Understanding the big data of video with AI

Yan Ye

Machine Intelligence Technology Lab, DAMO Academy

Alibaba Group
阿里巴巴集团

阿里巴巴机器智能技术实验室
Alibaba Machine Intelligence Technology Lab

Big Data of the Alibaba Ecosystem

Big data:	EBs (10^{18}) of video data
Diverse source:	e-commerce, live streaming, entertainment, sports, UGC, etc
Cloud computing:	Tens of millions of servers

OUTLINE

1. The challenges of big data of video
2. AI-powered video understanding
3. AI-powered video fingerprinting and search
4. AI-powered video content production

The challenges of big data of video

Big data of video...

Problem #1: the gap between how video is captured, transmitted, and stored, and how video is consumed

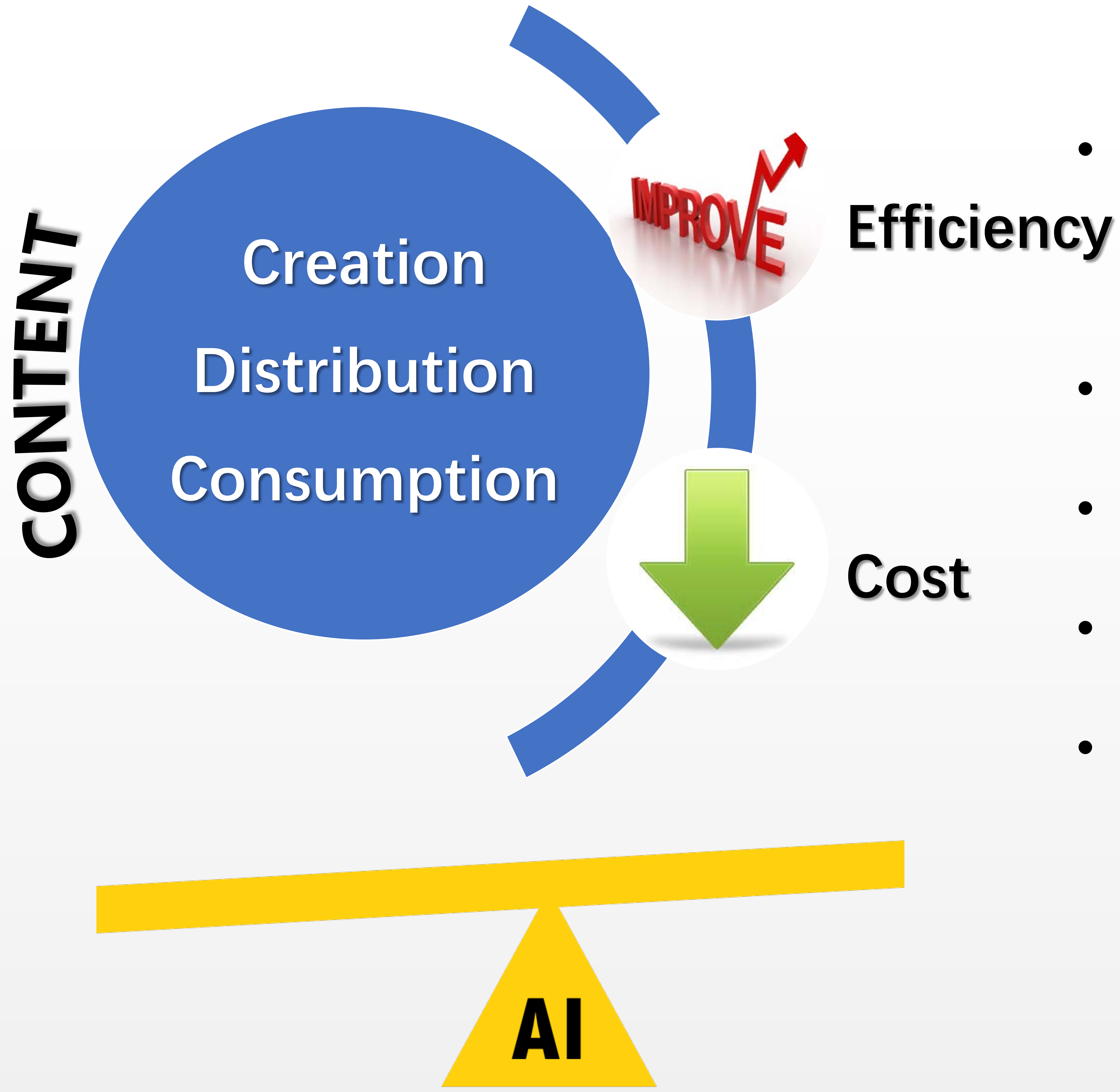
- Video is captured, transmitted and stored as a signal
- However, video is not (just) consumed as a signal, consumption happens at the semantics and emotional levels too
- Need to learn/understand the underlying structure in the video signal

Problem #2: managing the big data of video cost-effectively

- With an ever increasing video content database, need to increase efficiency and reduce cost
- Considering the diversity of the video source and specific applications, efficient content management must be *intelligent*

video classification	video indexing and search	cover image generation	highlight video generation	copyright management	multi-modality

AI Reshapes Video Content Management



AI-powered Video Understanding

- Object and scene recognition: who, what, where
- Action recognition
- Classification
- UGC labeling
- Fingerprinting & copyright

AI-powered Content Generation

- Video summarization
- Personalized cover image generation
- Audio editing
- Sports highlights
- Virtual content

