Concept Search Versus Embedding Search

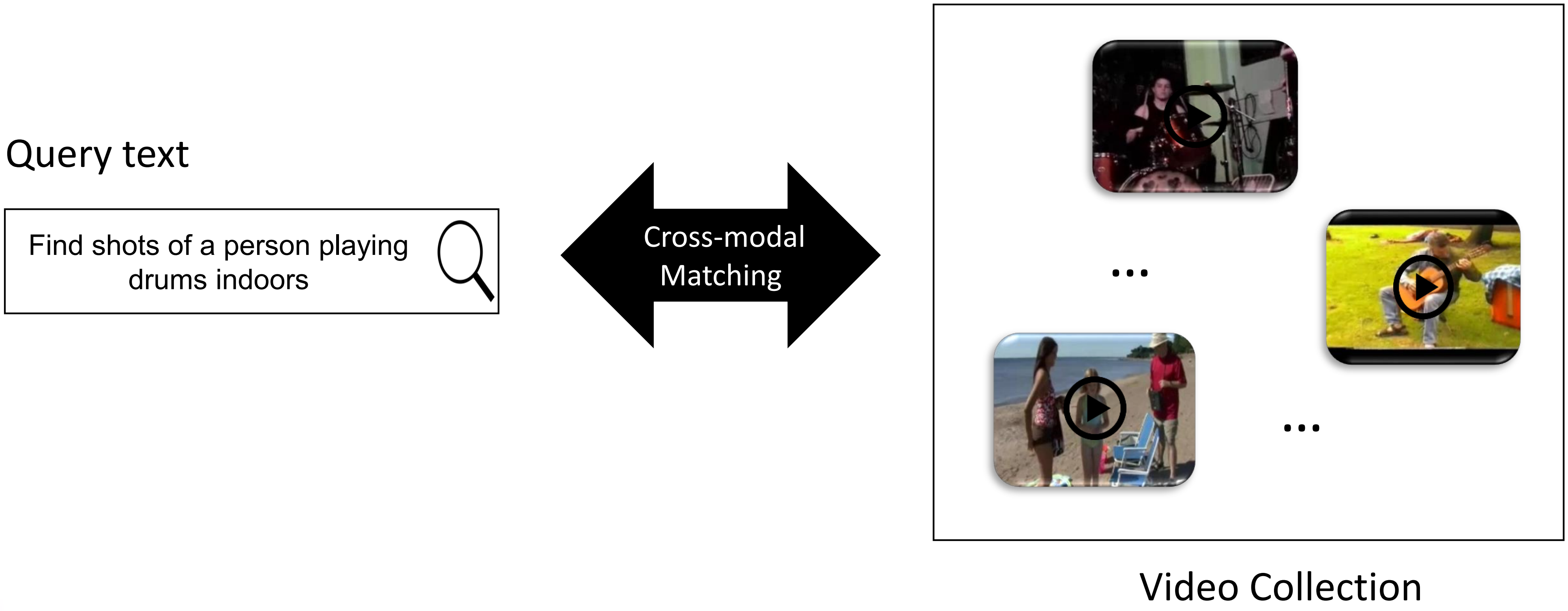
Jiaxin Wu, Phuong Anh Nguyen, and Chong-Wah Ngo
City University of Hong Kong

