Attend and Interact: Higher-Order Object Interactions for Video Understanding

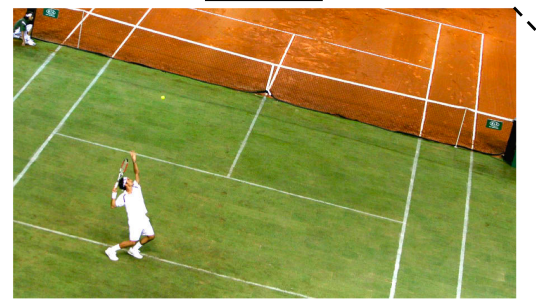
Chih-Yao Ma[1], Asim Kadav[2], Iain Melvin[2], Zsolt Kira[3], Ghassan AlRegib[1], and Hans Peter Graf[2]

[1]Georgia Institute of Technology, [2]NEC Laboratories America, [3]Georgia Tech Research Institute

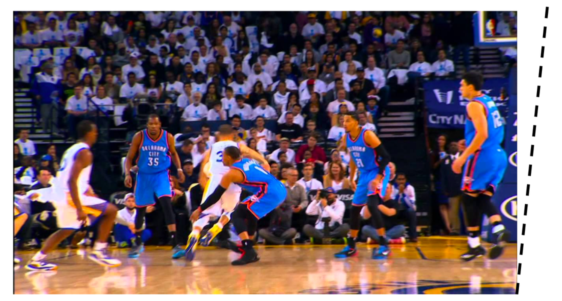
2018 CVPR SALT LAKE CITY • JUNE 18-22

1 Background: Video Understanding

Tennis



Basketball



Existing approaches:

- Ignore the fine-grained details of the scene.
- Do not infer interactions between various objects in the video.

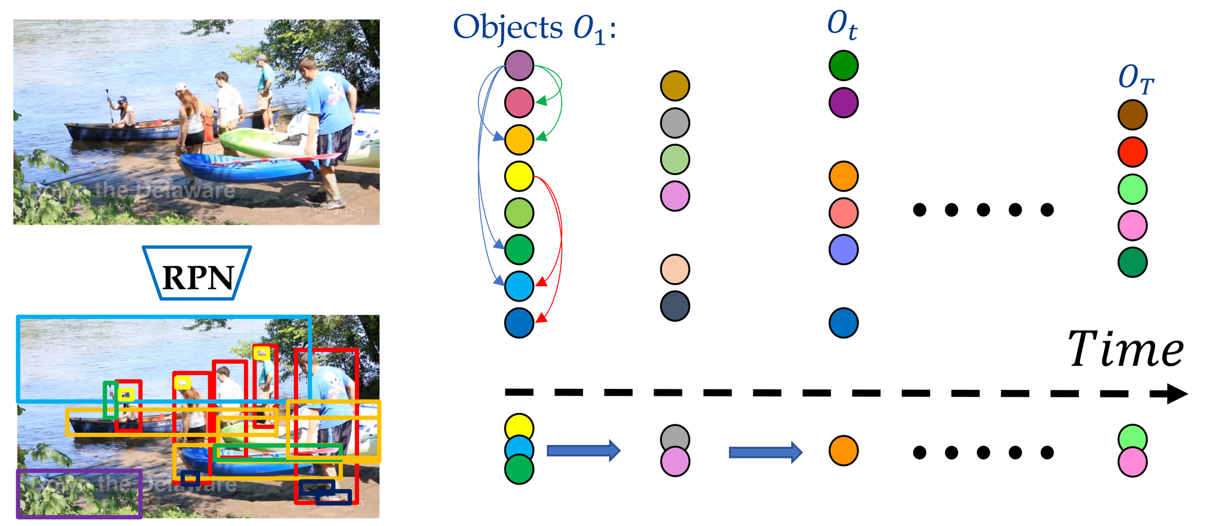
Object interactions matter:

- Human actions often involve complex interactions with various objects in the scene.

