DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization

Safa Messaoud

Ismini Lourentzou

Assma Boughoula*

Mona Zehni*

Zhizhen Zhao

Chengxiang Zhai

Alex Schwing

sigir21

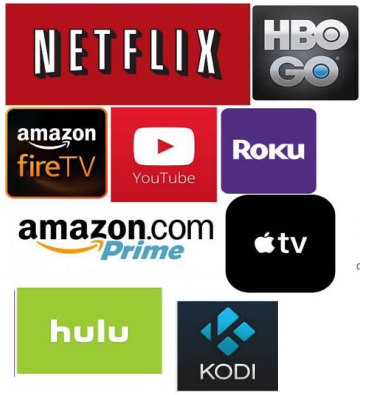
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

COLLEGE OF ENGINEERING COMPUTER SCIENCE VIRGINIA TECH

Growing Popularity of Video and Explosion of Video Sharing Platforms

82% of internet traffic will be from video streaming and downloads in 2022
(Cisco, 2019)

Video watch time:
ca. **10 hours per week**
per internet user in 2020
(Statistica, 2020)



Avg. U.S. consumer pays
4 different streaming video subscriptions
(Deloitte, 2021)

87% of the marketing professionals use video as a marketing tool
(Wyzowl, 2019)

https://cedcommerce.com/video-marketing

How can we make this exponentially growing video content easier to consume?

Query-Aware Multi-Video Summarization (QAMVS)

Query: Prince William Wedding

Retrieved Videos

Video 1

Video 2

Video 3

...

Video N

Summary: a subset of the input videos frames

Summary Criteria

1. Conciseness
2. Representativeness of query-relevant events
3. Chronological soundness

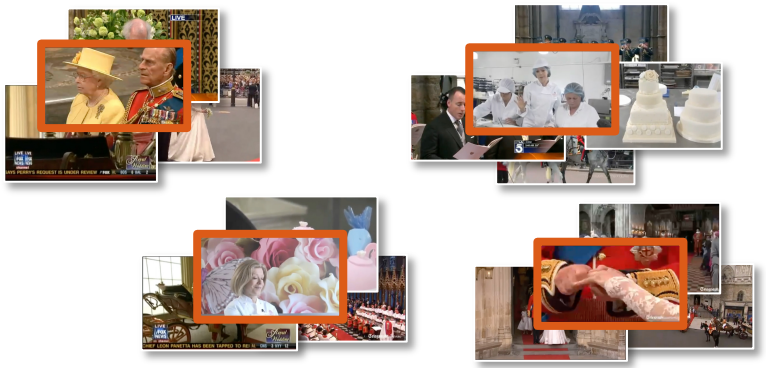
QAMVS Models in the Literature

Multi-staged pipelines optimizing for summary criteria **sequentially**

Conciseness → **Query-relevance** → **Chronological soundness**

Select diverse frames

Sparse Coding (*e.g.* QUASC)
Clustering (*e.g.* HDS)

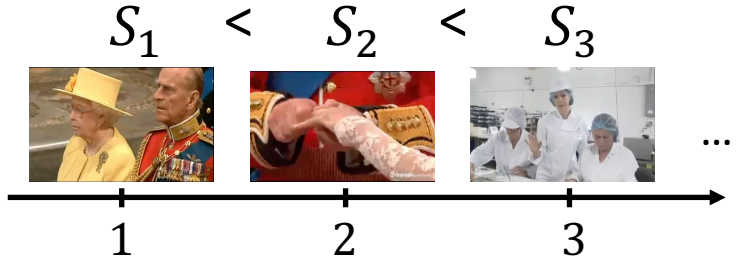


Discard frames dissimilar to web-images retrieved with the same query

Textual Query ⟷ **Image space** (Web-images)

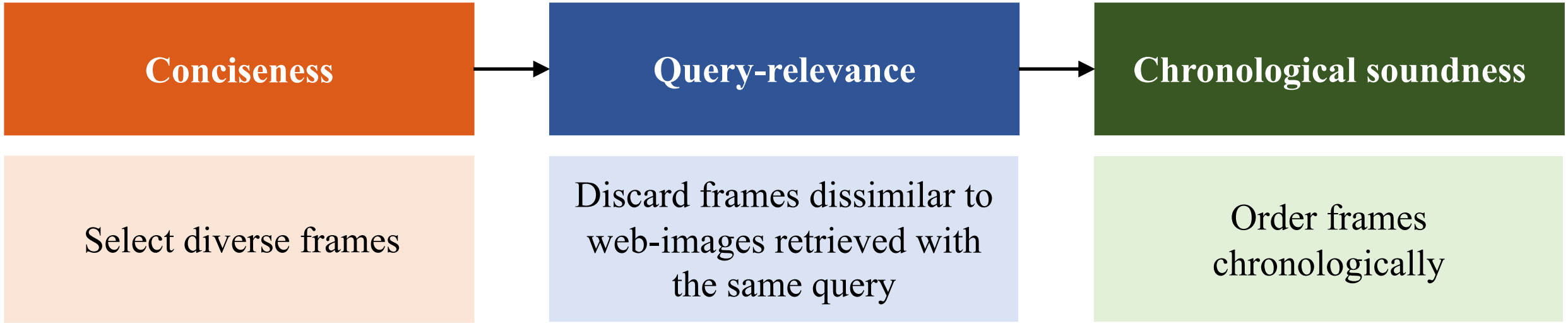
Order frames chronologically

Scores based on topic closeness/videos time tags



QAMVS Models in the Literature

Multi-staged pipelines optimizing for summary criteria **sequentially**



Limitations:

× Error propagation through sequential stages
× Polynomial complexity w.r.t. number of input frames

Can we learn efficient end-to-end trainable models for QAMVS?

× Scarcity of annotated data
× Subjectivity of ground-truth summaries

Reinforcement Learning for DeepQAMVS

Summary s_0 = [ ]

Agent (Policy Network)

F_1 F_2 F_3 F_1 F_2 F_3 F_1 F_2 F_3

Video 1 | Video 2 | Video 3

Videos ($\mathcal{V}$)

Web-Images ($\mathcal{I}$)