TRECVID 2020 Workshop

Outlines

- Introduction
- Our dual-task model
- Experiments and results
- Limitations

Ad-hoc Video Search



Query text

Find shots of a person playing drums indoors

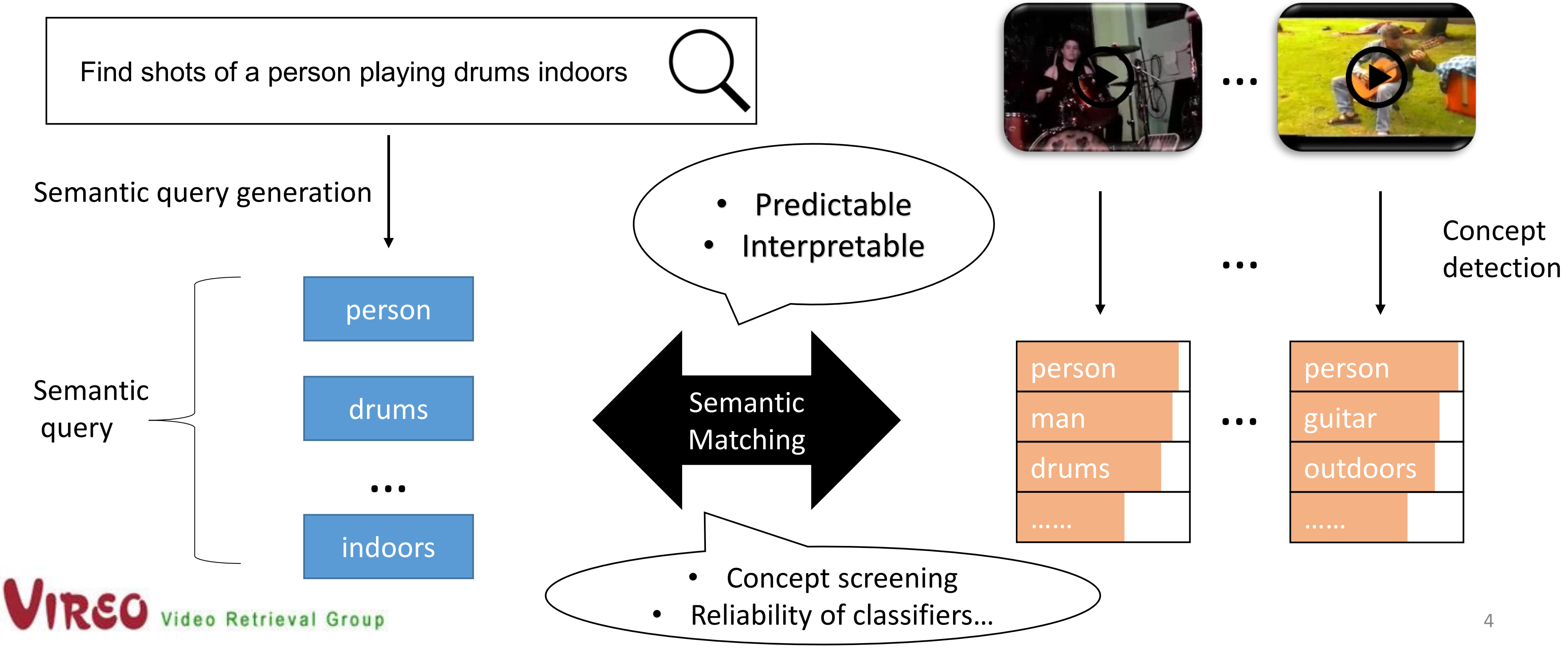
Cross-modal Matching

…

…

Video Collection

Two mainstreams -- Concept Search



Find shots of a person playing drums indoors

Semantic query generation

Semantic query

person

drums

...

indoors

- Predictable
- Interpretable

Semantic Matching

- Concept screening
- Reliability of classifiers…

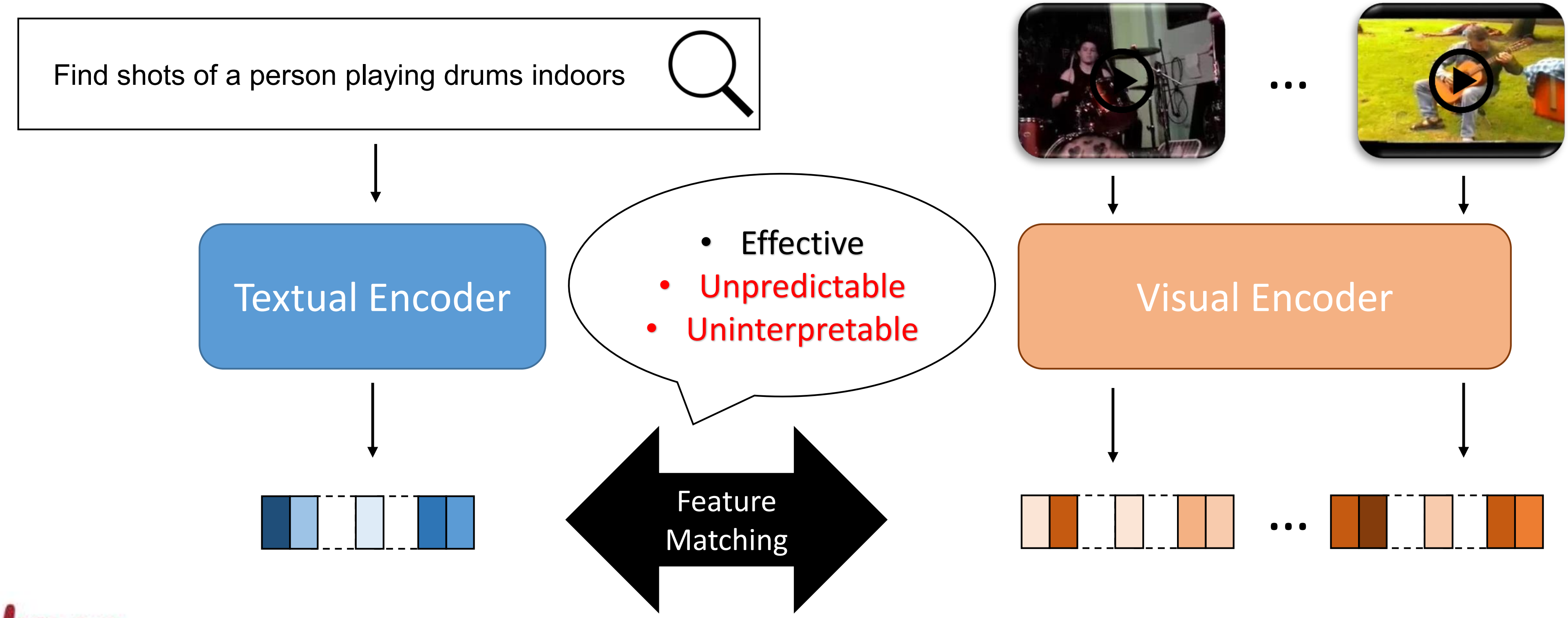
Concept detection

person
man
drums
……

...

