UNIVERSITY OF AMSTERDAM

2021 ICCV OCTOBER 11-17 VIRTUAL

# Social Fabric: Tubelet Compositions for Video Relation Detection



Shuo Chen

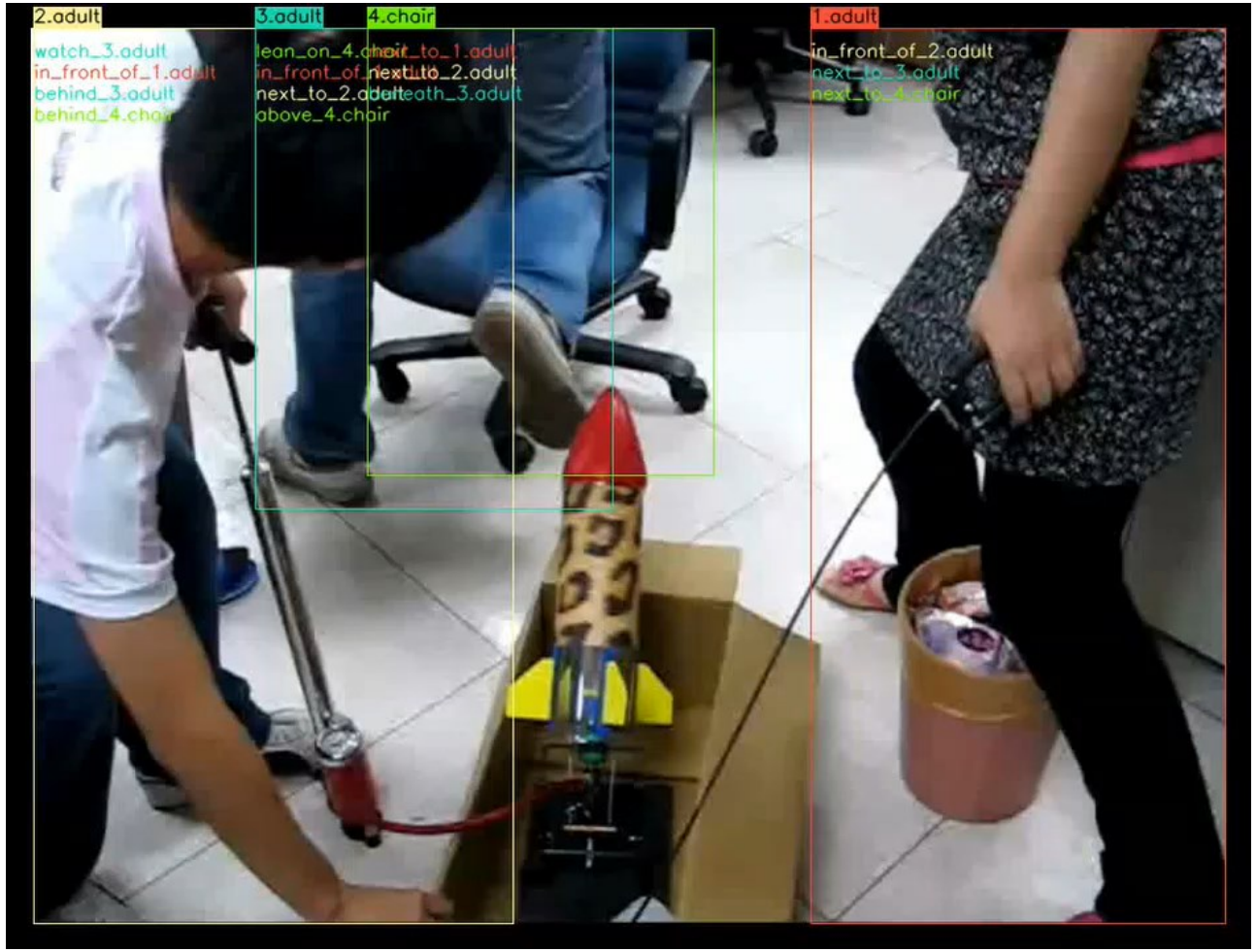
Zenglin Shi

Pascal Mettes

Cees G. M. Snoek

VIS Lab, University of Amsterdam

# Goal: video relation detection




**Input** – video

**Output** – relation triplet

- subject trajectory with class
- predicate class
- object trajectory with class

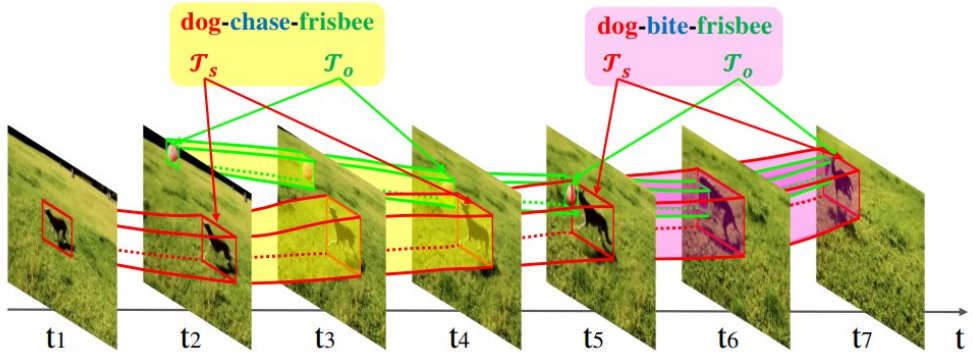
**Challenge**

- spatiotemporal localization
- modeling the interactions

# Related works

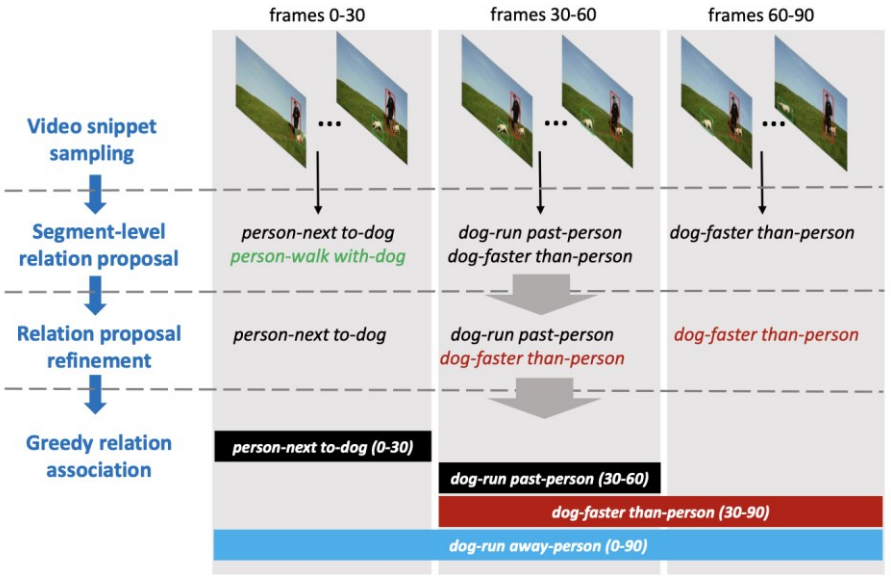
## Shang et al. MM 2017

Pose the problem and introduce first dataset




## Leading approaches [Qian et al. MM 2019, Xie et al MM 2020, Su et al. MM 2020, etc.]

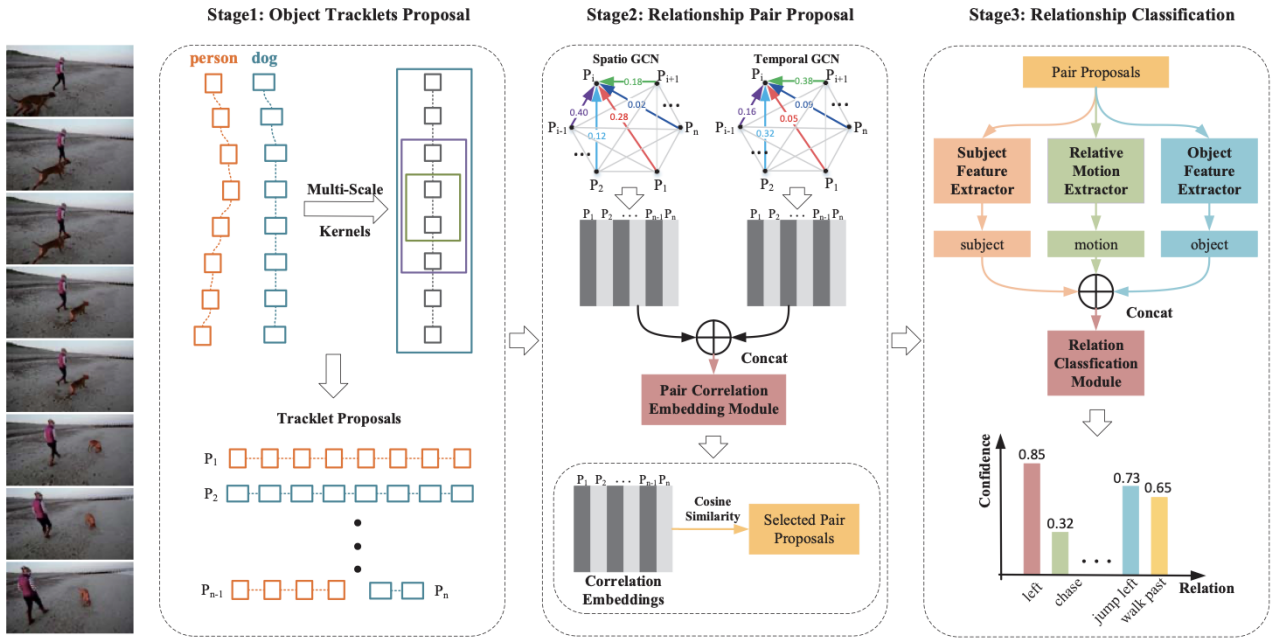