AI-powered Content Distribution

- Recommendations
- Multi-modal search
- Multi-lingual search

AI-powered Video Understanding

Multi-modal video structuring

Multi-modal video understanding

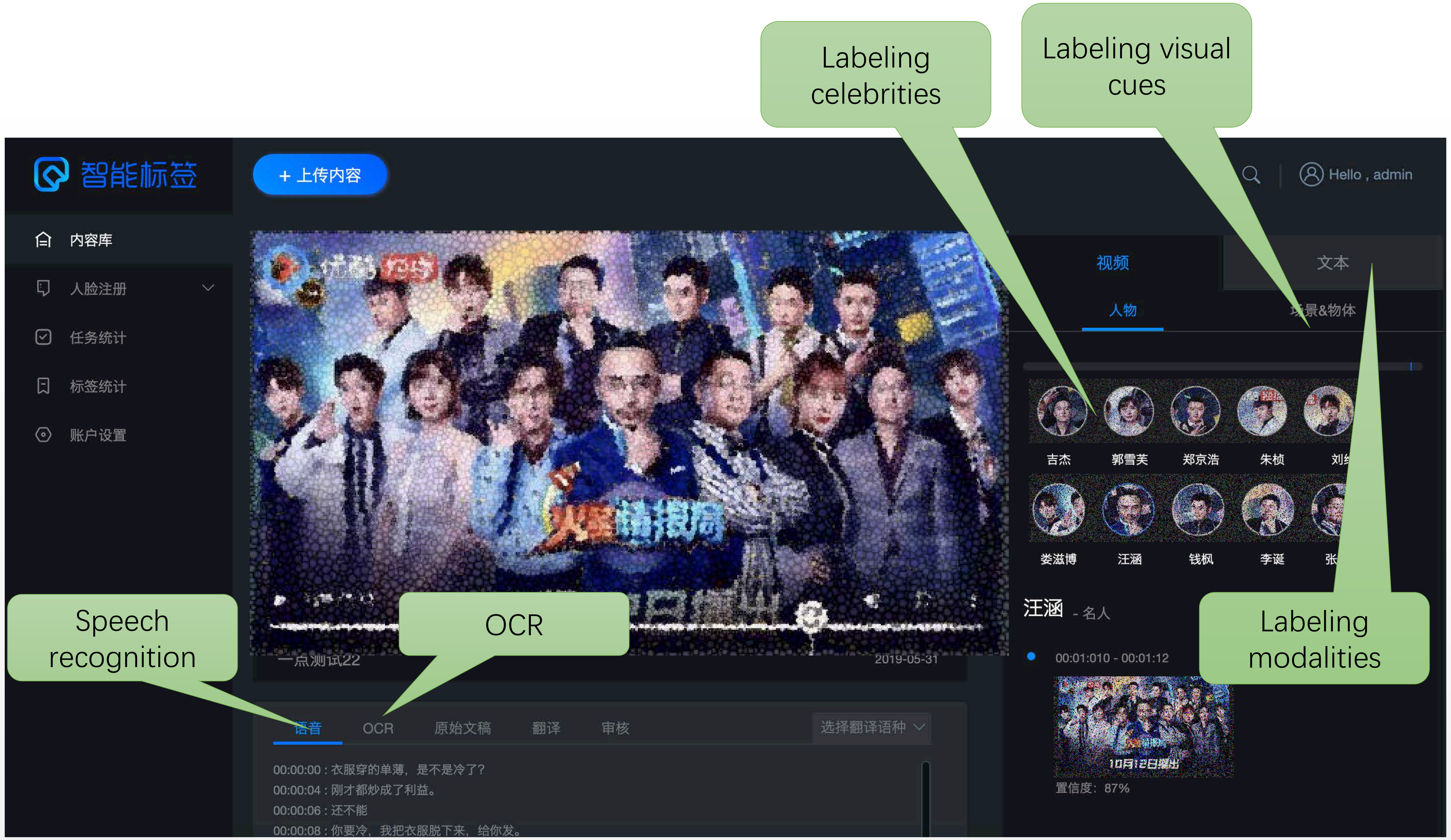
- vision, speech, audio, text

Multi-dimensional structuring

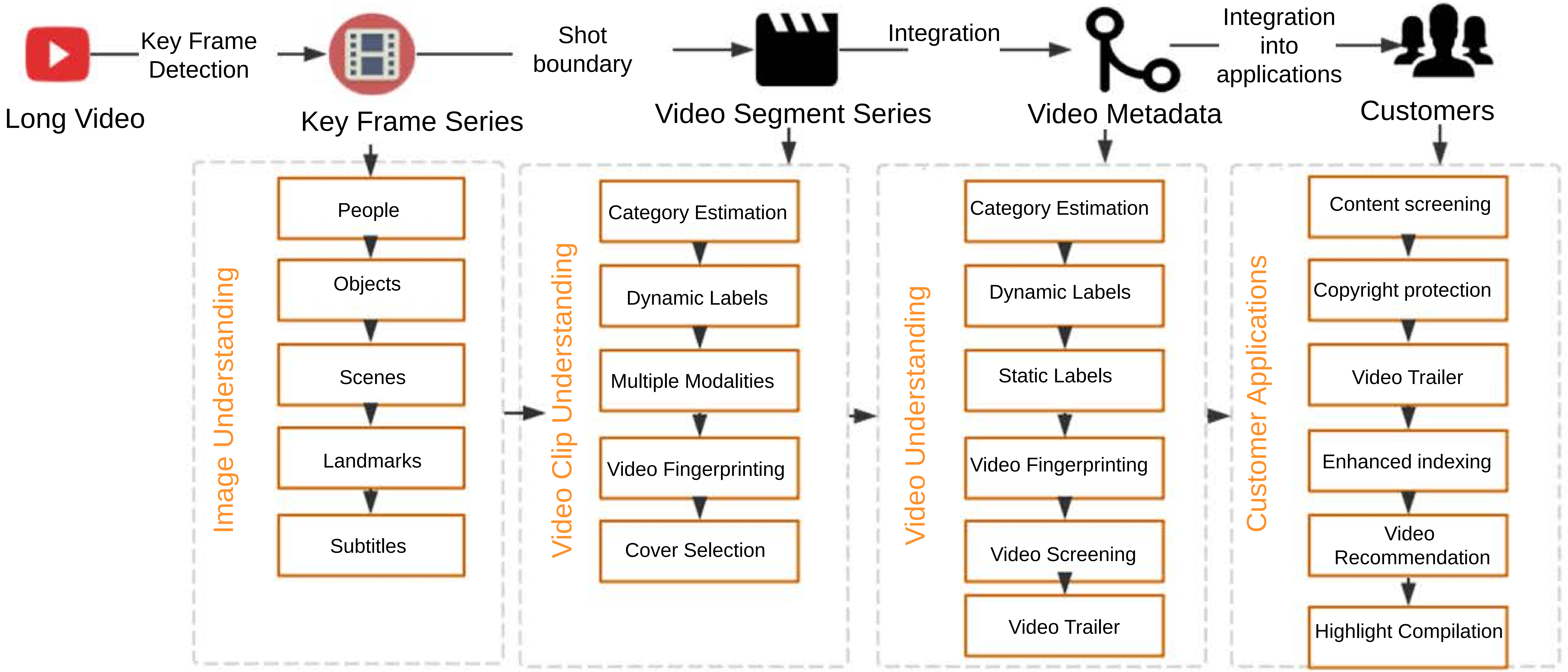
- people, objects, scenes, costume, action, events, emotion