person
guitar
outdoors
……

Two mainstreams -- Embedding Search



Find shots of a person playing drums indoors

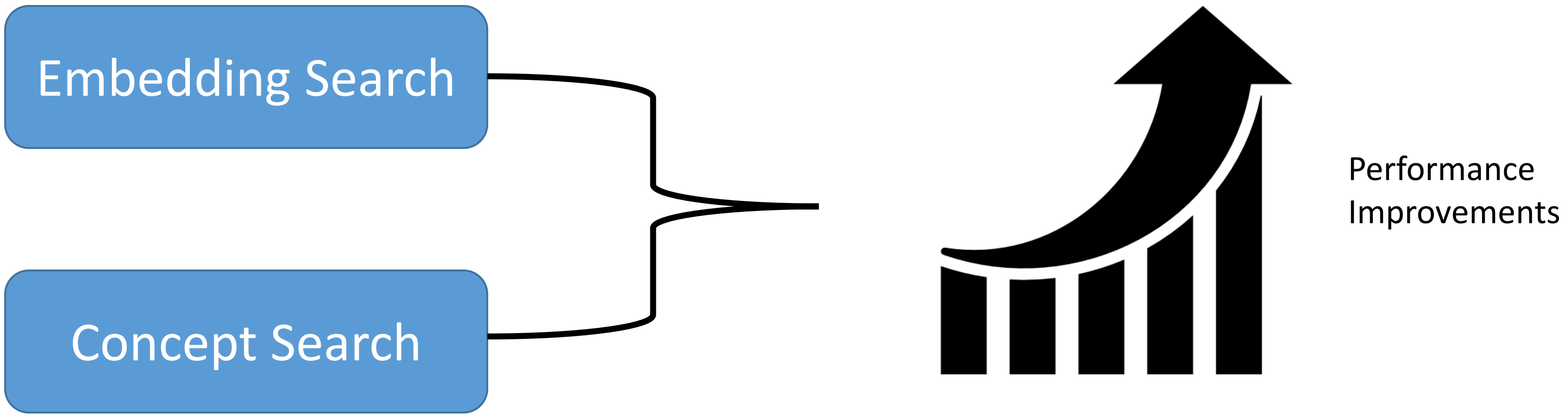
Textual Encoder

- Effective
- Unpredictable
- Uninterpretable

Visual Encoder

Feature Matching

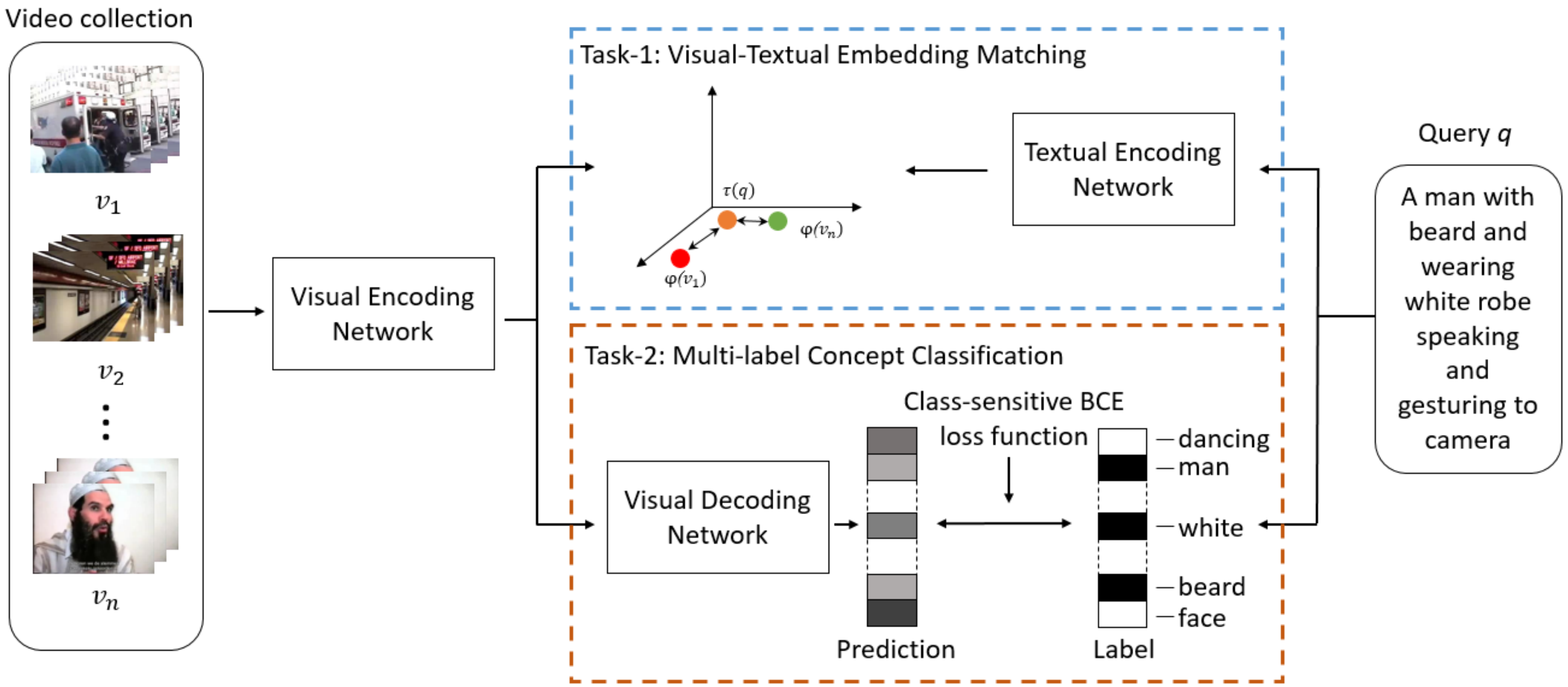
Previous work



But two models are trained on different datasets.