Skiing

Snowboarding

2 Fine-grained Object Interactions



Goal:

- Efficiently detect object interactions
- Temporal reasoning

Challenges:

- Unordered variable-lengths of object sets that span across time
- Video contains hundreds or thousands of frames
- Object-object pairs are too large to fully represented by a finite-capacity neural network

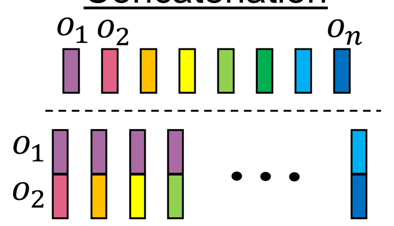
Higher-order interactions:

Object Interactions are **not** always between two objects (one-to-one)

3 Contribution: From Pairwise to Higher-Order Interactions

Interactions/relationships:

$$RN(O) = f_\phi\left(\sum_{i,j} f_\theta(o_i, o_j)\right)$$

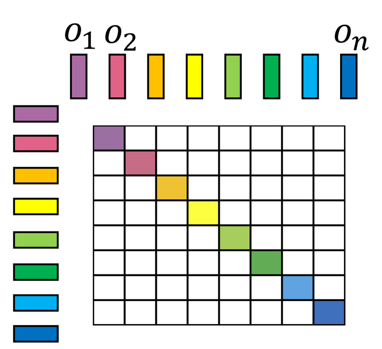


Concatenation [1]:

$$f_\theta(o_i, o_j) = W_{f_\theta}^T(o_i \parallel o_j)$$

Dot-product:

$$f_\theta(o_i, o_j) = \theta(o_i)^T\phi(o_j)$$
$$\rightarrow O^T W_\theta^T W_\phi O$$



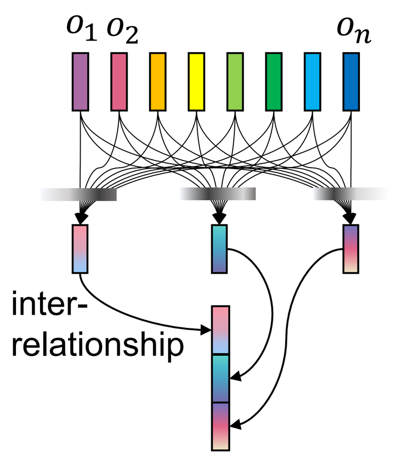
Higher-order interactions:

- Interactions over groups of inter-related objects
- Covers pair-wise or triplet object relationships as a special case

Goal:

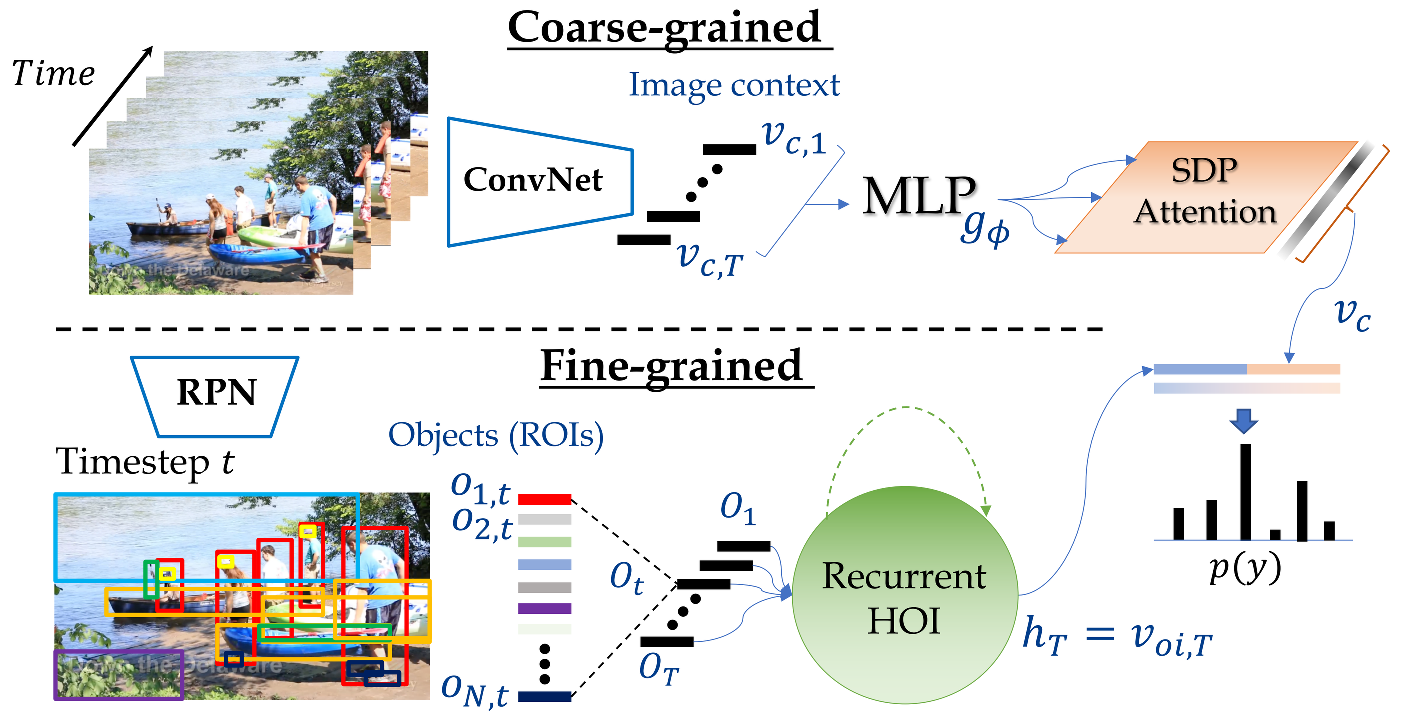
- Detect inter-object relationships
- Objects with significant relationships are selected
- Groups of selected object relationships are concatenated.

Higher-Order Interaction



[1] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *NIPS* 2017.

4 Action Recognition – *SINet*



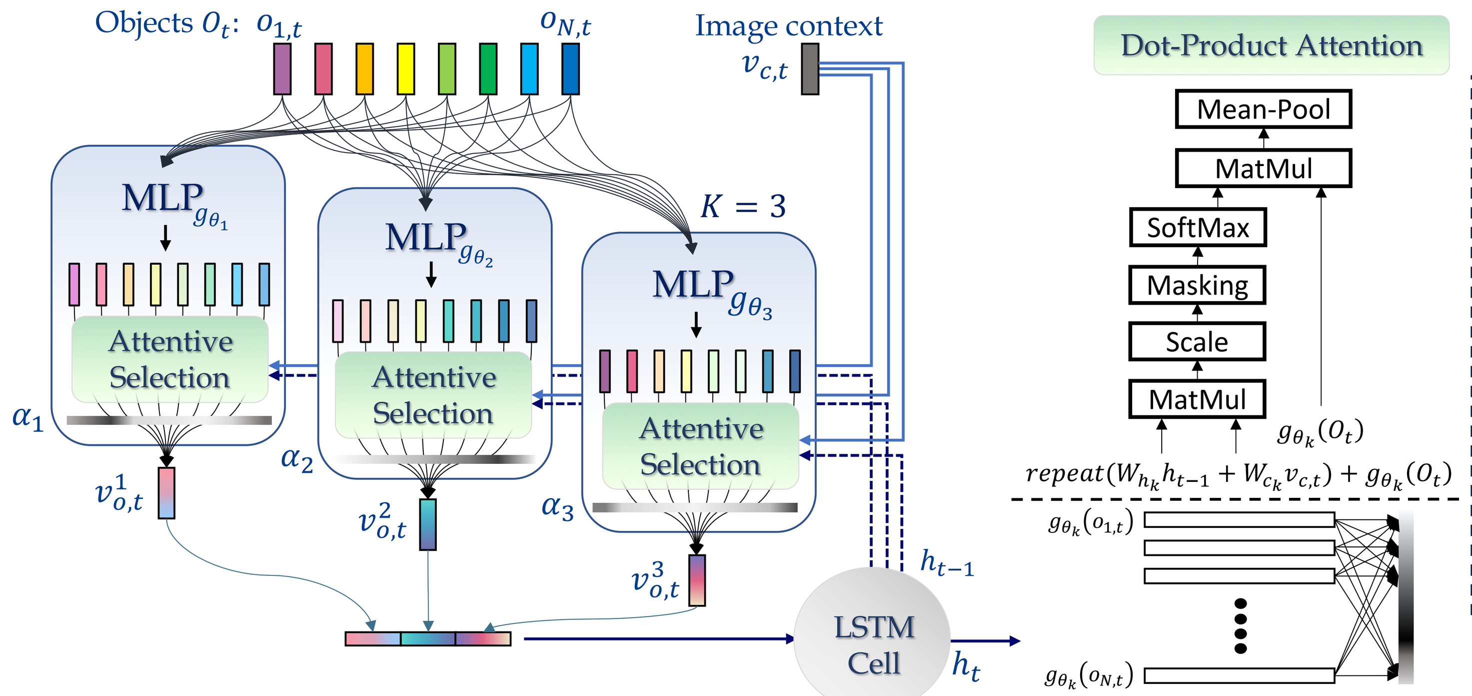
Coarse-grained:

- Video frames are encoded via a ConvNet
- Temporal pooling via SDP-Attention

Fine-grained:

- Objects (ROIs) are obtained from a RPN
- Progressively detect higher-order interactions via the HOI module

5 Recurrent Higher-order Interaction (HOI)

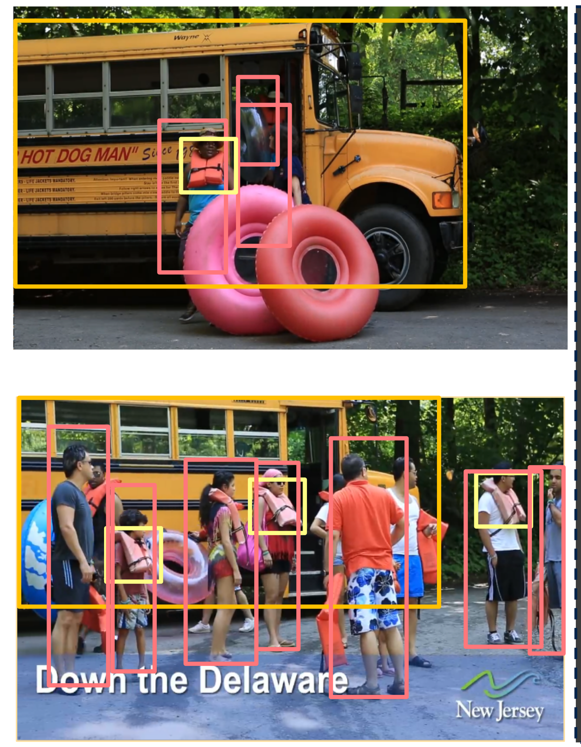


Selects K groups of arbitrary objects with inter-object relationships

Condition on:

- Image context $v_{c,t}$
- Object sets O_t
- Previous interactions h_{t-1}

6 Video Captioning – *SINet-Caption*

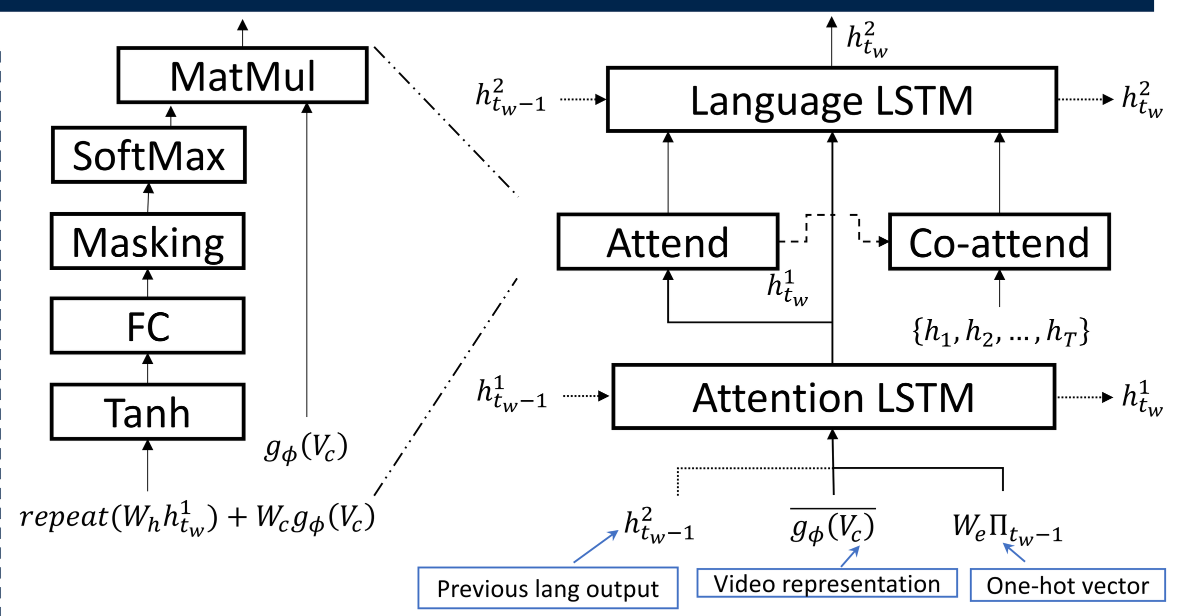


Caption:

A group of people get off of a yellow school bus with life rafts around their neck.

Relationships:

- [group of people, get off, yellow school bus]
- [group of people, with, life rafts]



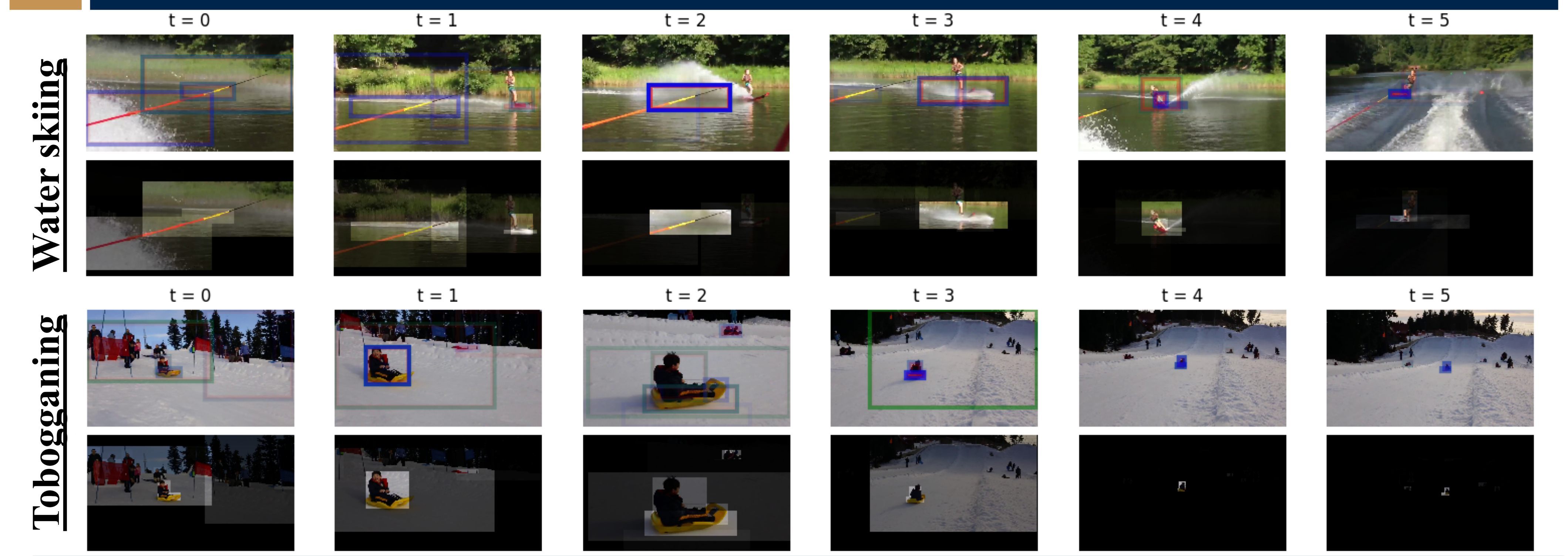
The **Attention LSTM** identifies which parts of the video in spatiotemporal feature space are needed for **Language LSTM** to generate the next word.