Open-set recognition

- light-weight search engine

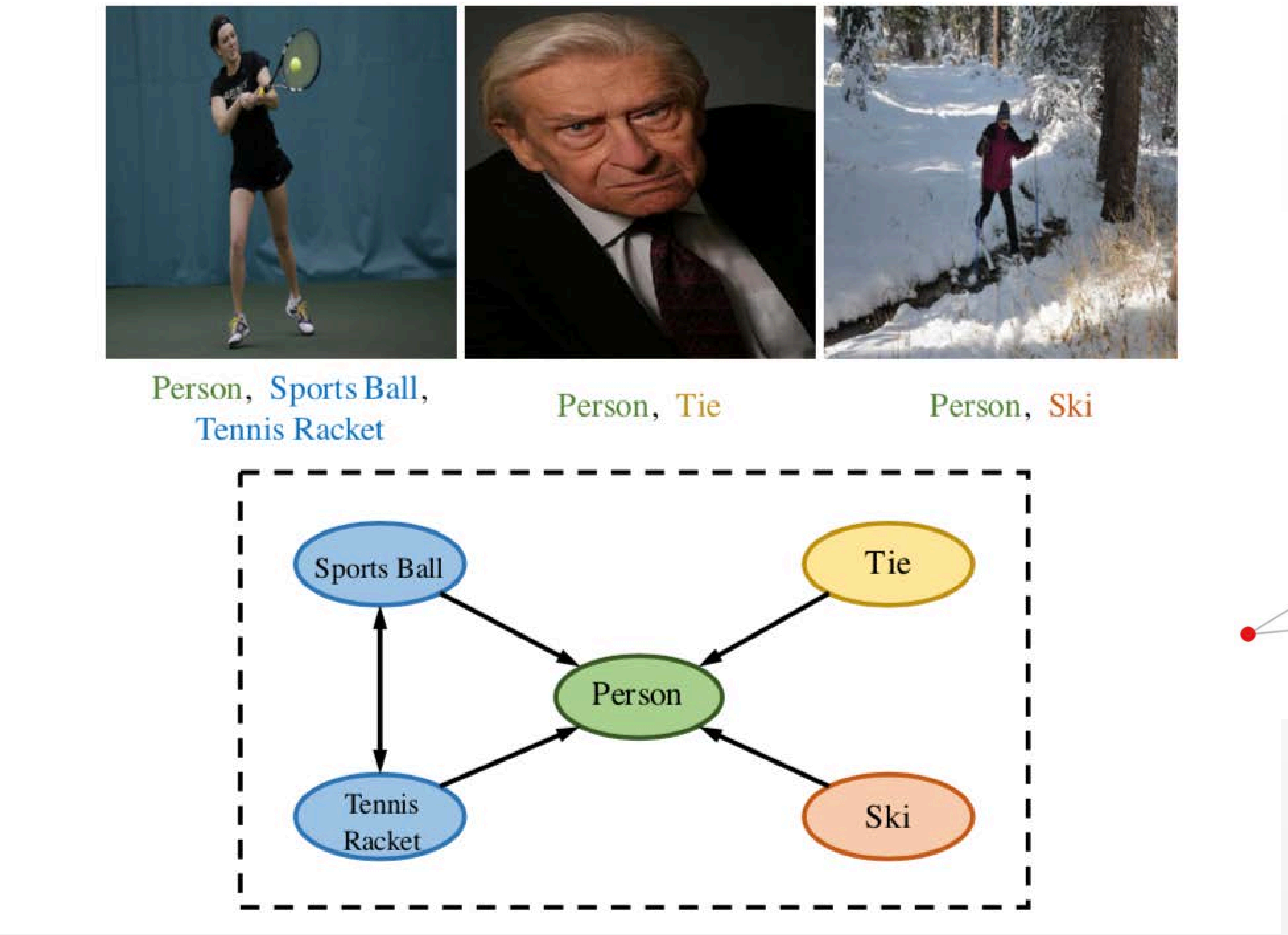


AI-powered multi-level analysis of video content

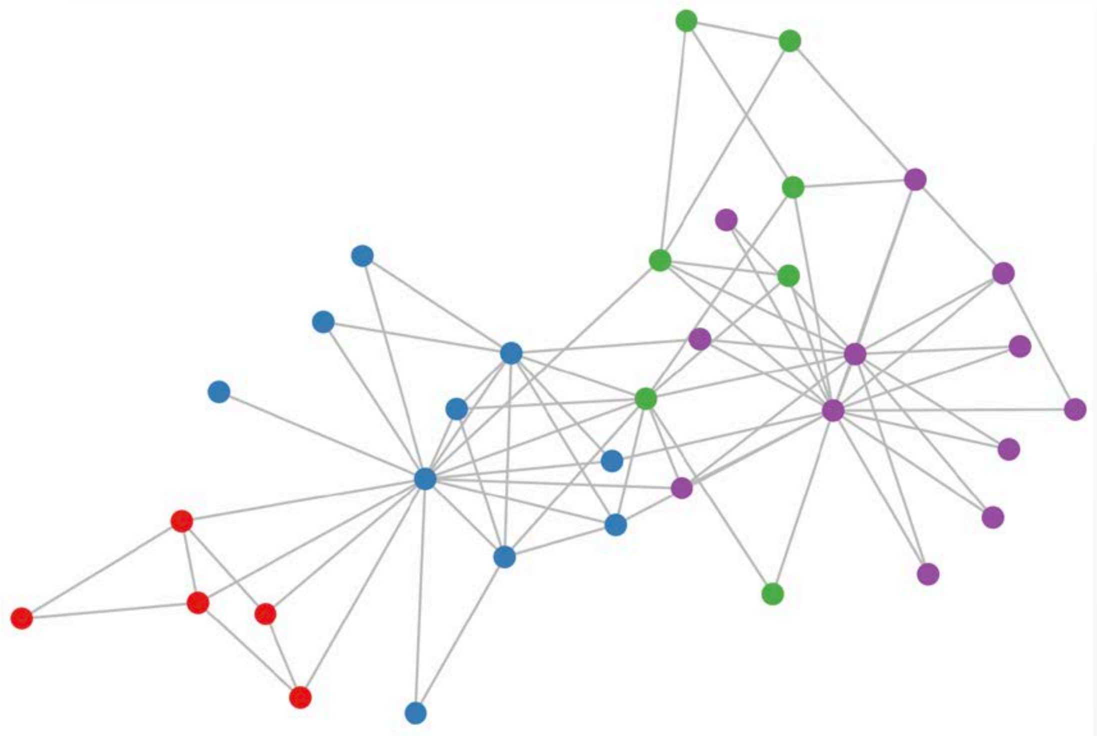


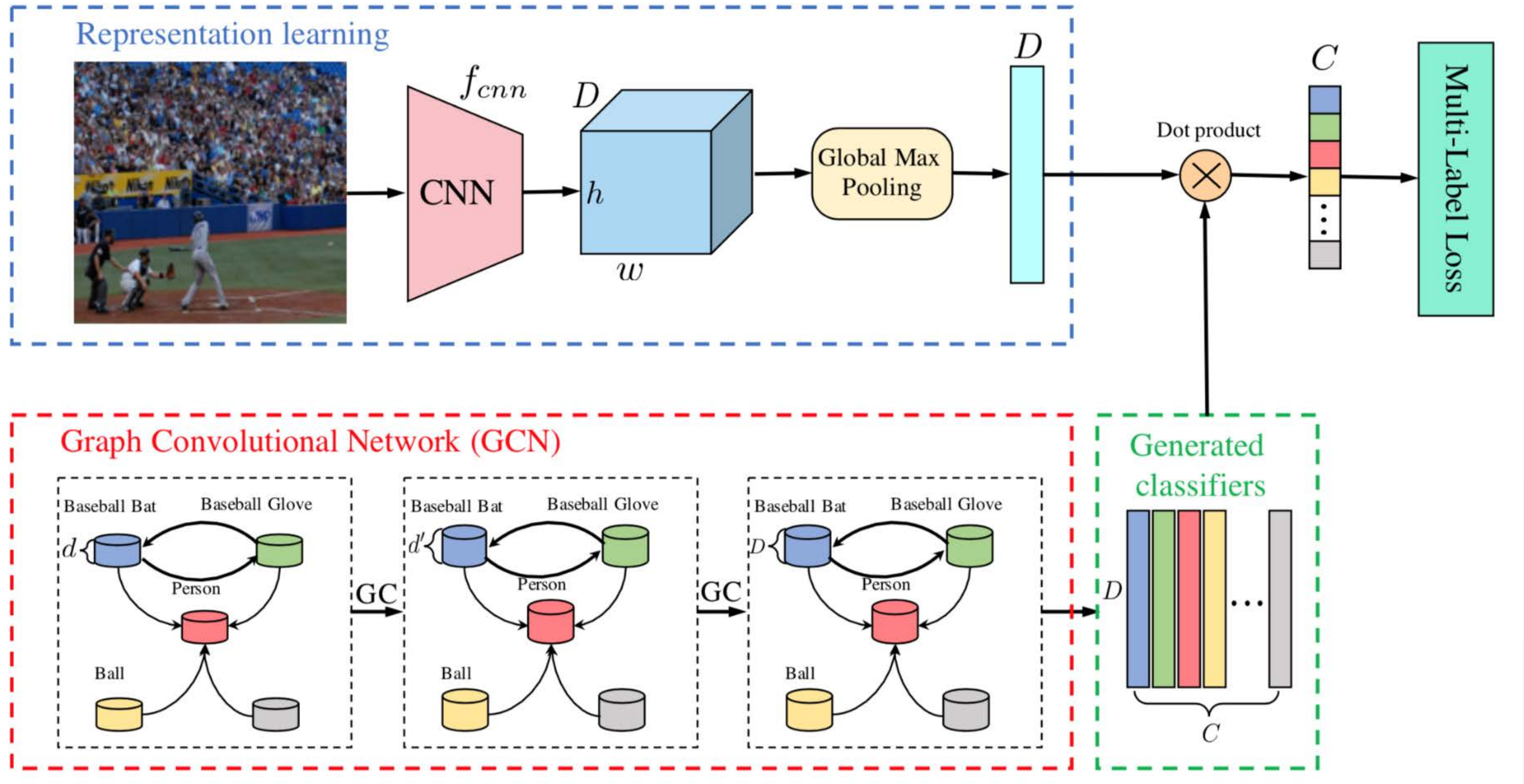
Improving Video Understanding

Label Correlation

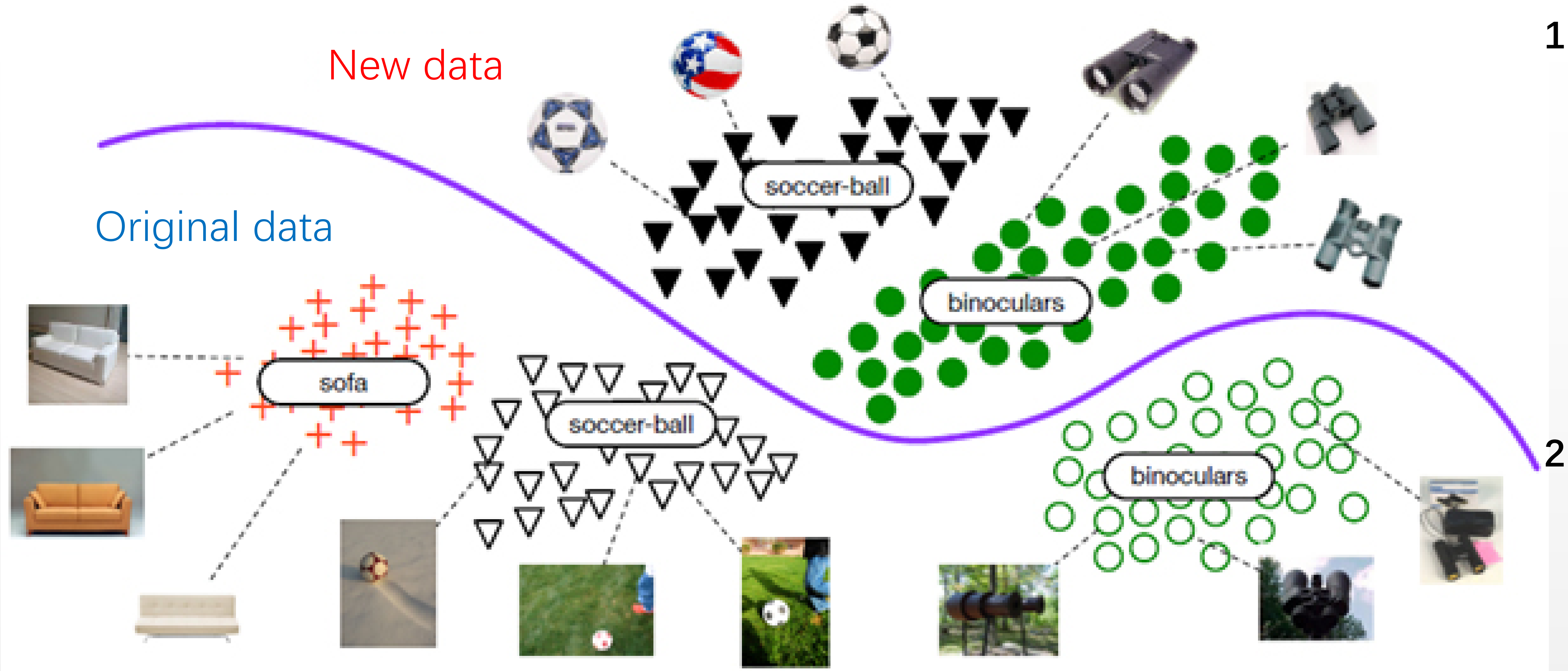


Pairwise relationship modeling



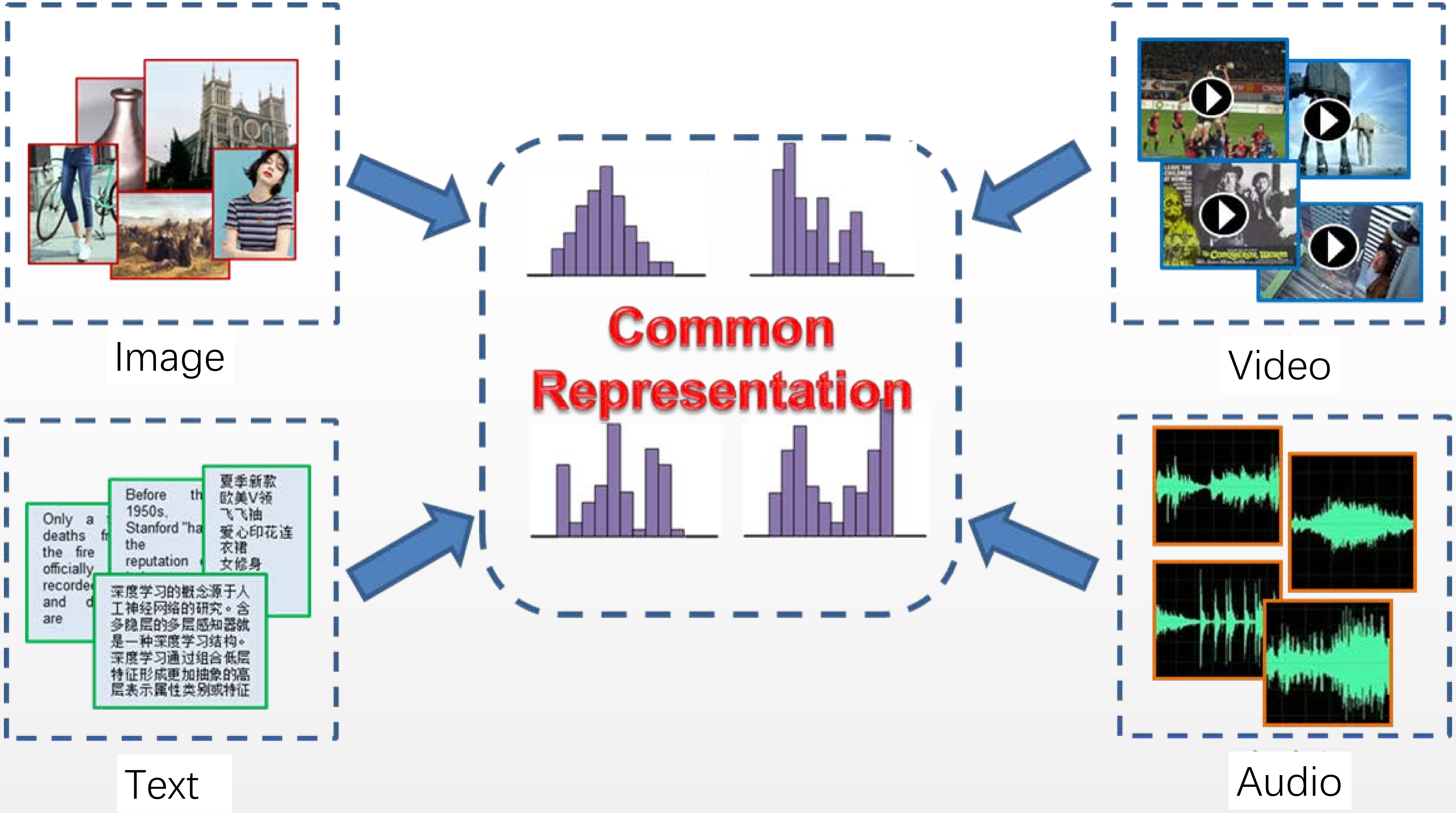


Large-scale classification with incremental learning



1. New data vs. original data: improving performance of the former while keeping the latter the same (no degradation)