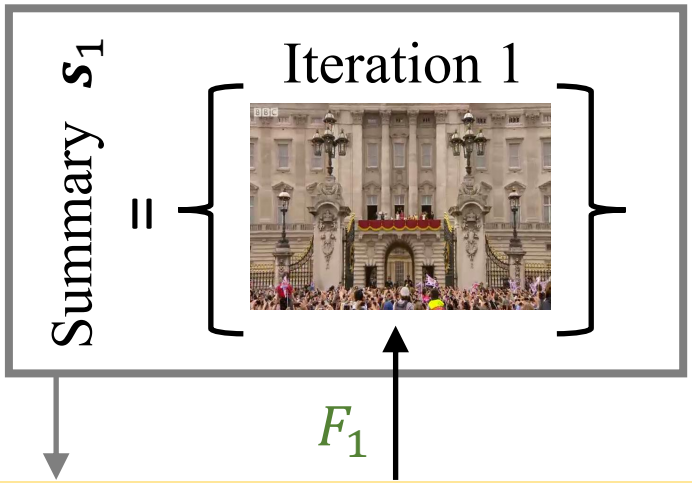
Query: Prince William Wedding

Videos' Titles & Descriptions

1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Reinforcement Learning for DeepQAMVS



Agent (Policy Network)

Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Reinforcement Learning for DeepQAMVS

Agent (Policy Network)

Video 1 | Video 2 | Video 3

Videos ($\mathcal{V}$)

Web-Images ($\mathcal{I}$)

Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Reinforcement Learning for DeepQAMVS

Summary s_3 = { Iteration 1  , Iteration 2  , Iteration 3  }

F_3

Agent (Policy Network)

F_1 F_2 F_3 | F_1 F_2 F_3 | F_1 F_2 F_3

Video 1 | Video 2 | Video 3

Videos ($\mathcal{V}$)

Web-Images ($\mathcal{I}$)

Query: Prince William Wedding

Videos' Titles & Descriptions

1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Reinforcement Learning for DeepQAMVS

Summary s_4 = { Iteration 1  , Iteration 2  , Iteration 3  , Iteration 4  }

State s_t
Summary at iteration t

F_3

Agent (Policy Network)

Video 1 | Video 2 | Video 3

Videos ($\mathcal{V}$)

Web-Images ($\mathcal{I}$)

Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Reinforcement Learning for DeepQAMVS

Summary s_4 = { Iteration 1  , Iteration 2  , Iteration 3  , Iteration 4  }

$a_4 = \{F_3\}$

Agent (Policy Network)

State s_t
Summary at iteration t

Action a_t
Select a new frame from the video collection

Video 1 | Video 2 | Video 3

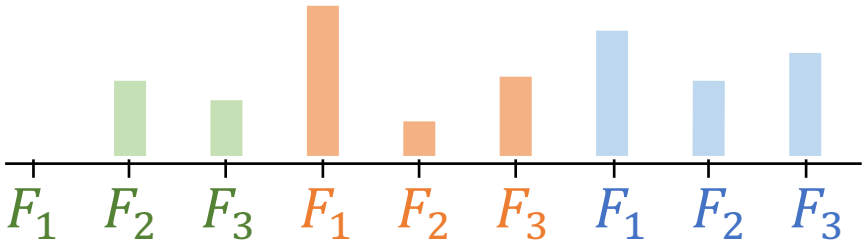
Videos ($\mathcal{V}$)

Web-Images ($\mathcal{I}$)

Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Policy Network: Pointer Network



Video Frame Distribution

$\pi_{\boldsymbol{\theta}}(a_2|\boldsymbol{s}_1, \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{D}})$

Summary = $\boldsymbol{s}_1$

F_1

Policy Network

F_1 F_2 F_3

Video 1

F_1 F_2 F_3

Video 2

F_1 F_2 F_3

Video 3

Videos ($\boldsymbol{\mathcal{V}}$)

Web-Images ($\boldsymbol{\mathcal{I}}$)