1. Generate object and subject proposals **on short video snippets**
2. Encode the proposals and predict their relation
3. Associate the relations over the entire video



# Related works

Liu et al. CVPR 2020 forego the need for snippets.

1. Localize individual object and subjects **tubelets** throughout the **entire** video
2. Filter out unlikely pairs
3. Predict predicates for the remaining ones



# Key idea

All existing works on video relation prediction treat object proposals or tubelets as **single entities** and model their relations ***a posteriori.***

## Key idea

All existing works on video relation prediction treat object proposals or tubelets as **single entities** and model their relations ***a posteriori.***

We classify and detect predicates for **pairs** of object tublets ***a priori***.

# Key idea

All existing works on video relation prediction treat object proposals or tubelets as **single entities** and model their relations ***a posteriori.***

We classify and detect predicates for **pairs** of object tublets ***a priori***.

As a result, we only need two stages:

1. Generating interaction proposals from tubelet pairs
2. Predicting the appropriate predicate.

# Key idea

All existing works on video relation prediction treat object proposals or tubelets as **single entities** and model their relations ***a posteriori.***

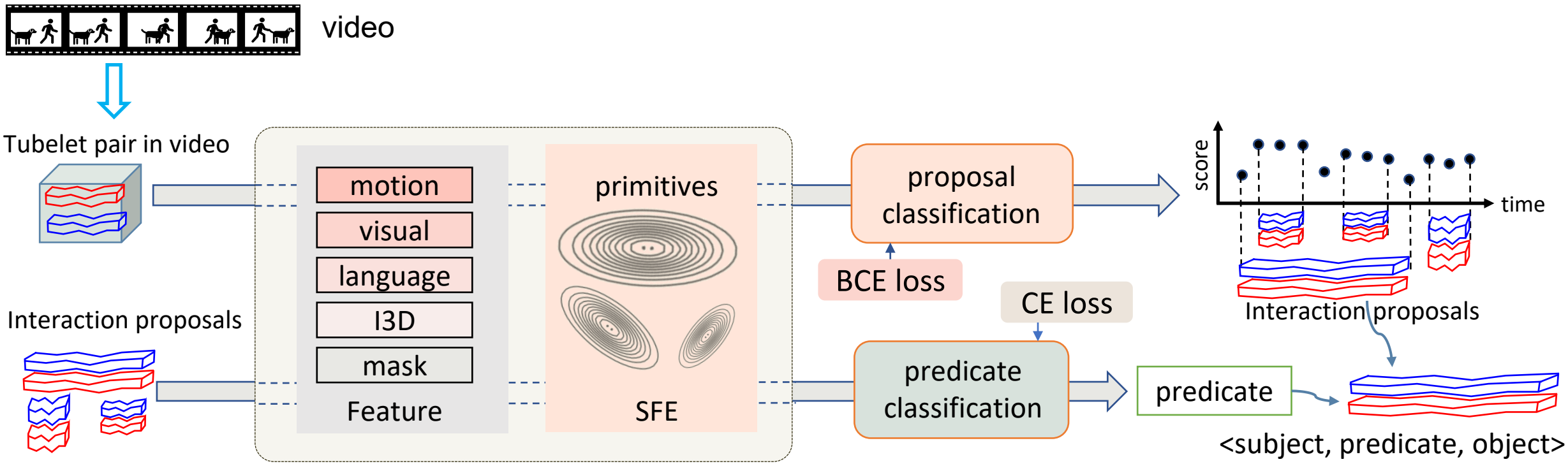
We classify and detect predicates for **pairs** of object tublets ***a priori***.