2. Fast learning: no need to re-train

Multi-modality indexing and search



AI-powered Video Fingerprinting and Search

Video Fingerprinting and Content Search

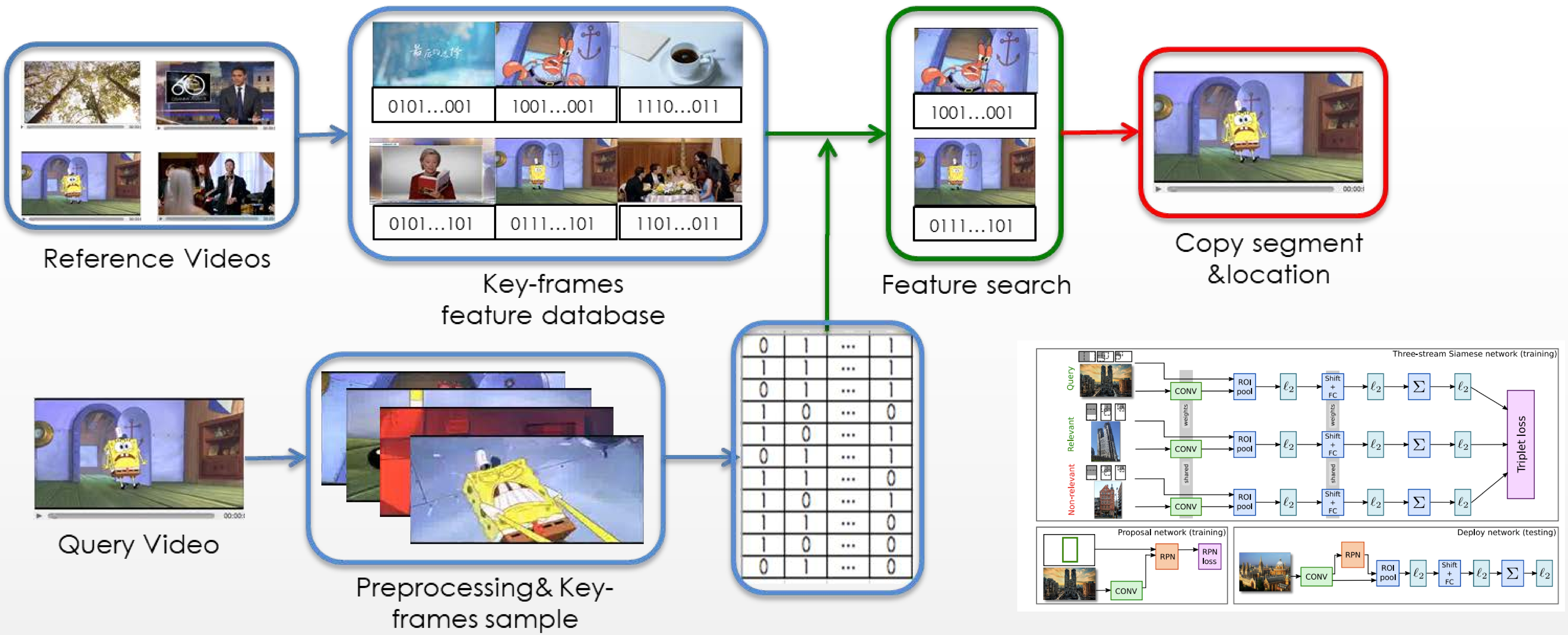
Same Source

Similar Source

Same-source video transformations

1. Quality change: noise, contrast, blur, re-encoding …
2. Spatial transformation: PIP, text insertion, mirroring, aspect ratio, rotation, crop, shift …
3. Editing: timeline
4. Combination of the above