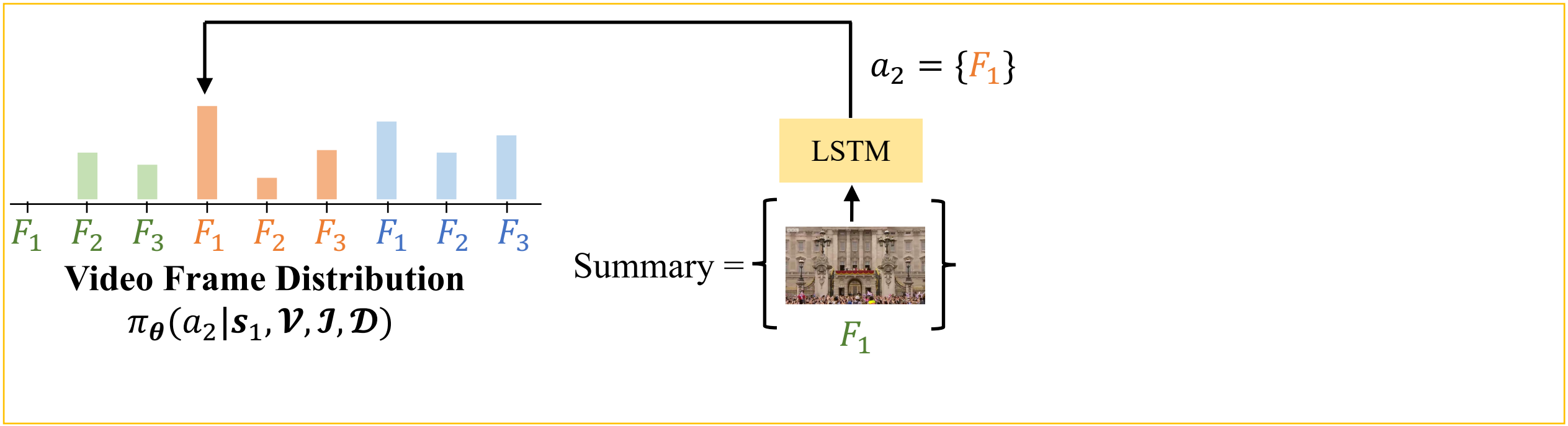
Query: Prince William Wedding

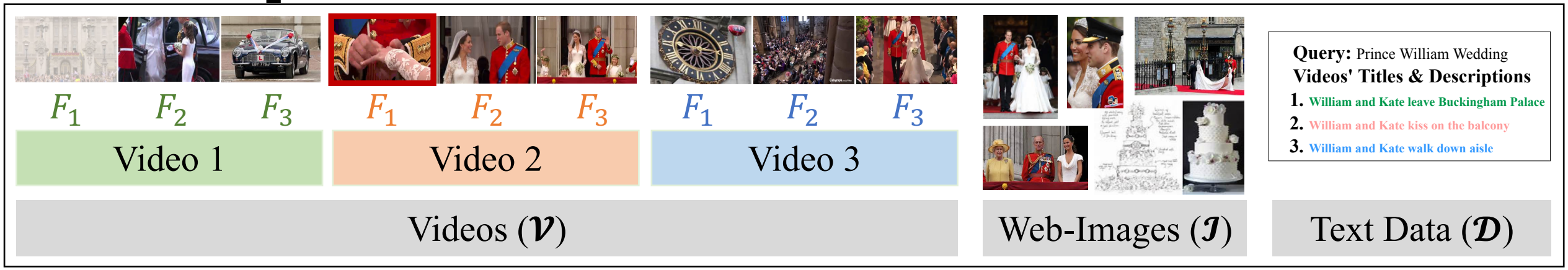
Videos' Titles & Descriptions

1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

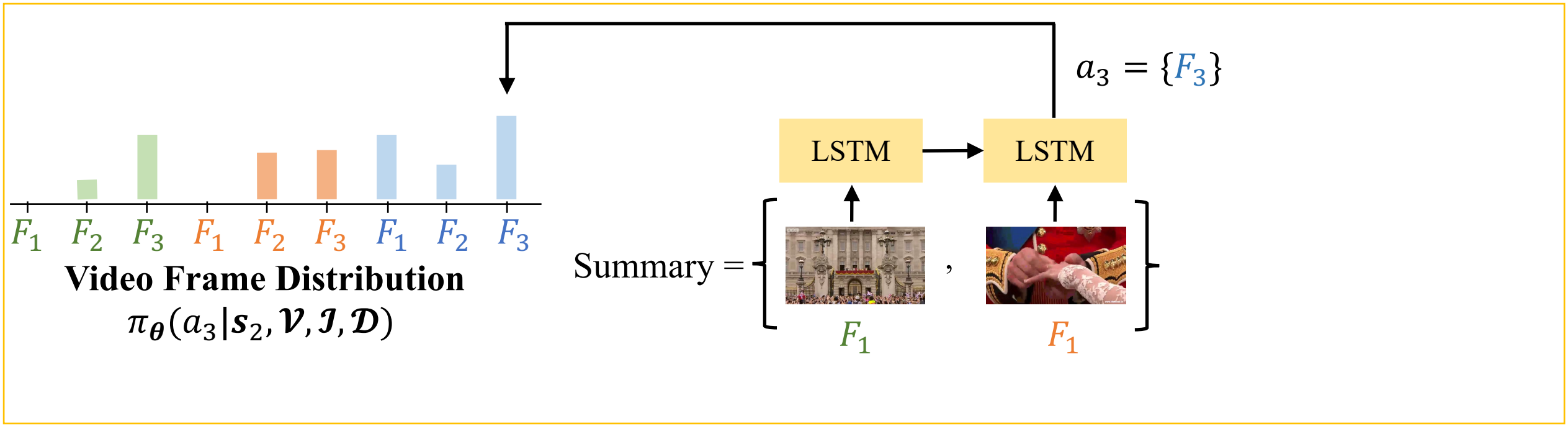
Text Data ($\boldsymbol{\mathcal{D}}$)

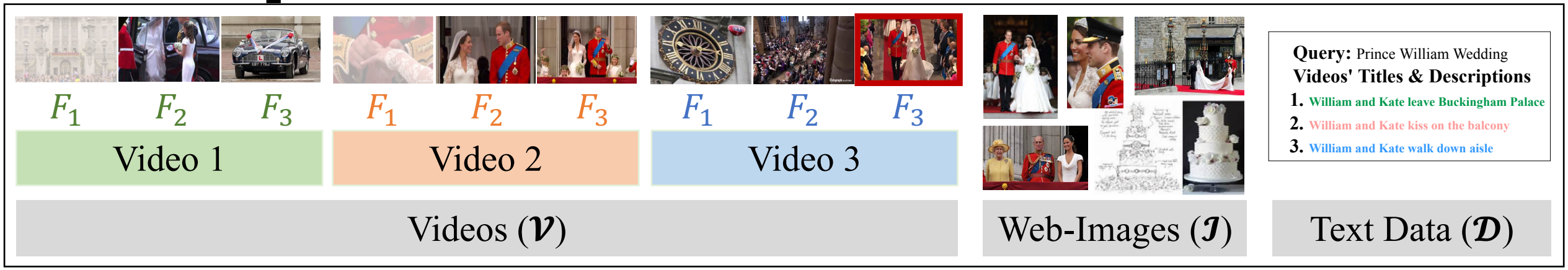
Policy Network: Pointer Network





Policy Network: Pointer Network





Policy Network: Pointer Network

$a_4 = \{F_3\}$

LSTM → LSTM → LSTM

F_1 F_2 F_3 F_1 F_2 F_3 F_1 F_2 F_3

Video Frame Distribution

$\pi_{\theta}(a_4|s_3, \mathcal{V}, \mathcal{I}, \mathcal{D})$

Summary = {  ,  ,  }

F_1 F_1 F_3

Pointer networks learn an order over the frames

F_1 F_2 F_3 F_1 F_2 F_3 F_1 F_2 F_3

Video 1 | Video 2 | Video 3

Videos ($\mathcal{V}$)

Web-Images ($\mathcal{I}$)

