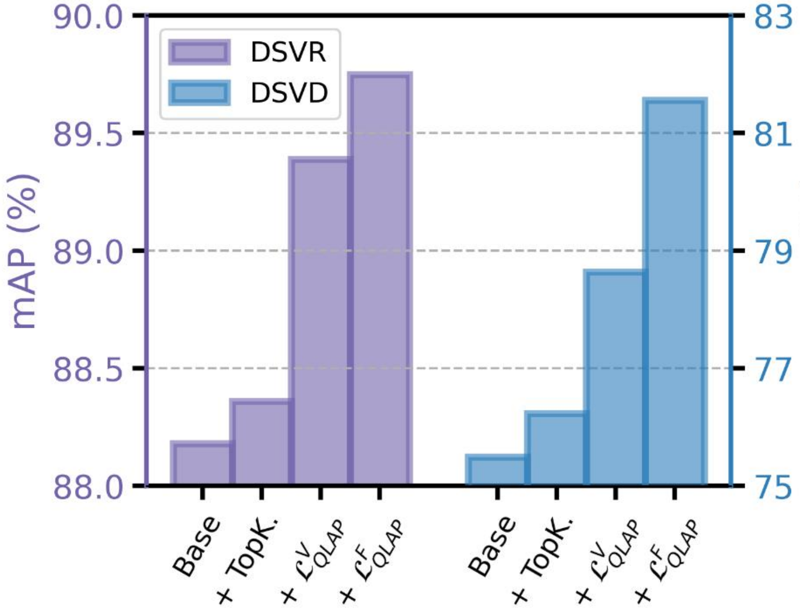
- ➢ A clearer margin between scores of relevant and irrelevant pairs
- ➢ Each component makes progress on the AP-based metrics



(a) Relative pair distribution



(b) Irrelative pair distribution



(c) Effect of components

Thanks for your listening!

E-mail: liuyang232@mails.ucas.ac.cn

Session 2, Poster P107, 30 Oct.

Paper **Code**