As a result, we only need two stages:

1. Generating interaction proposals from tubelet pairs
2. Predicting the appropriate predicate.

We call our tubelet representation '**social fabric**'.

# Two-stage network



# Two-stage network

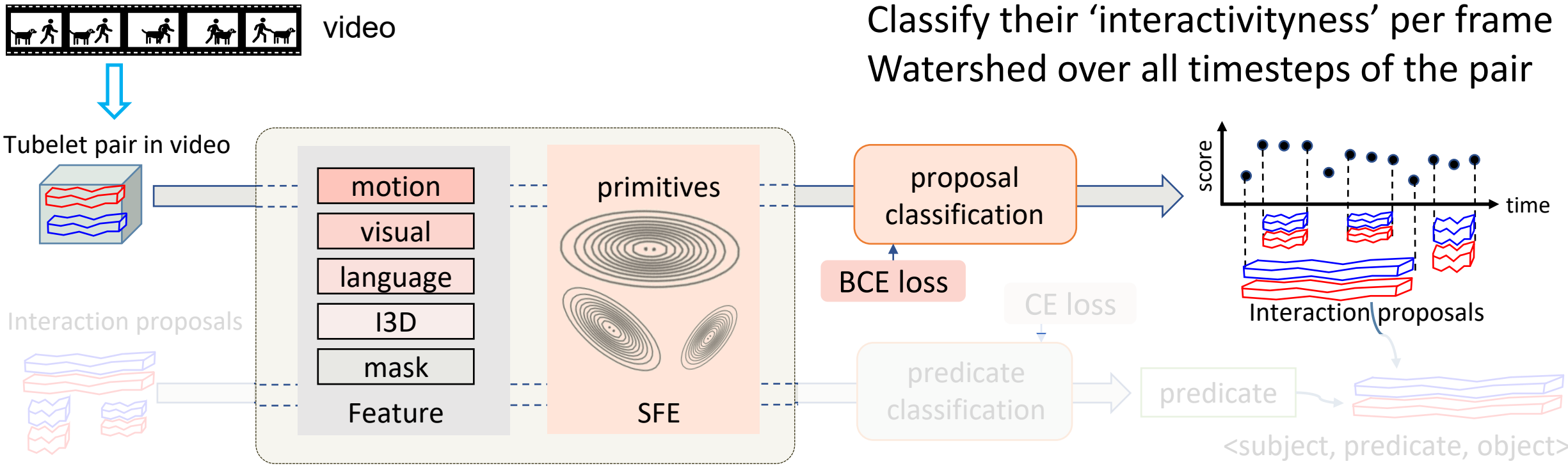
**Stage 1: Interaction proposals**

Object detection and linking

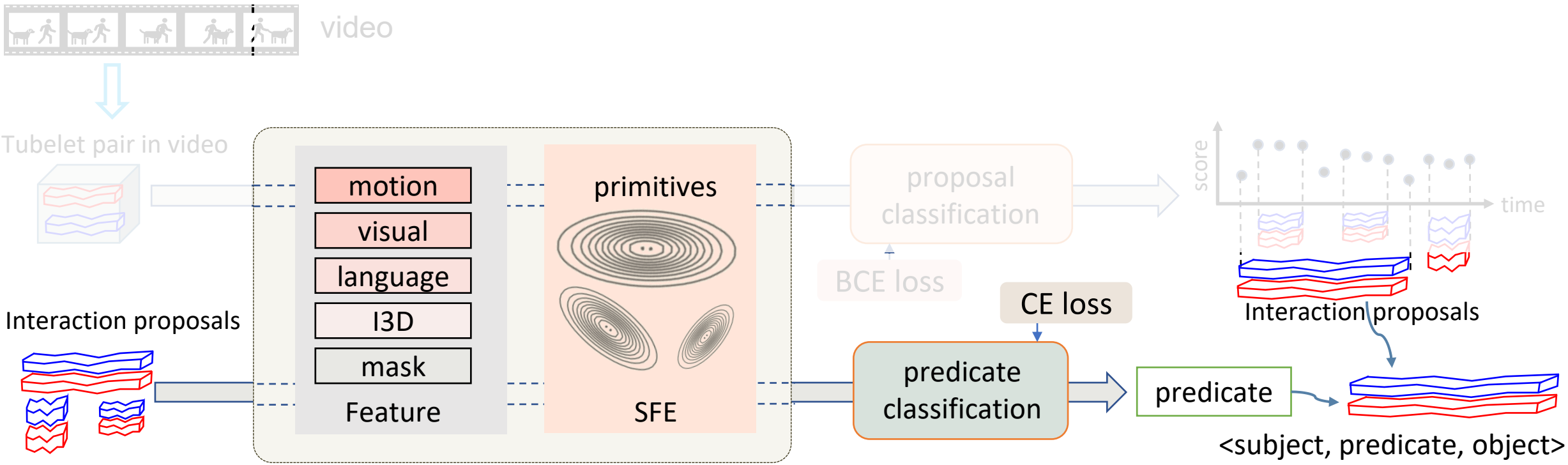
Consider all tubelet combinations

Classify their ‘interactivityness’ per frame

Watershed over all timesteps of the pair



# Two-stage network




**Stage 2: Predicate prediction**

Encode each proposal (pair of tubelets)

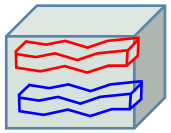
Final linear layer obtains predicate score