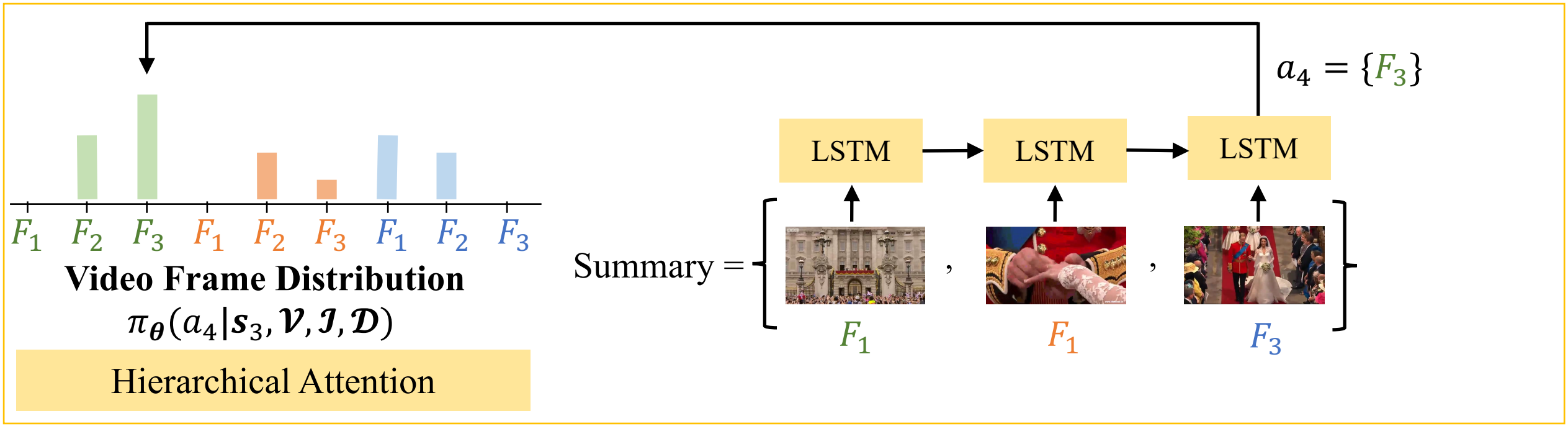
Query: Prince William Wedding

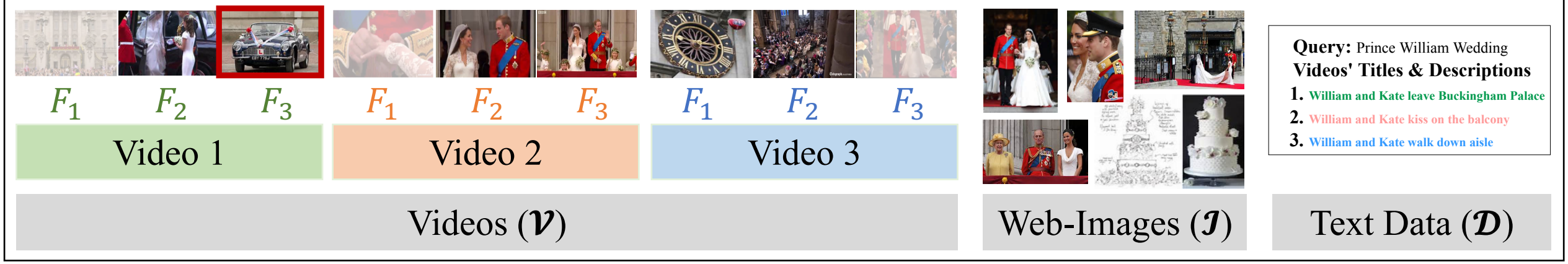
Videos' Titles & Descriptions

1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

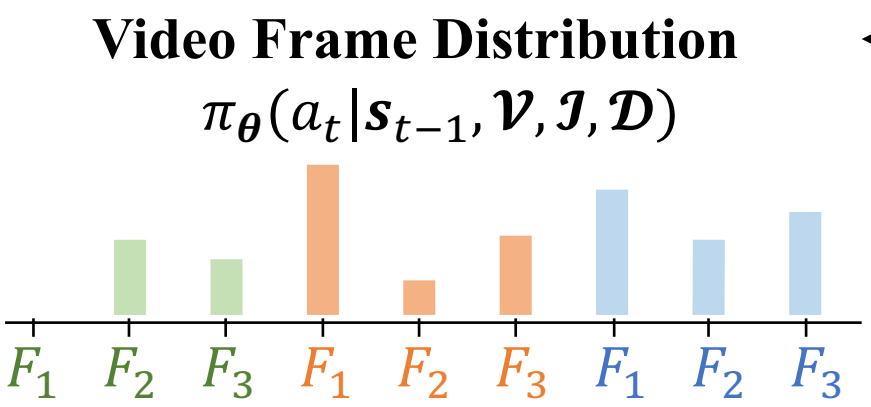
Text Data ($\mathcal{D}$)