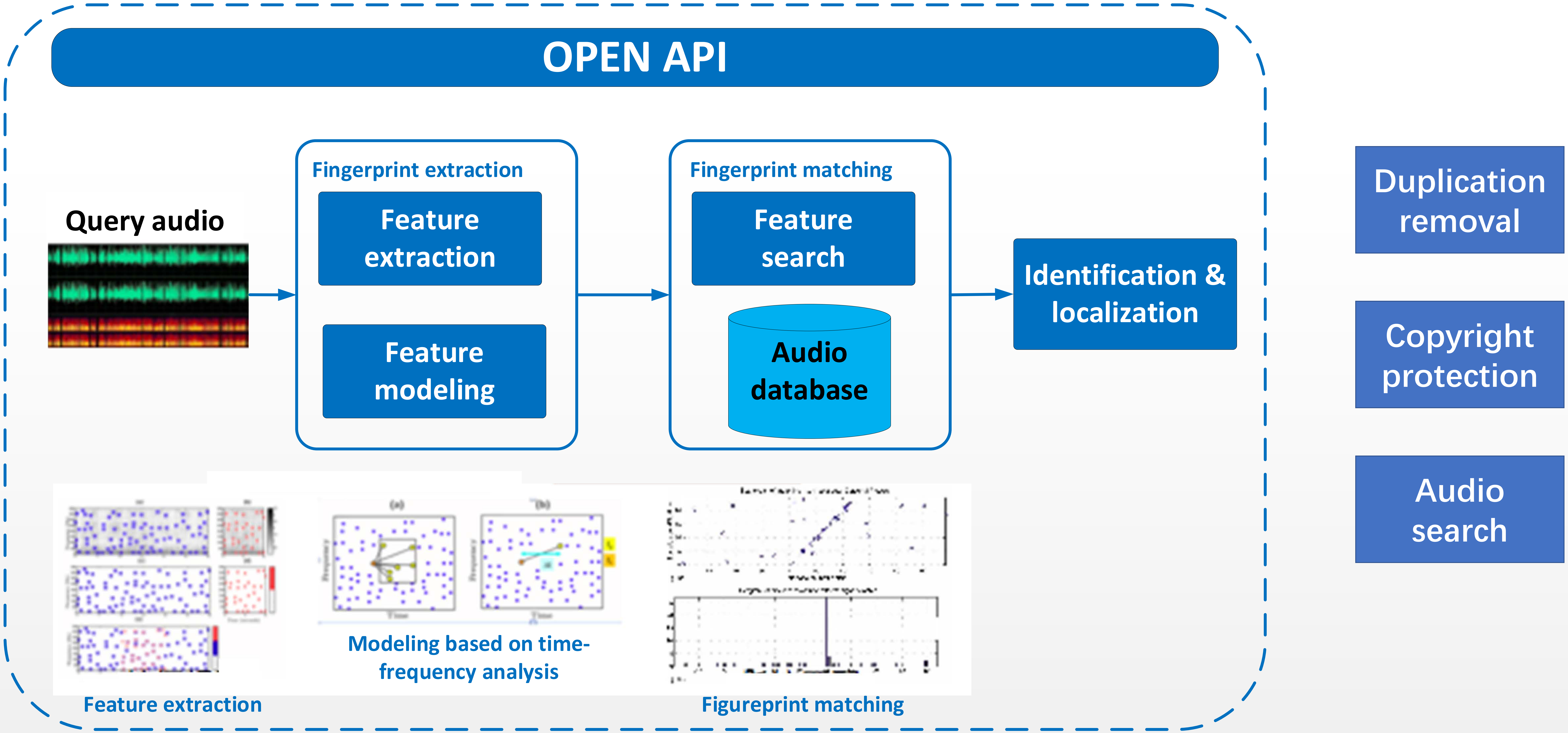
Video Fingerprinting: copyright & search



Training set: 10000*400 same-source transformed video

阿里巴巴机器智能技术实验室
Alibaba Machine Intelligence Technology Lab

Audio Fingerprinting



AI-powered Video Content Production

Cover Image/Video Generation

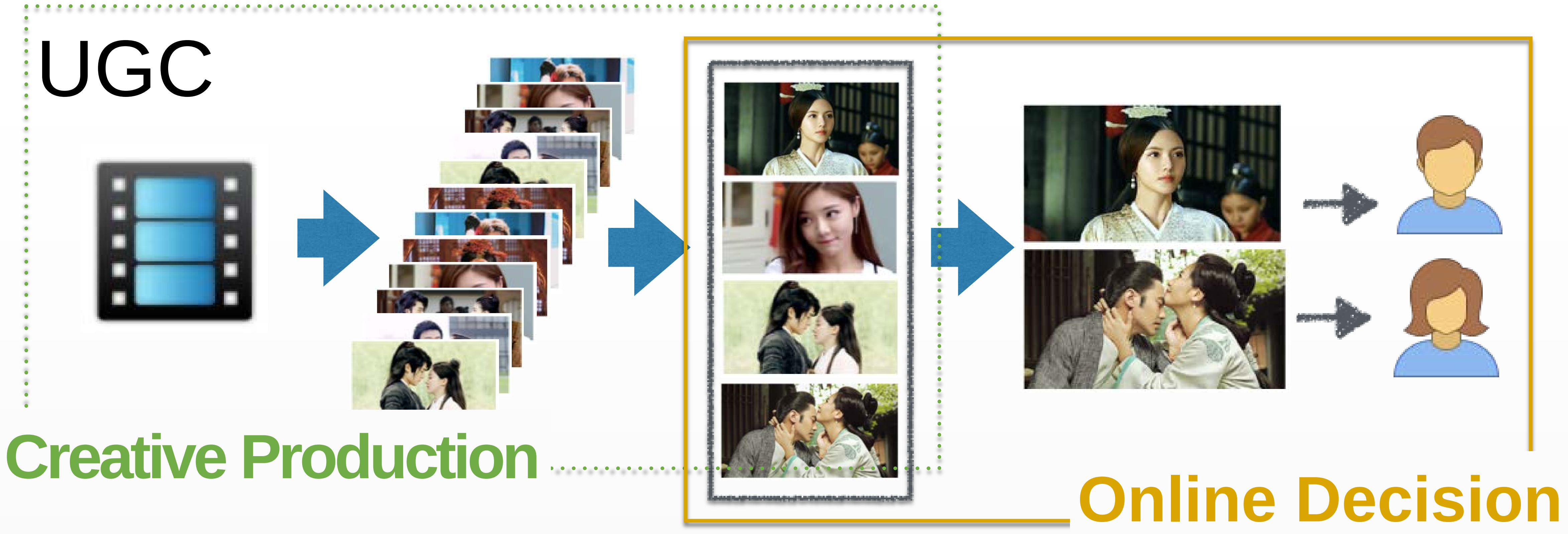
Cover image/video is directly related to user's click-through-rate (CTR)

Problem

- When we have massive amount of video from diverse sources, how do we produce the cover images/video using a general algorithm?
- How to personalize cover image/video?

Solution

- Joint video summarization + online decision optimization with bandwidth cost consideration



Results @ Youku :