Multiply with object and subject scores

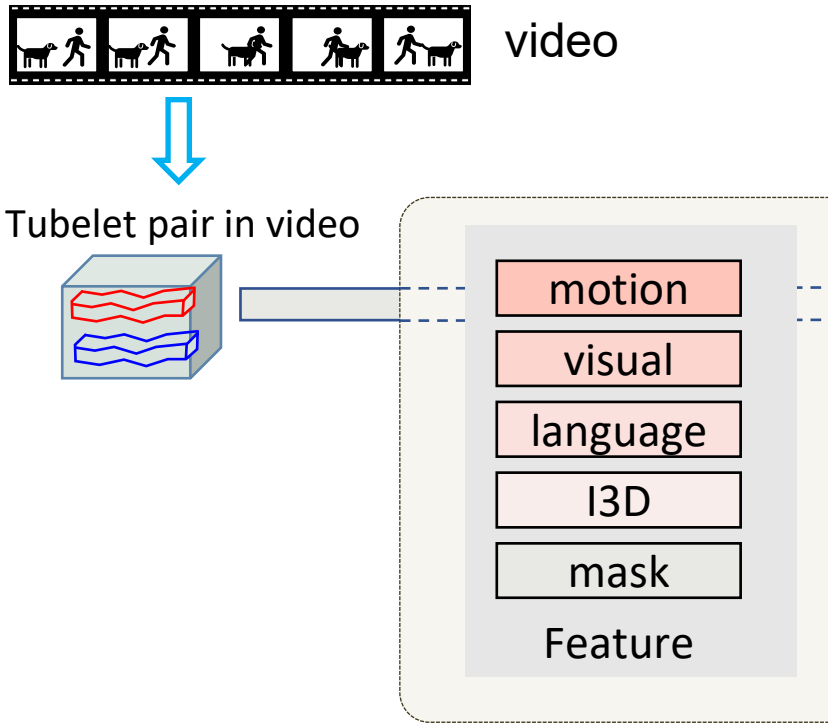
# Object detection and linking

video

Tubelet pair in video



# Multi-modal features




## Video-Language Features

**Motion:** spatial location features by [Su et al. MM 2020]

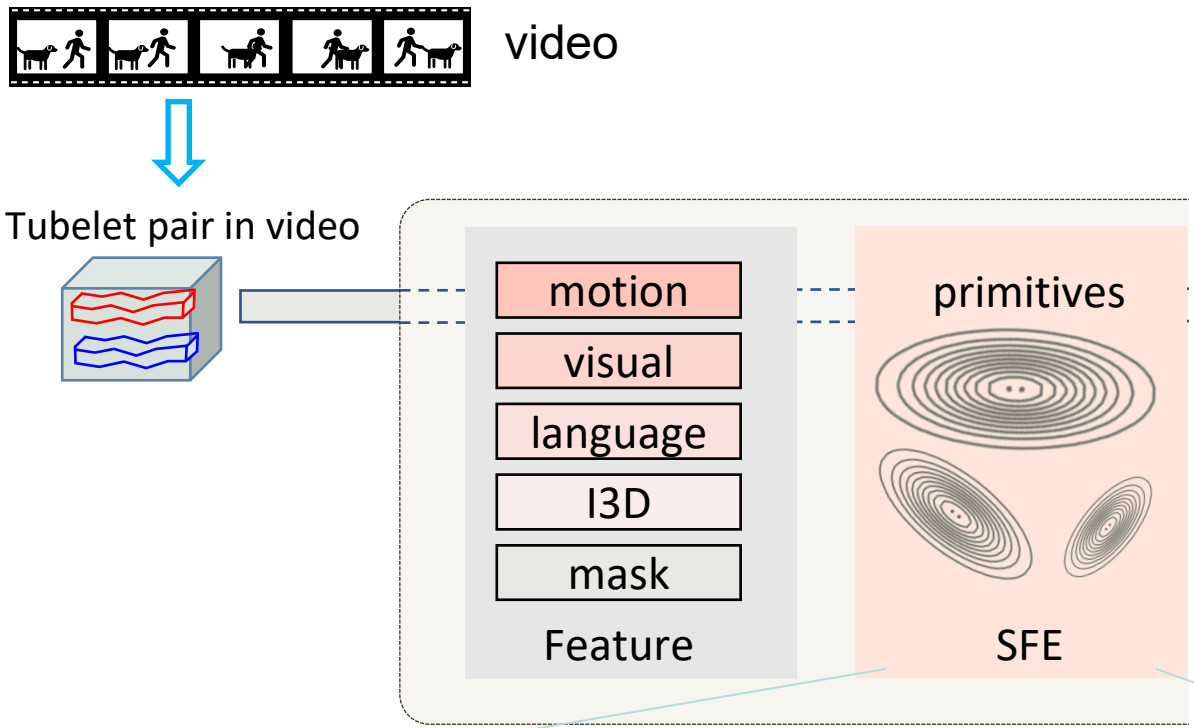