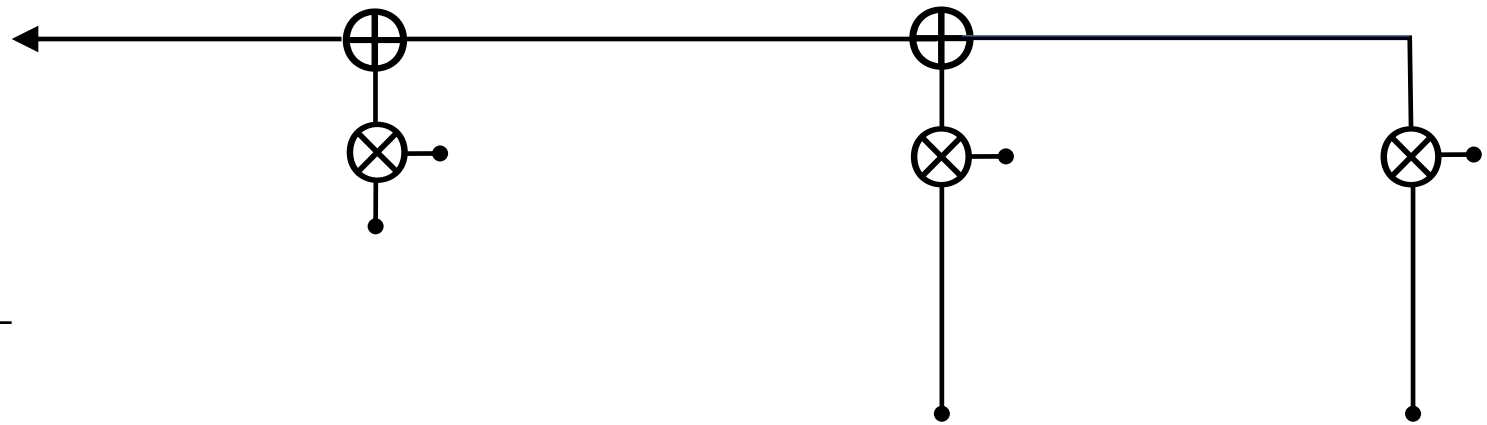
Policy Network: Pointer Network





Policy Network: Hierarchical Attention

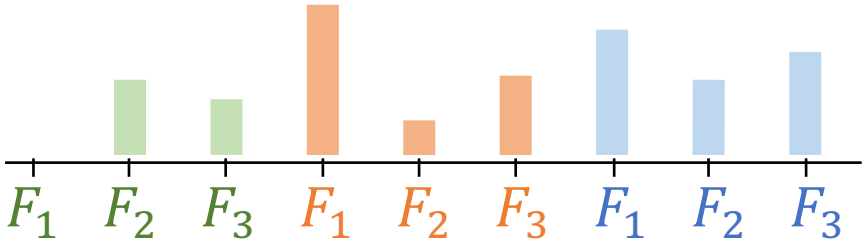


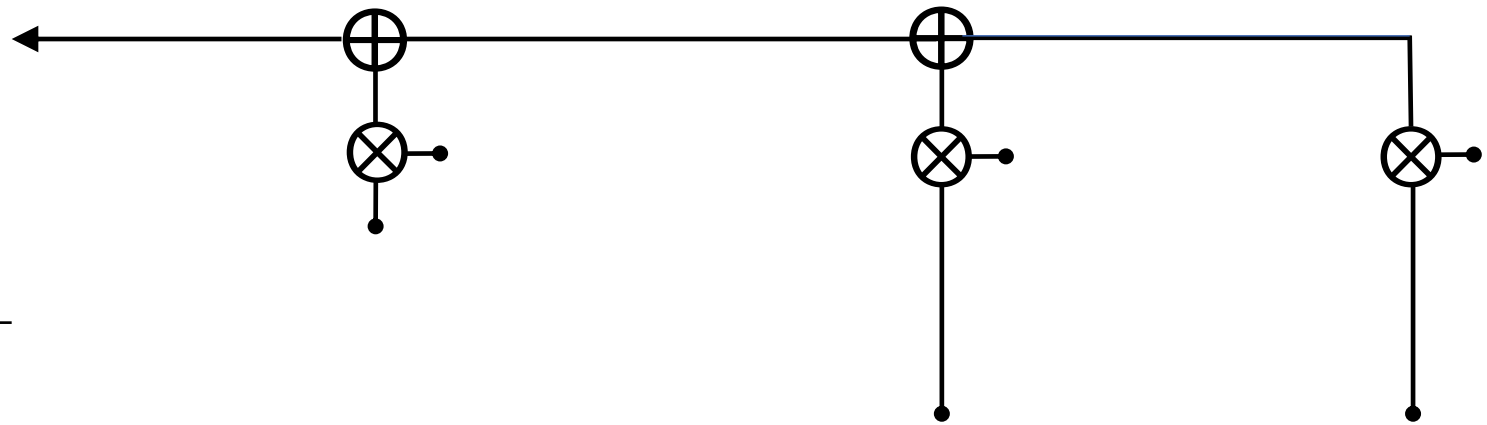


Policy Network: Hierarchical Attention

Video Frame Distribution

$\pi_{\theta}(a_t|s_{t-1}, \mathcal{V}, \mathcal{I}, \mathcal{D})$





$\pi_{\theta}(a_t|s_{t-1}, \mathcal{D})$

Query Attention

Query: Prince William Wedding

Videos' Titles & Descriptions

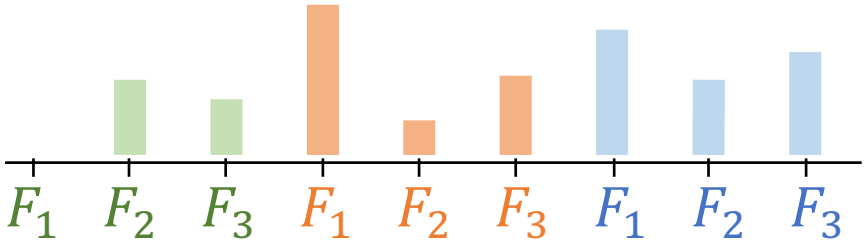
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

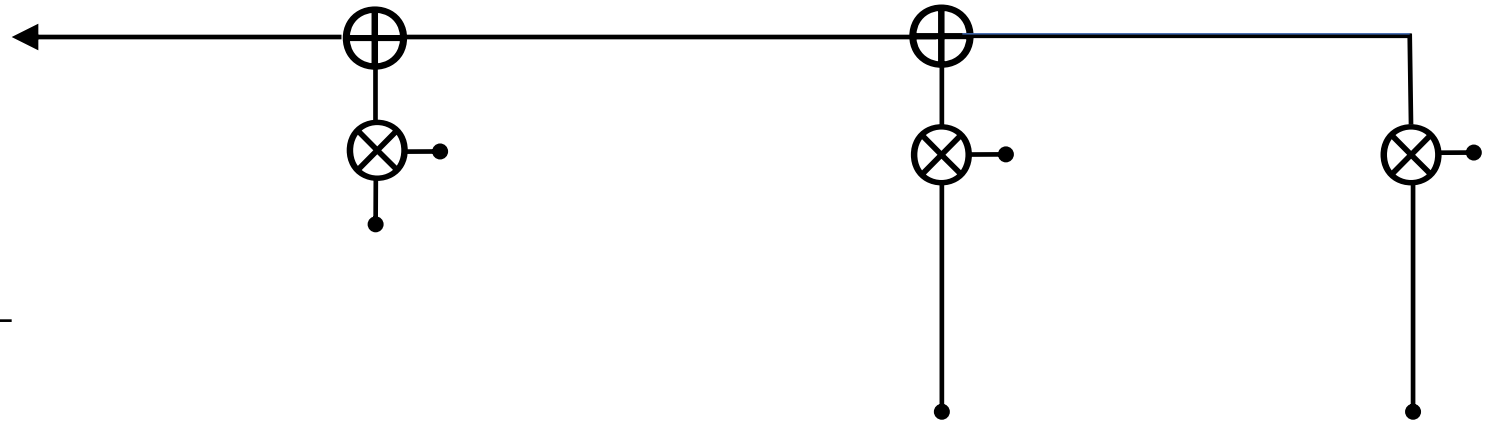
Text Data ($\mathcal{D}$)

Policy Network: Hierarchical Attention

Video Frame Distribution

$\pi_{\theta}(a_t|s_{t-1}, \mathcal{V}, \mathcal{I}, \mathcal{D})$





$\pi_{\theta}(a_t|s_{t-1}, \mathcal{I})$

Image Attention

$\pi_{\theta}(a_t|s_{t-1}, \mathcal{D})$

Query Attention

Web-Images ($\mathcal{I}$)

Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Policy Network: Hierarchical Attention

Video Frame Distribution

$\pi_{\theta}(a_t|s_{t-1}, \mathcal{V}, \mathcal{I}, \mathcal{D})$

F_1 F_2 F_3 F_1 F_2 F_3 F_1 F_2 F_3

μ_1 μ_2 μ_3

$\pi_{\theta}(a_t|s_{t-1}, \mathcal{V})$

Video Frame Attention

$\pi_{\theta}(a_t|s_{t-1}, \mathcal{I})$

Image Attention

$\pi_{\theta}(a_t|s_{t-1}, \mathcal{D})$

Query Attention

Video Attention

F_1 F_2 F_3 | F_1 F_2 F_3 | F_1 F_2 F_3

Frame Attention | Frame Attention | Frame Attention



Video 1 | Video 2 | Video 3

Web-Images ($\mathcal{I}$)

Query: Prince William Wedding

Videos' Titles & Descriptions

1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle

Text Data ($\mathcal{D}$)