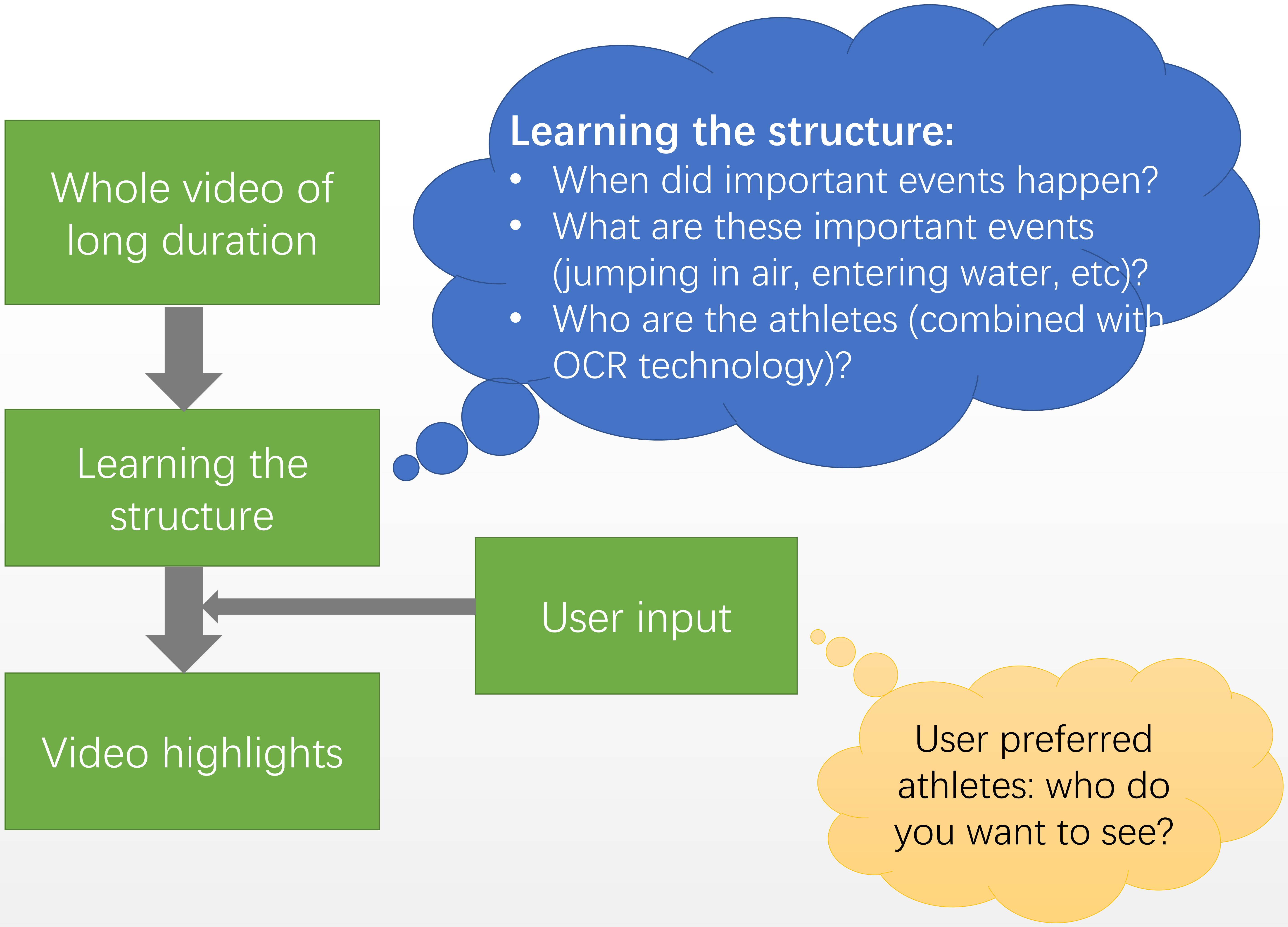
CTR +15%

Dwell-Time +12%

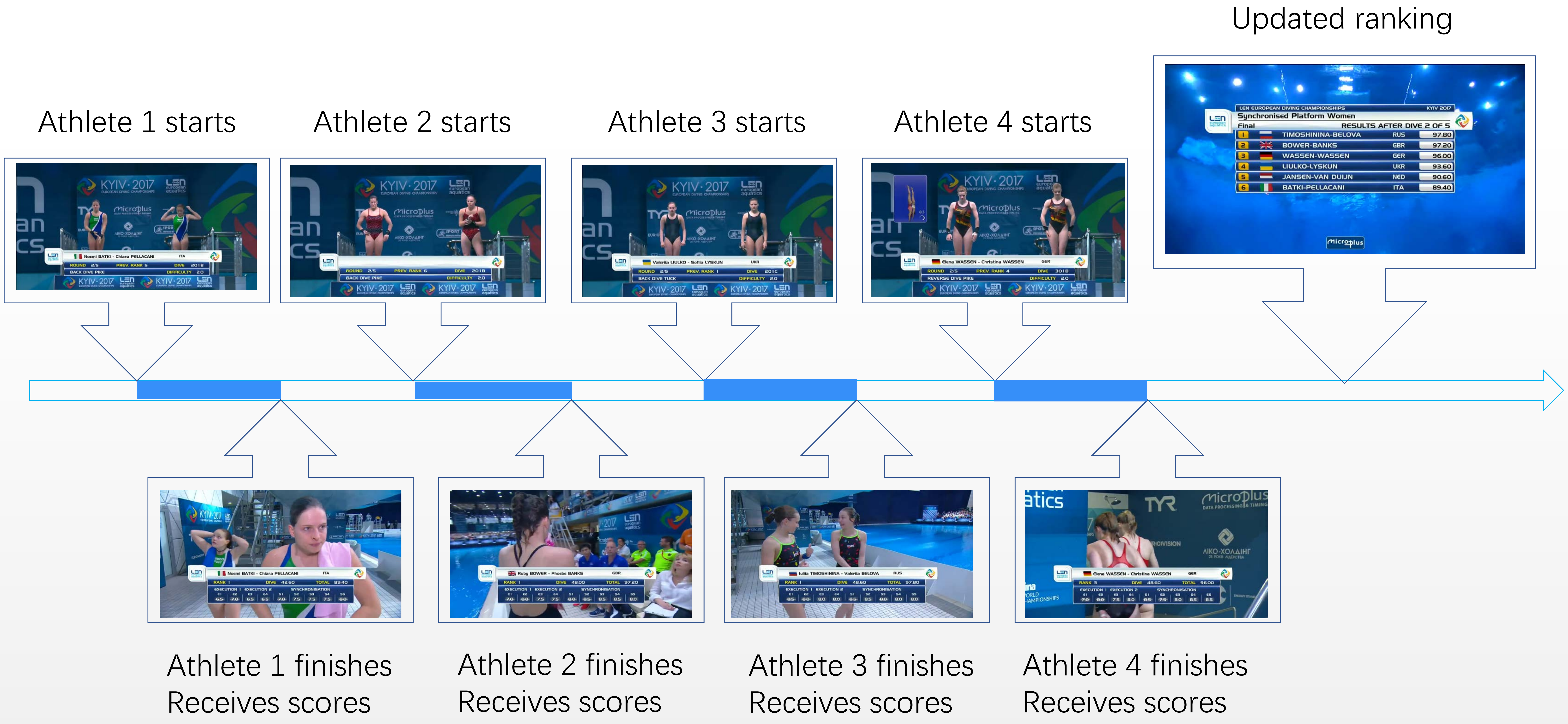
Video Production: Sporting Event Highlights Generation

Diving highlights

- **Why diving:**
  - High viewership in China
  - Relatively simple video structure provides an easier starting point to deliver commercial-quality product
- **Goal:**
  - Using DL technologies to understand the video structure
  - *When, what, and who of the key events along the timeline?*
  - Combine with OCR technology to allow users to create highlights of specific athletes