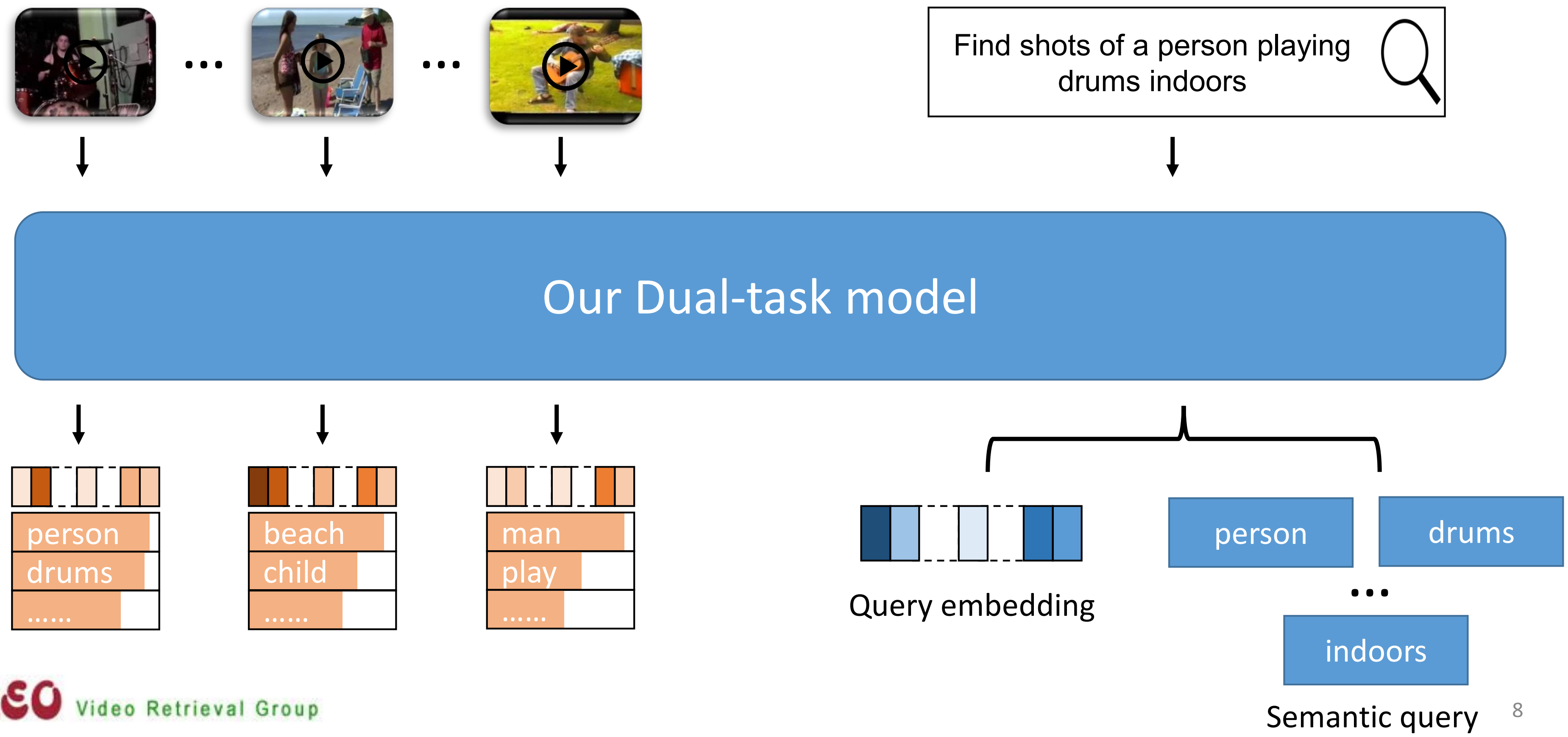
- needs more effort
- Fusion result could be unpredictable

Our solution


Video collection
v_1
v_2
v_n
Visual Encoding Network
Task-1: Visual-Textual Embedding Matching
$\tau(q)$
$\varphi(v_n)$
$\varphi(v_1)$
Textual Encoding Network
Query q
A man with beard and wearing white robe speaking and gesturing to camera
Task-2: Multi-label Concept Classification
Class-sensitive BCE loss function
Visual Decoding Network
Prediction
Label
dancing
man
white
beard
face

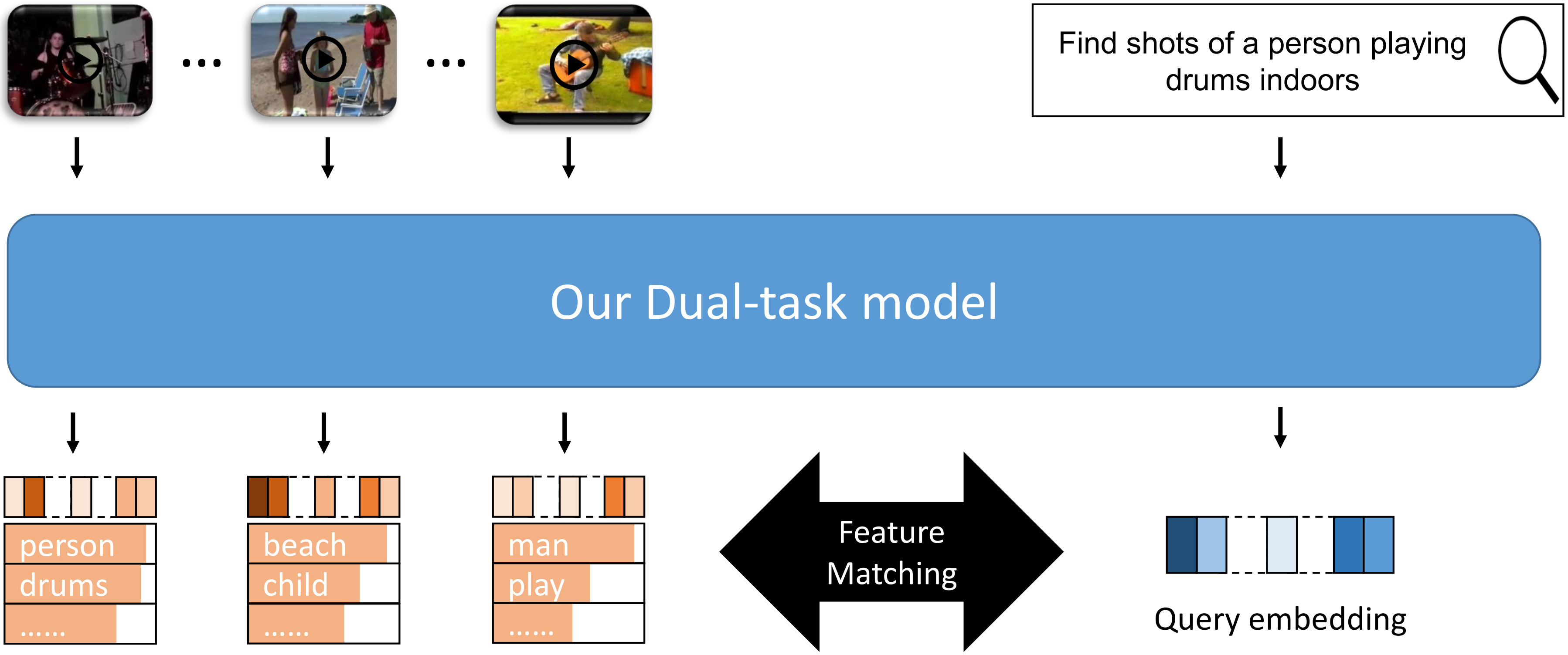
Wu and Ngo, Interpretable Embedding for Ad-hoc Video Search, *ACMMM,* 2020