Rewards

1. **Diversity Reward (R_{div}):** Selected frames are diverse

R_{div} {  ,  ,  } > R_{div} {  ,  ,  }

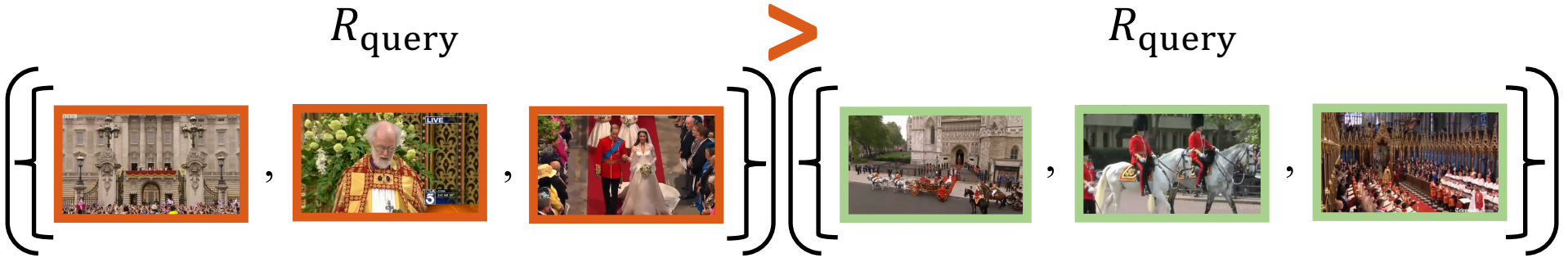
Rewards

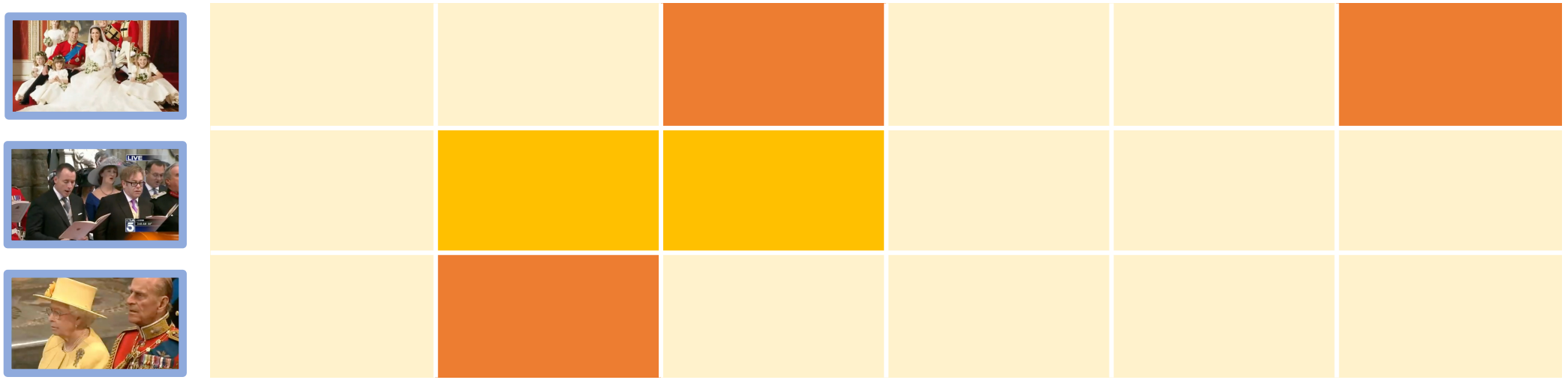
1. **Diversity Reward (R_{div}):** Selected frames are diverse
2. **Representativeness Reward (R_{rep}):** Selected frames are cluster centers of the input frames

$$R_{\text{rep}}\left(\left\{\ ,\ ,\ \right\}\right) > R_{\text{rep}}\left(\left\{\ ,\ ,\ \right\}\right)$$

Rewards

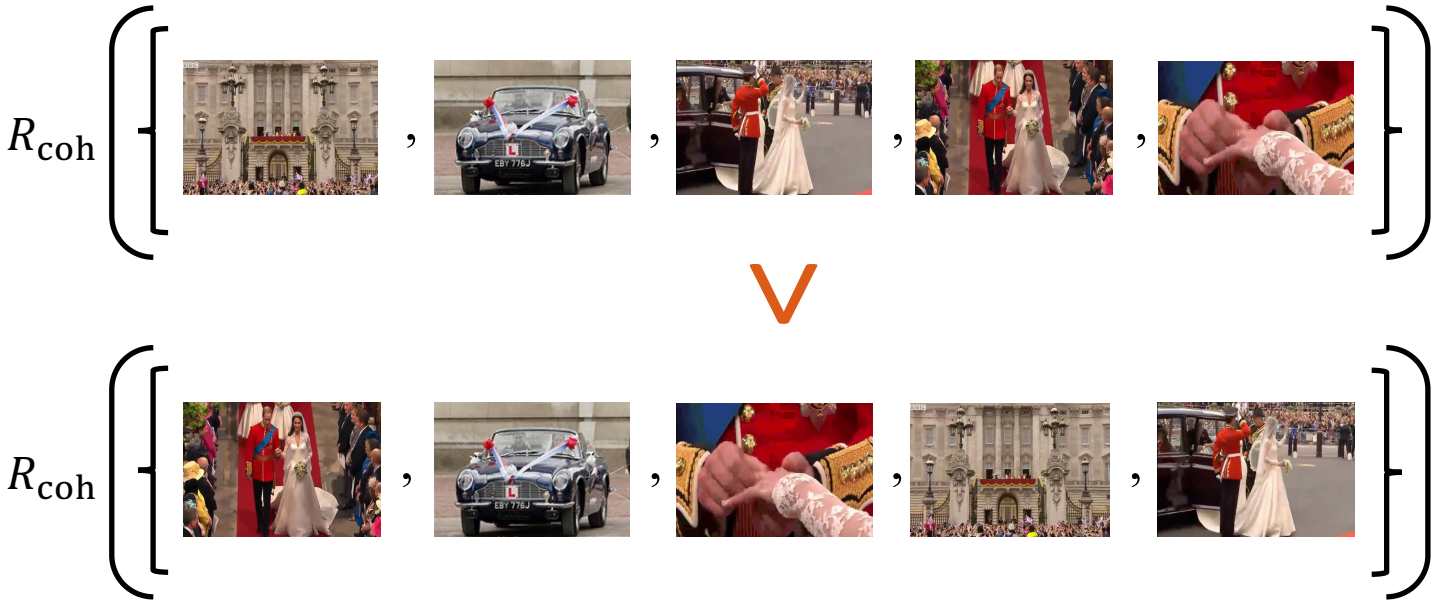
1. **Diversity Reward (R_{div}):** Selected frames are diverse
2. **Representativeness Reward (R_{rep}):** Selected frames are cluster centers of the input frames
3. **Query-adaptability Reward (R_{query}):** Selected frames are similar to retrieved web-images





Rewards

1. **Diversity Reward (R_{div}):** Selected frames are diverse
2. **Representativeness Reward (R_{rep}):** Selected frames are cluster centers of the input frames
3. **Query-adaptability Reward (R_{query}):** Selected frames are similar to retrieved web-images
4. **Temporal Coherence Reward (R_{coh}):** Summaries are visually coherent