One round of diving competition



阿里巴巴机器智能技术实验室
Alibaba Machine Intelligence Technology Lab

Creating Diving Highlights

Input long video

↓

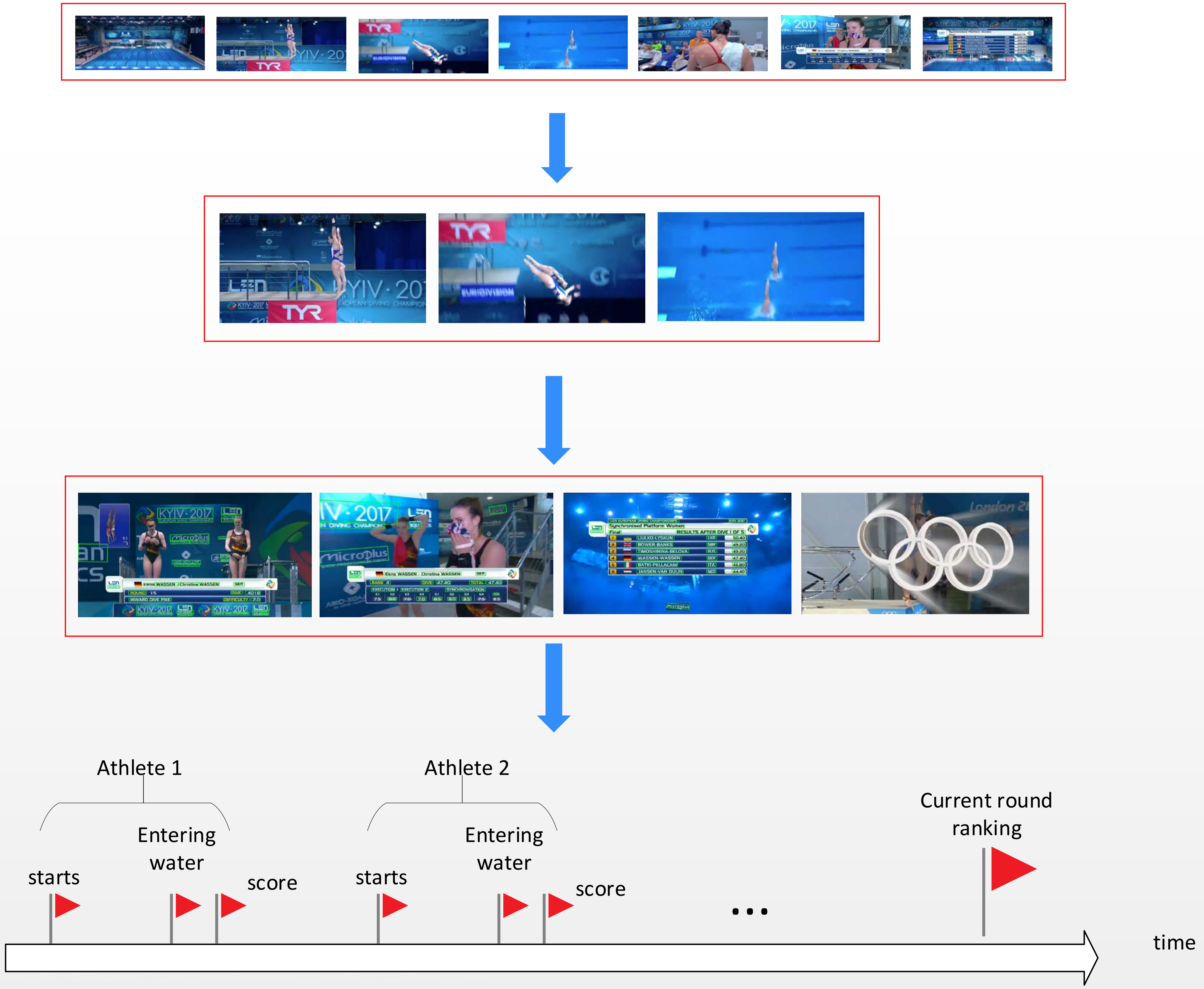
Diving detection

↓

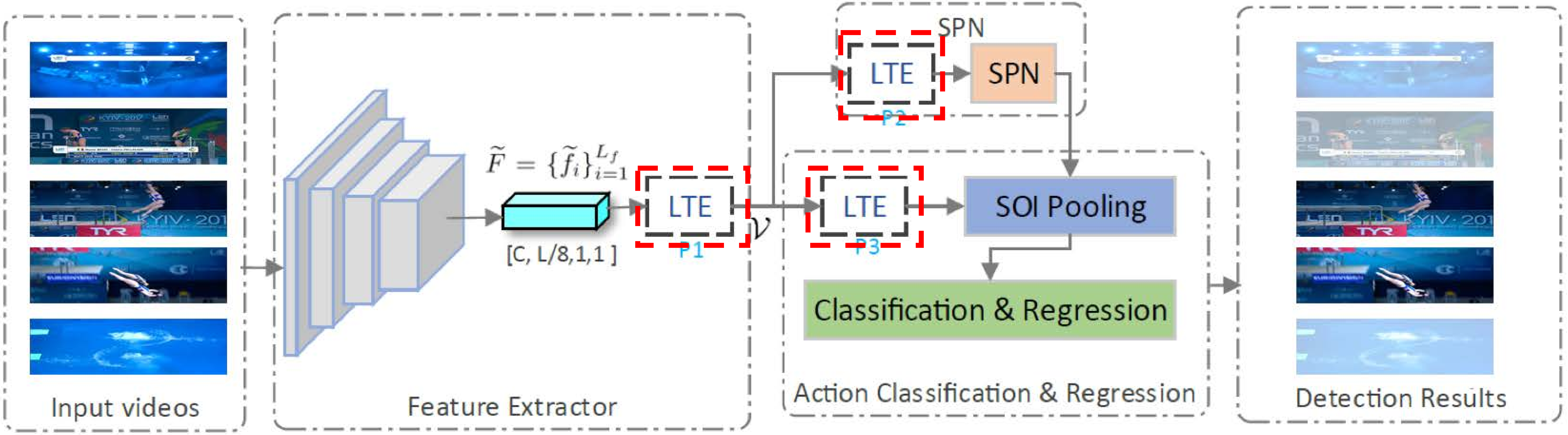
Replay detection | Key caption frame detection | OCR