Our solution



VIREO Video Retrieval Group

Our run1- embedding search



Find shots of a person playing drums indoors

Our Dual-task model

person
drums
……

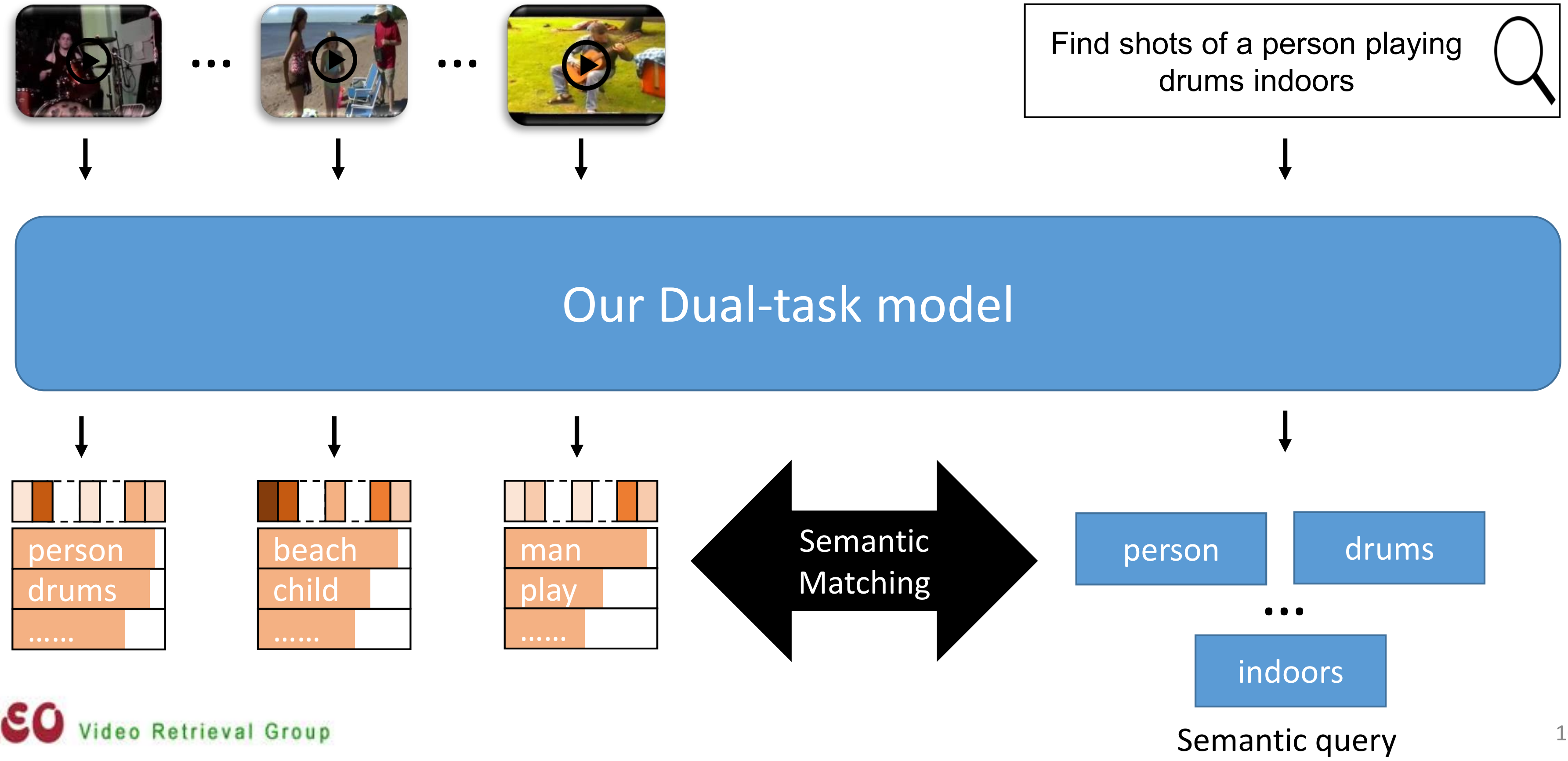
beach
child
……

man
play
……

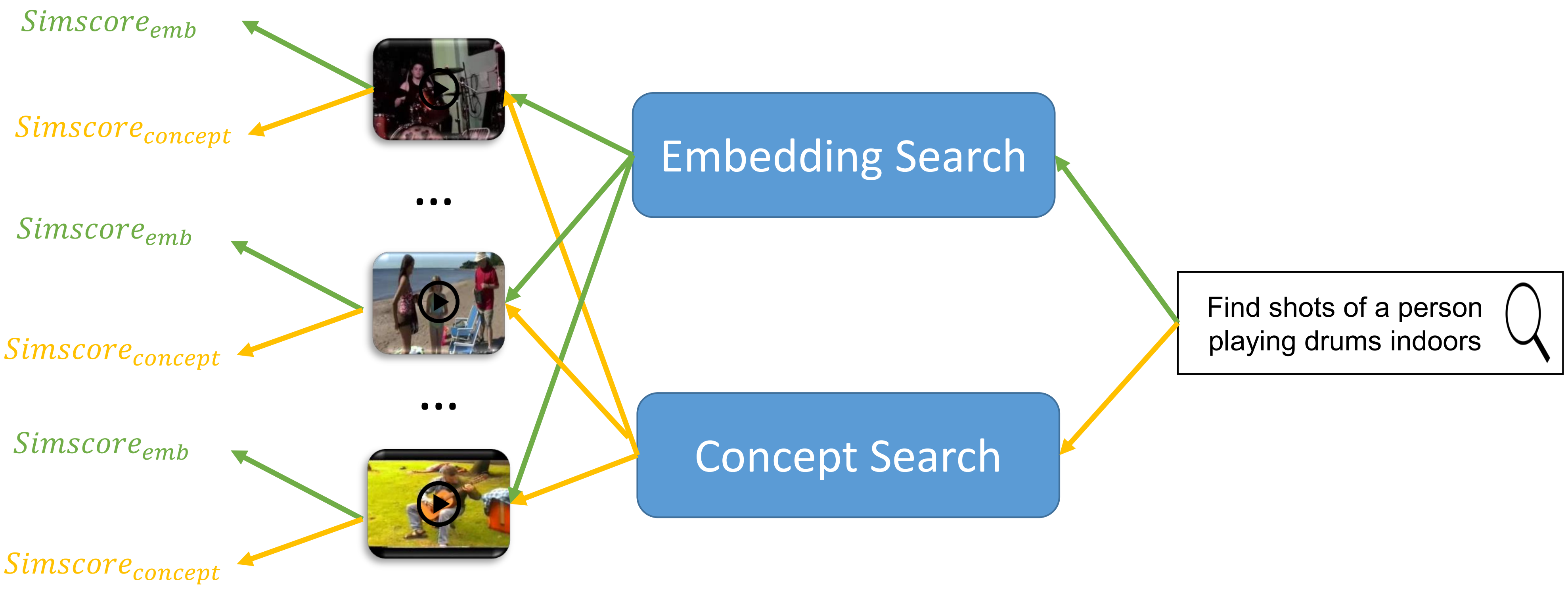
Feature Matching

Query embedding

Our run2 - concept search



Our run3 - fusion search



$$Simscore_{fusion} = (1 - θ) * Simscore_{emb} + θ*Simscore_{concept}$$

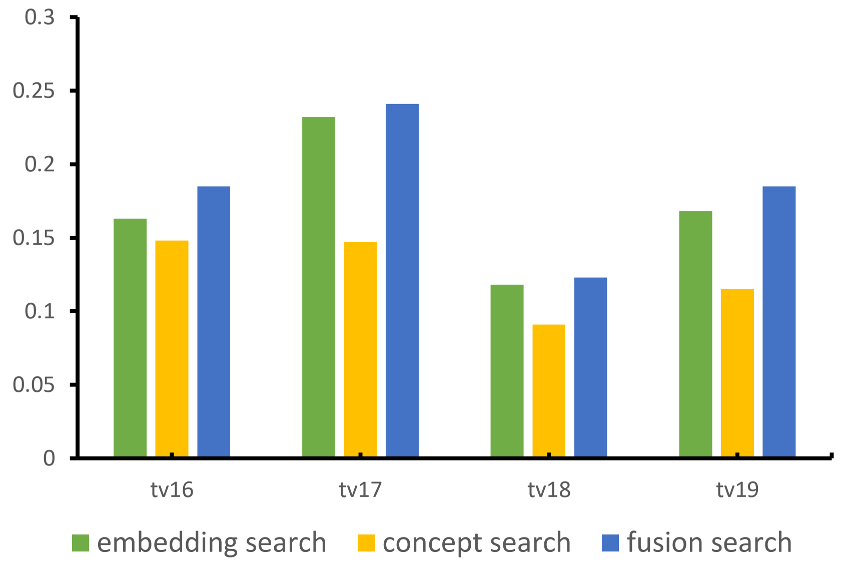
VIREO Video Retrieval Group

Our solution

	#video	#caption	#AVS test query
Training set:			
MSR-VTT	10,000	200,000	
TGIF	100,855	124,534	
VidOR-MPVC	2,496	32,466	
Validation set:			
TV16 VTT training set	200	400	
Test set:			
IACC.3	335,944		90 (tv16-tv18)
V3C1	1,082,659		30 (tv19)

Concept bank is built on all training caption sentences by using words appear more than 5 times.

AVS comparison on tv16-19



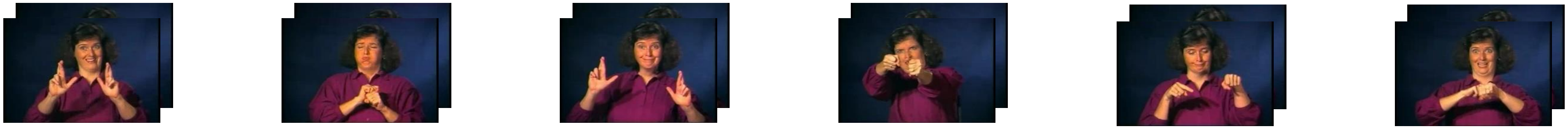
- Embedding search is better than concept search
  - Embedding search performs better at compositional semantics, e.g., sign language
  - Concept search is good at finding individual semantics, e.g., scarf, fountains
- Fusion search attains the best results
  - Consistence brings forward true positive results.
  - Complementary enables fusion search can solve different kinds of queries.