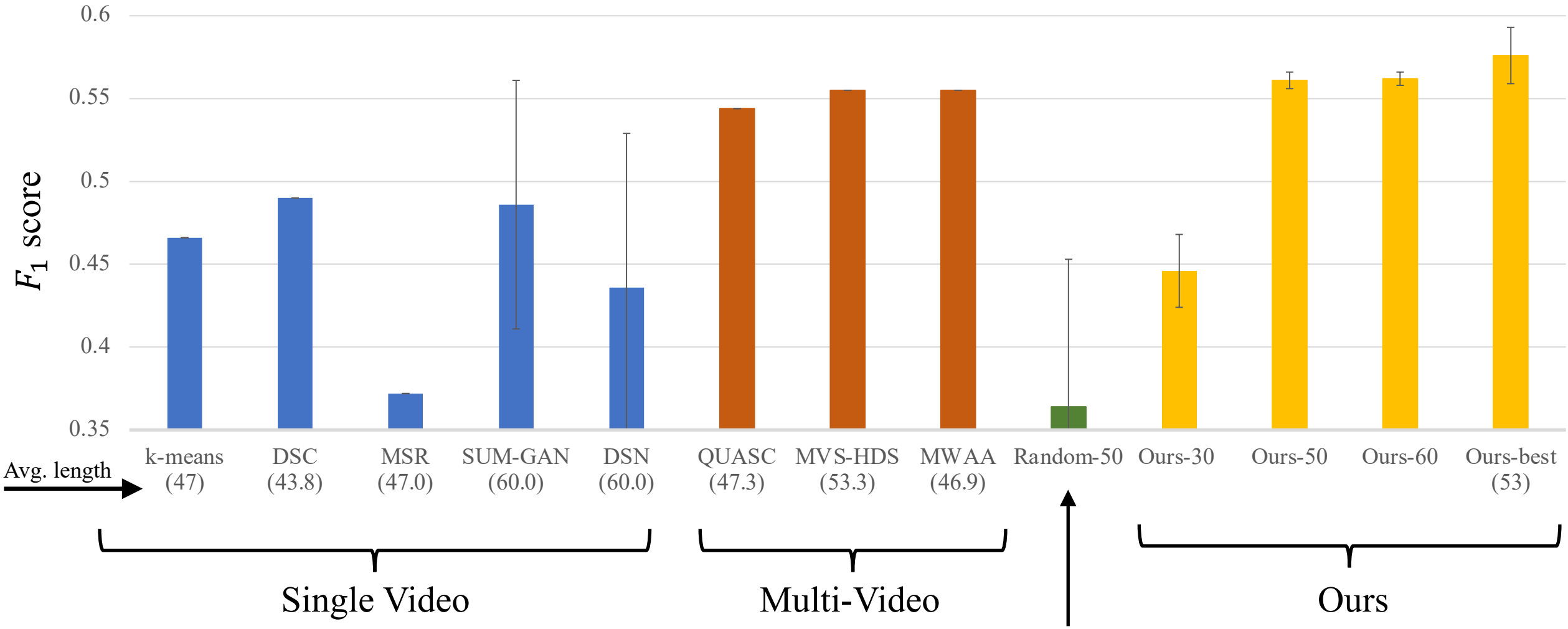


Experiments

Dataset: MVS1K Dataset (4 ground-truth summaries per Query ID)

Query ID	Query	# Videos	# Frames	# Images
1	Britains Prince William wedding 2011	90	1124	324
2	Prince death 2016	104	1549	142
3	NASA discovers Earth-like planet	100	1349	226
4	American government shut-down 2013	82	962	177
5	Malaysia Airline MH370	109	1330	435
6	FIFA corruption scandal 2015	90	785	177
7	Obama re-election 2012	85	1263	207
8	Alpha go vs Lee Sedo	84	976	118
9	Kobe Bryant retirement	109	1140	221
10	Paris terror attacks	83	857	651
Total	-	936	-	2678

Quantitative Results



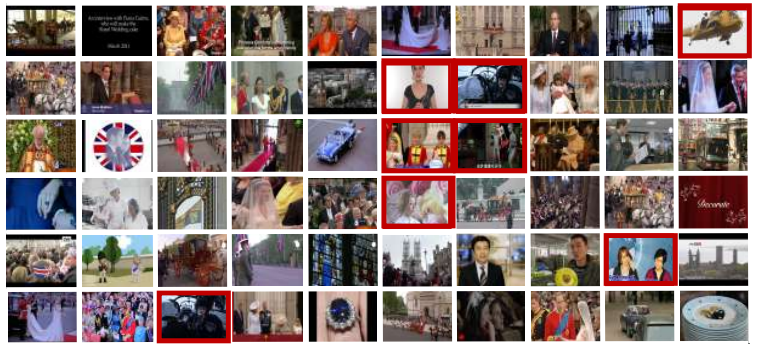
Qualitative Results

K-Means (6|7)

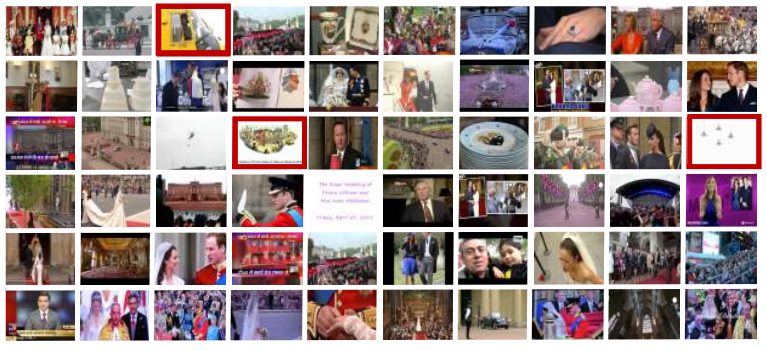
DSC (0|4)

MSR (0|7)

SUM-GAN (0|8)



DSN (0|3)



HDS (2|4)