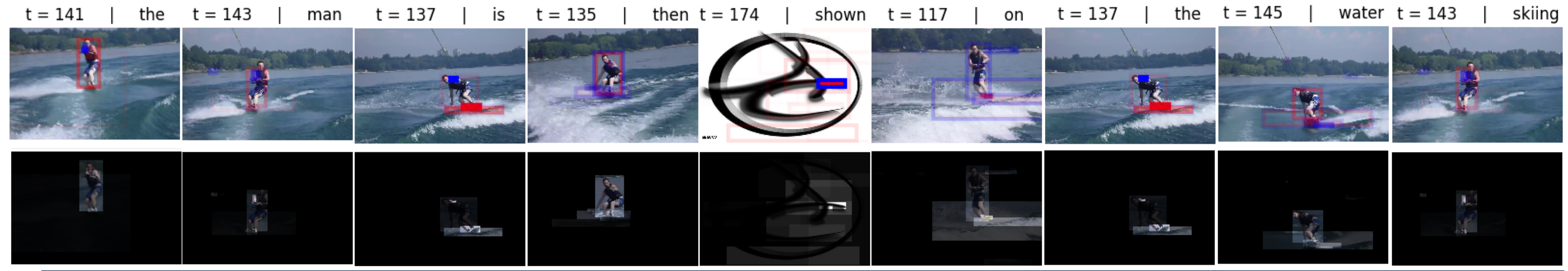
7 Experiment – Kinetics & ActivityNet Captions

Method	Top-1	Top-5	FLOP(e^9)
Prior Arts			
I3D (25 FPS) (test)	71.1	89.3	
TSN (Inception-ResNet-v2) (2.5 FPS)	73.0	90.9	
Ours (1 FPS)			
Img feat + LSTM (baseline)	70.6	89.1	
Img feat + temporal SDP-Attn	71.1	89.6	
Obj feat (mean-pooling)	72.2	90.2	
Obj pairs (mean-pooling)	73.4	90.8	18.3
Img + obj feat (mean-pooling)	73.1	91.1	
SINet (K = 1)	73.9	91.3	**2.7**
SINet (K = 2)	74.2	91.5	5.3
SINet (K = 3)	**74.2**	**91.7**	8.0

Method	B@4	R	M	C
Test set				
LSTM-YT (C3D)	1.24	-	6.56	14.86
S2VT (C3D)	2.62	-	7.85	20.97
H-RNN	2.53	-	8.02	20.18
S2VT + Full context	3.98	-	9.46	24.56
LSTM-A + policy gradient + retrieval (ResNet + P3D ResNet)	-	-	12.84	-
Validation set (Avg. 1st and 2nd)				
LSTM-A + policy gradient + retrieval (ResNet + P3D ResNet)	**3.13**	14.29	8.73	14.75
SINet-Caption – img (ResNeXt)	1.84	20.46	9.56	43.12
SINet-Caption – obj (ResNeXt)	1.92	20.67	9.56	44.02
SINet-Caption – img + obj – no co-attn	2.03	21.08	9.79	44.81
SINet-Caption – img + obj – co-attn	1.98	**21.25**	**9.84**	**44.84**

8 Qualitative Results





Distinguish interactions when actions with common objects presented – *horse*:

- People are *riding* horses.
- A woman is *brushing* a horse.
- People are playing *polo* on a field.
- The man ties *up* the calf.