VIREO Video Retrieval Group

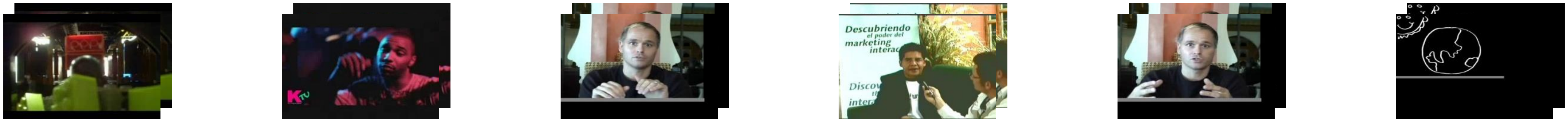
Embedding Search > Concept Search

543 Find shots of a person communicating using sign language

Embedding search: 0.469



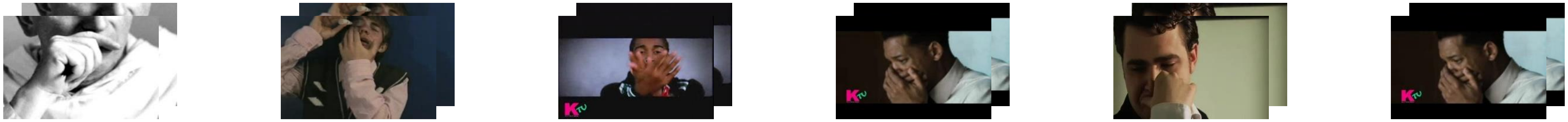
Concept search: 0.000



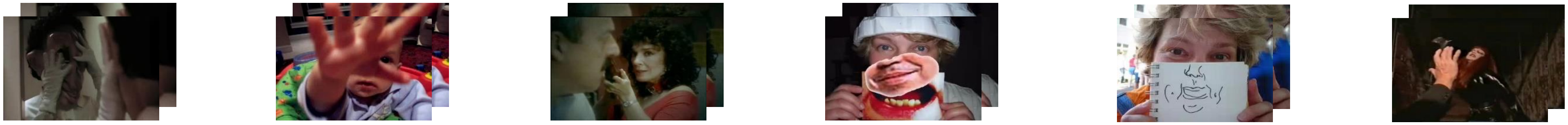
Embedding Search > Concept Search

576 Find shots of a person holding his hand to his face

Embedding search: 0.138



Concept search: 0.000

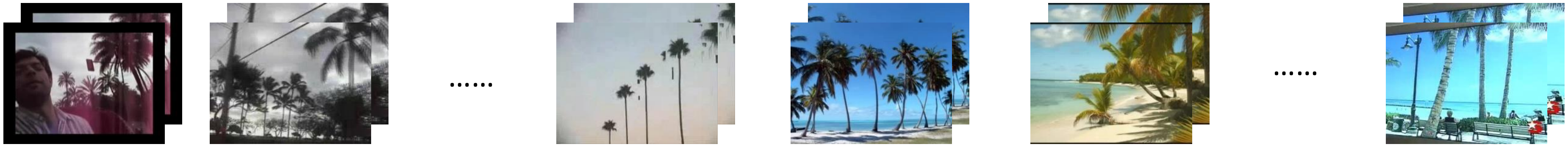


Concept Search > Embedding Search

512 Find shots of palm trees

Embedding search: 0.135

Concept search: 0.335



Concept Search > Embedding Search

558 Find shots of a person wearing a scarf

Embedding search: 0.151

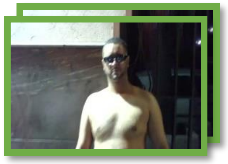
Concept search: 0.516

Fusion search is better

- Fusion search benefits from the consistency in two searches

637 Find shots of a shirtless man standing up or walking outdoors

Embedding search: 0.1985   Concept search: 0.1981   Fusion search: 0.2777



Fusion search	**1**	**2**	**3**	**4**	**7**	**30**	**152**
Embedding search	1	5	7	2	24	84	440
Concept search	33	112	82	393	1	2	3

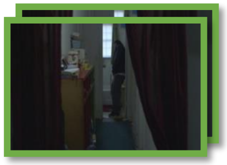
Rank in each search scheme

Fusion search is better

- Fusion search benefits from the complementary in two searches

624 Find shots of a person in front of a curtain indoors

Embedding search: 0.099 Concept search: 0.091 Fusion search: 0.184



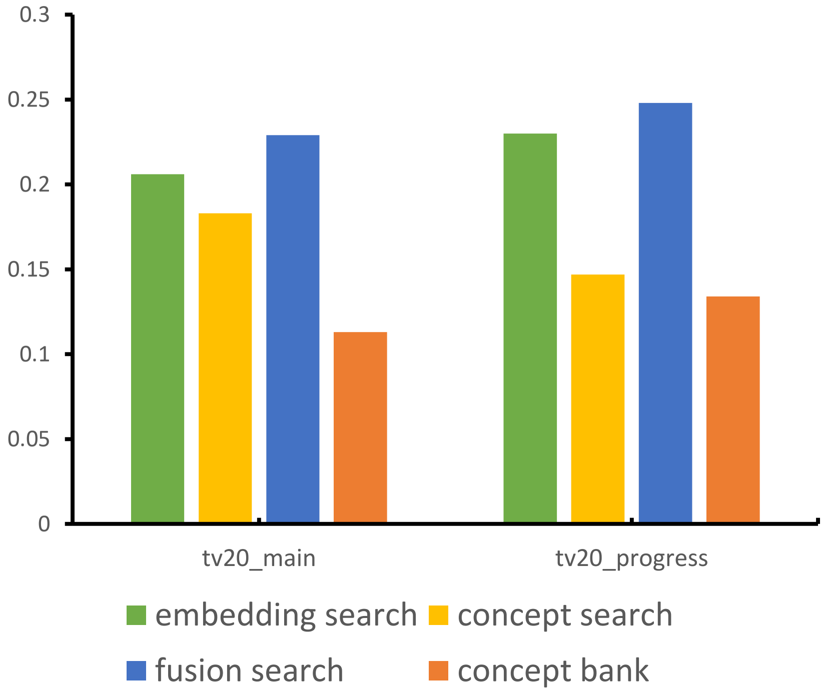
Fusion search	1	2	3	4	6	7	875
Embedding search	1	2	10	11	74	6	Not in top1000
Concept search	531	Not in top1000	101	208	1	Not in top1000	3

Rank in each search scheme

Our run4 – our previous concept model

Type	model	Dataset	#Concepts
Object/person	ResNet152	ImageNet Shuffle	12988
		ImageNet	1000
		TRECVid SIN	346
		Research Collection	497
	FasterRCNN	OpenImage	600
Action	P3D	Kinetics	600
Place	ResNet152	MIT Place	365

AVS comparison on tv20



- The dual-task concept search achieves better performances than our previous concept model.
  - Almost on all queries, except for those OOV queries.
- Fusion search attains the best result.
  - Embedding search outperforms concept search almost on all queries.
  - Concept search shows better results in finding static individual semantics, e.g., necklace, jeans.