Redundant

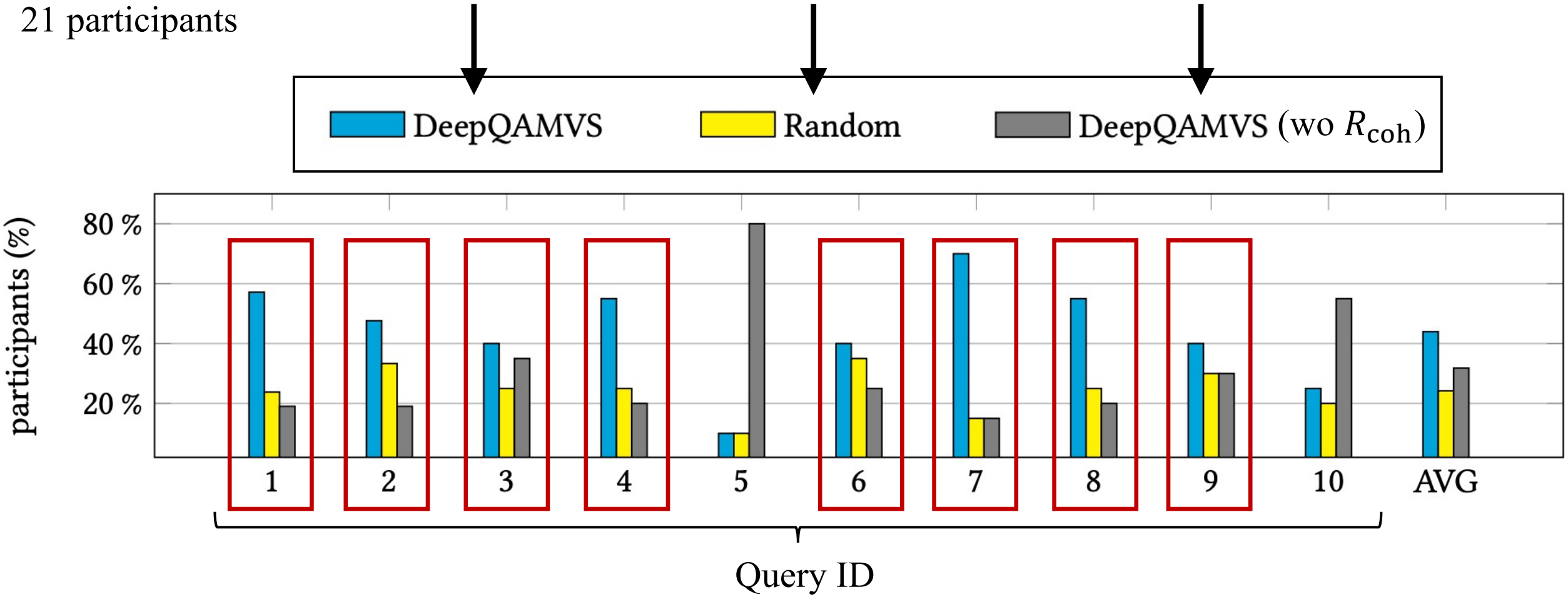
Unimportant

QUASC (0|3)

OURS (0|1)

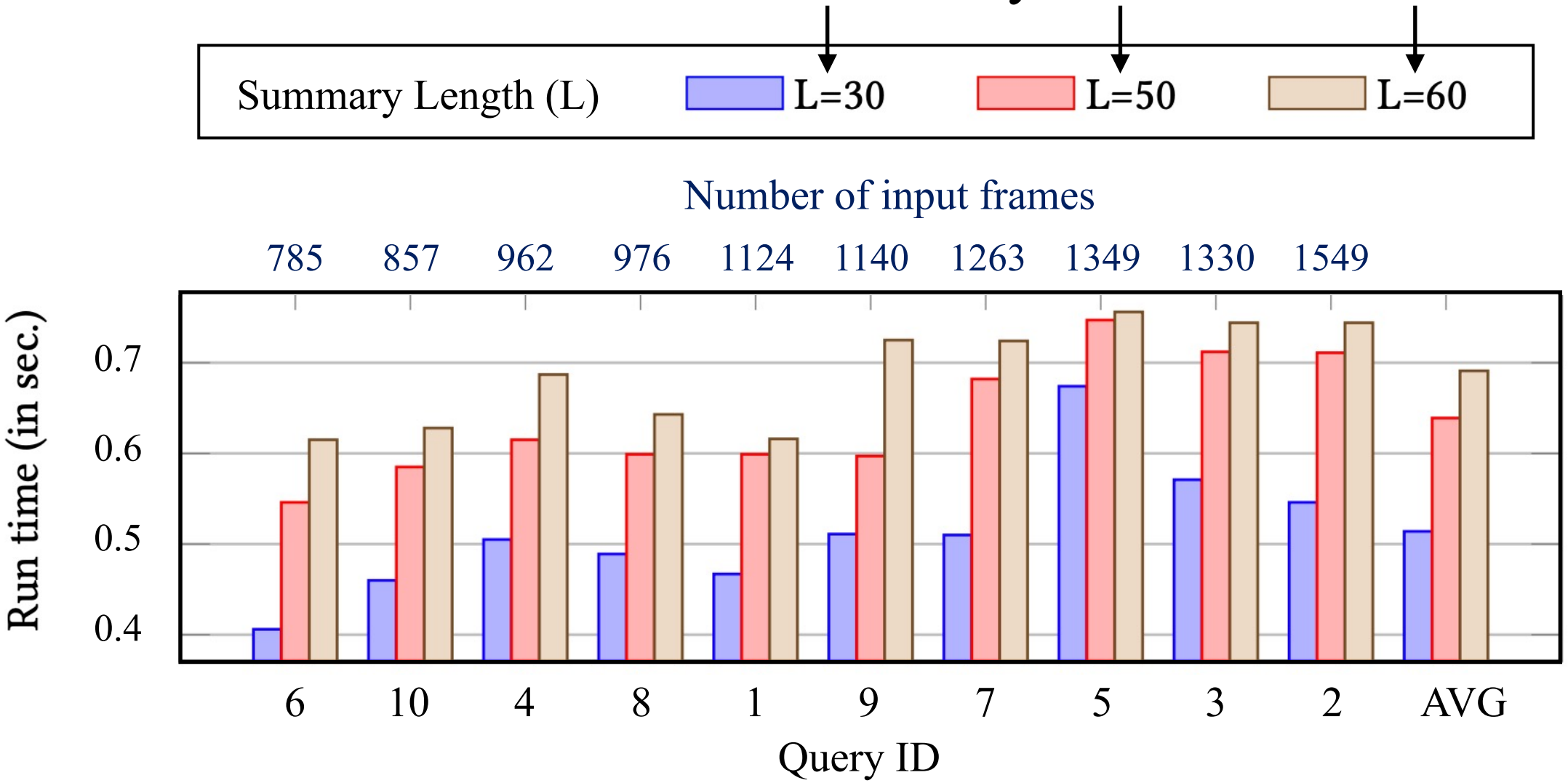
User Study for Temporal Coherence Assessment

21 participants



Users preferred our DeepQAMVS summary in 8 out of 10 queries

Run-Time Analysis



Run-time scales linearly with the number of input frames and summary length

Conclusion

Advantages

✓First end-to-end trainable model for QAMVS

✓SOTA results on MVS1K dataset

✓Scales linearly with the number of input video frames and summary length

Future QAMVS would benefit from:

Better evaluation metric combining visual, textual and temporal order overlap

Better rewards for temporal coherence

New dataset with segment based textual annotation

SCAN ME