↓

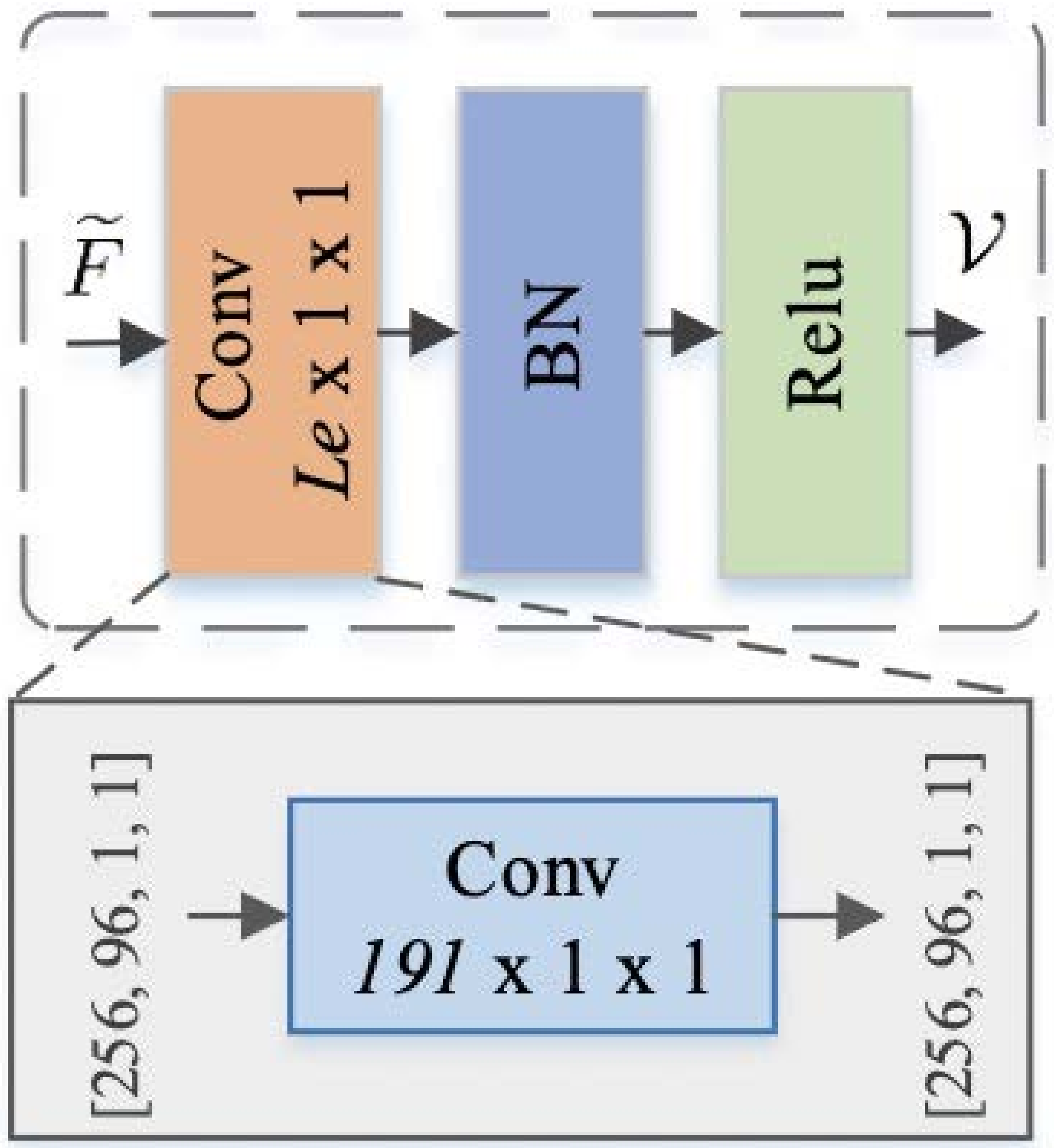
Video structuring



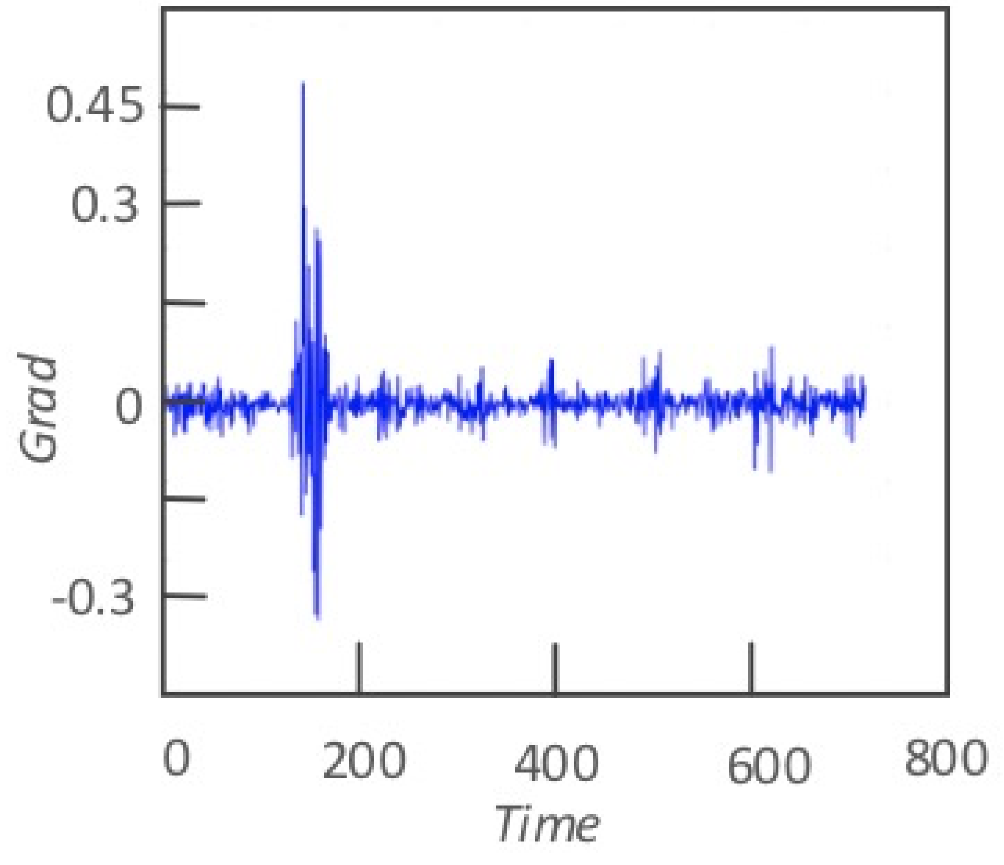
Diving detection

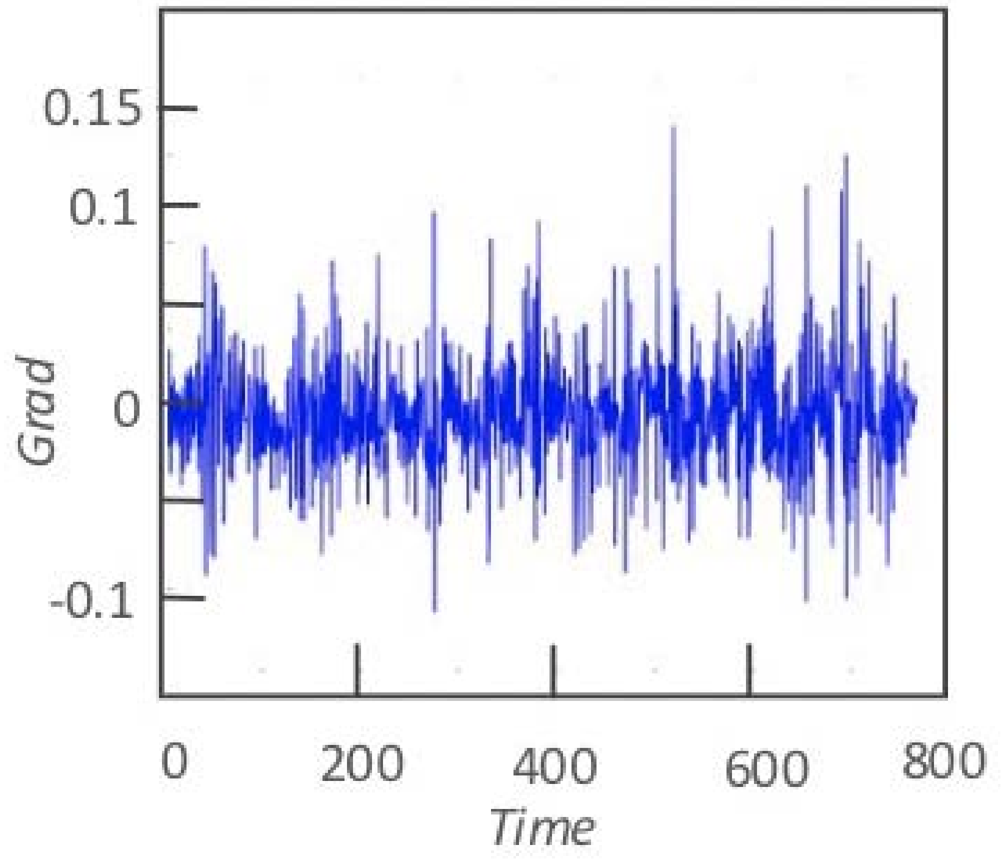


Long-term temporal encoding layer (LTE)



(1) Linear LTE layer





Compared to the baseline method (left), the proposed LTE layer (right) can significantly increase the effective temporal reception field (ERTF)

Detection results: public dataset

Methods requiring both RGB data and optical flow

tIoU	In[1]	0.1	0.3	0.4	0.5	0.6	0.7
Karaman *et al.* [45]	RF[2]	4.6	2.4	1.4	0.9	-	-
Richard et al [24]	RF	39.7	30.0	23.2	15.2	-	-
Shou *et al.* [46]	RF	47.7	36.3	28.7	19.0	10.3	5.3
Yeung *et al.* [47]	RF	48.9	36.0	26.4	17.1	-	-
Yuan *et al.* [22]	RF	51.4	33.6	26.1	18.8	-	-
Shou *et al.* [31]	RF	-	40.1	29.4	23.3	13.1	7.9
Yuan *et al.* [27]	RF	51.0	40.1	27.8	17.8	-	-
Gao *et al.* [32]	RF	60.1	50.1	41.3	31.0	19.1	9.9
Hou *et al.* [21]	RF	51.3	43.7	-	22.0	-	-
Dai *et al.* [48]	RF	-	-	33.3	25.6	15.9	9.0
Zhao *et al.* [28]	RF	**66.0**	51.9	41.0	29.8	-	-
Yang *et al.* [49]	RF	-	-	-	14.7	-	-
Gao *et al.* [34]	RF	54.0	44.1	34.9	25.6	-	-
Humam *et al.* [50]	RF	-	51.8	42.4	30.8	20.2	11.1
Lin *et al.* [6]	RF	-	-	45.0	36.9	28.4	20.0
Shou *et al.* [51]	RF	-	35.8	29.0	21.2	13.4	5.8
Liu *et al.* [8]	RF	-	**53.9**	46.8	37.4	29.5	21.3

Methods requiring only RGB data

End-to-end Method							
Xu *et al.* [11]	R[3]	54.5	44.8	35.6	28.9	-	-
Yu *et al.* [12]	R	49.3	42.6	-	31.9	-	14.2
LTENet	R	59.0	53.2	**48.1**	**41.1**	**32.2**	**22.1**

The proposed method only needs the RGB input, achieves the best action recognition results: For IoU = 0.5, 41% MAP

Detection results: private dataset

Dataset: 1.1k diving video including world tournament and Olympic games since 2010

tIoU	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	mMAP
mAP	0.955	0.944	0.937	0.914	0.881	0.856	0.763	0.554	0.229	0.011	0.7044

Caption Frame Detection

- What is a Caption Frame? The frame containing information about the athlete(s), scores, ranking, etc.
- Caption frame detection, combined with OCR technologies, can be used to generate diving highlight of specific athletes

Demo: diving highlights of specific athletes

[Original video 39 minutes: http://publicvideos.oss-cn-hangzhou-zmf.aliyuncs.com/d_Q3PH8jsD0.mp4）](http://publicvideos.oss-cn-hangzhou-zmf.aliyuncs.com/d_Q3PH8jsD0.mp4)

Concluding remarks

- Video is much more difficult than images
- Processing efficiency can be significantly increased if learning can be conducted in the compressed domain
- Can compression technologies be adapted to assist with AI-based video learning?


Q
A