VIREO Video Retrieval Group

Limitations

- Suffers from out-of-vocabulary problem

Query	Concept search	Embedding search	Fusion search
606 Find shots of people queuing	0.001	0.002	0.002
644 Find shots of sailboats in the water	0.016	0.000	0.001

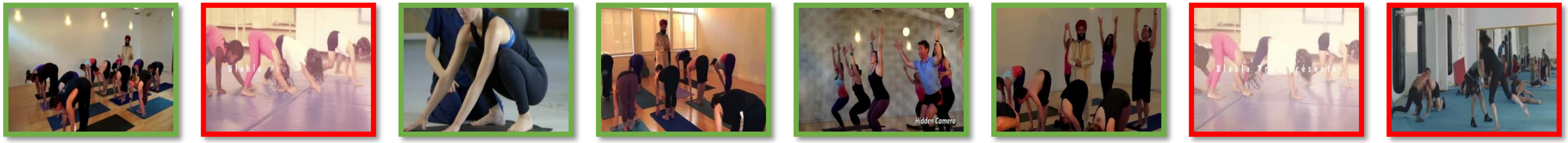
- Bad performances on these queries: contaminated results.

Query	Concept search	Embedding search	Fusion search
594 Find shots of people doing yoga	0.039	0.026	0.040
653 Find shots of group of people clapping	0.028	0.052	0.061
655 Find shots of one or more persons standing in a body of water	0.032	0.049	0.052

Improvement by manual query

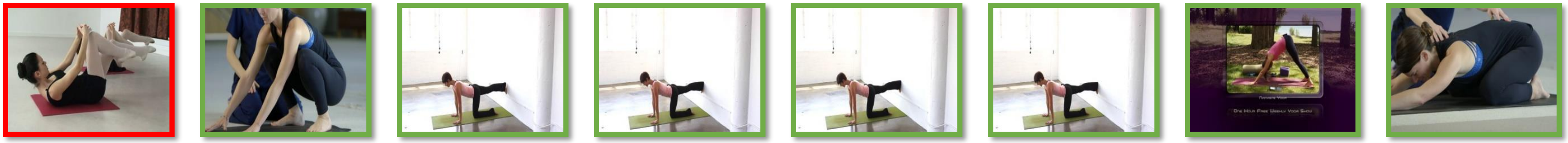
594 Find shots of people doing yoga

Embedding search--automatic run: 0.039



Manually modify the query to Find shots of yoga mat

Embedding search--manual run: 0.197



Improvement by manual query

655 Find shots of one or more persons standing in a body of water

Fusion search--automatic run: 0.052

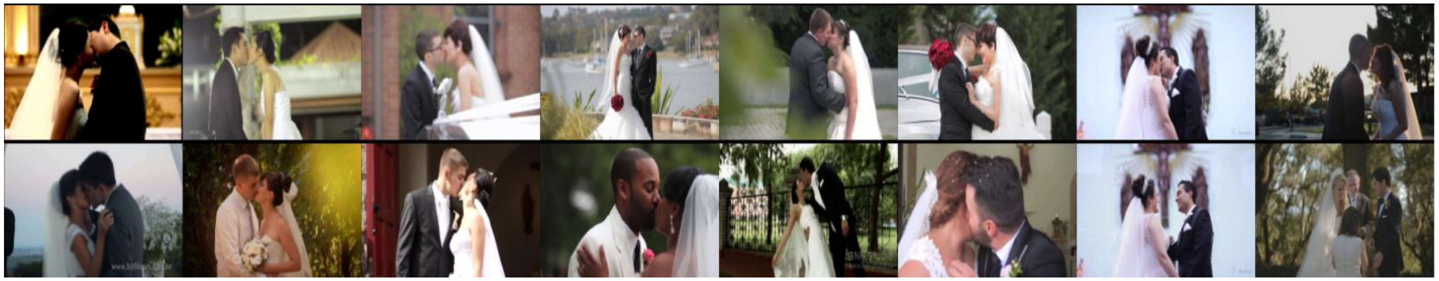
Manually modify the query <u>Find shots of persons standing in water</u>

Fusion search--manual run: 0.112

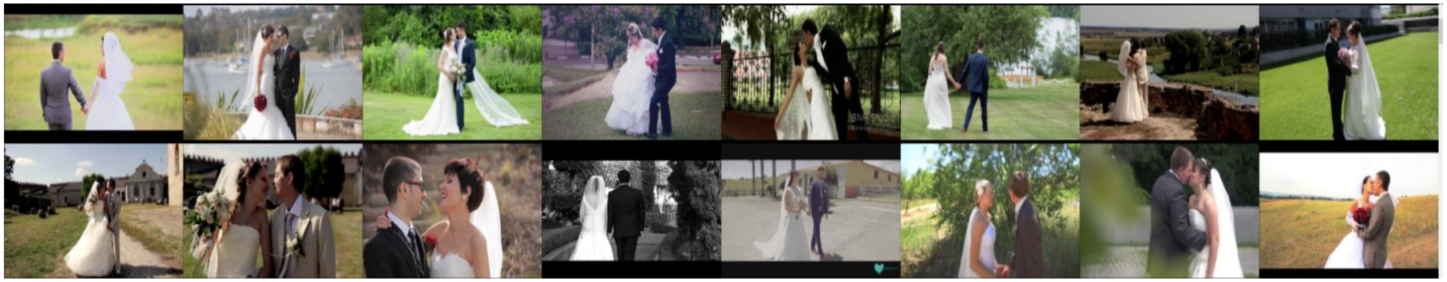
Limitations

- Unable to deal with negation

604 Find shots of bride and groom kissing



607 Find shots of two people kissing who are not bride and groom



Limitations

- Try to solve it by manually splitting the query

(two people kissing) – (bride and groom)

Thank you
Q&A