**Visual:** Faster-R-CNN backbone followed by RoI pooling layer.

**Language:** 600-dim word2vec, pre-trained on Google News.

**I3D:** I3D with fixed dimension of 832.

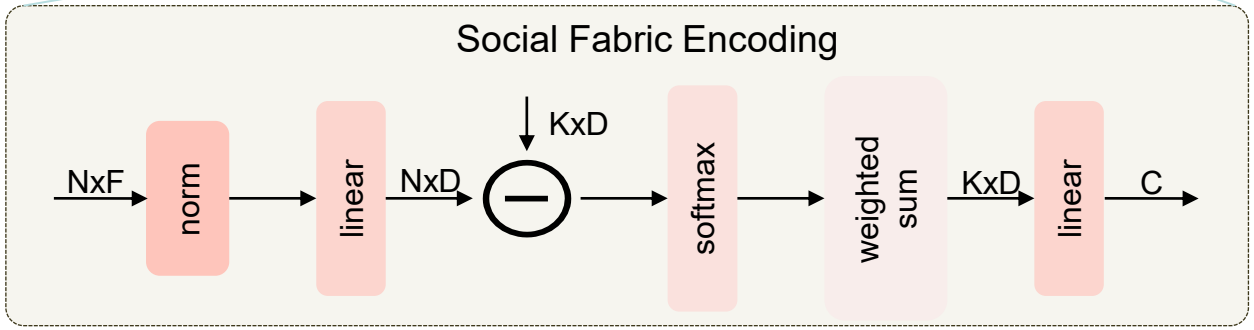
**Mask**: generated based on bounding boxes of subject and object in tubelet pairs, following [Xie et al. MM 2020]

# Social Fabric Encoding



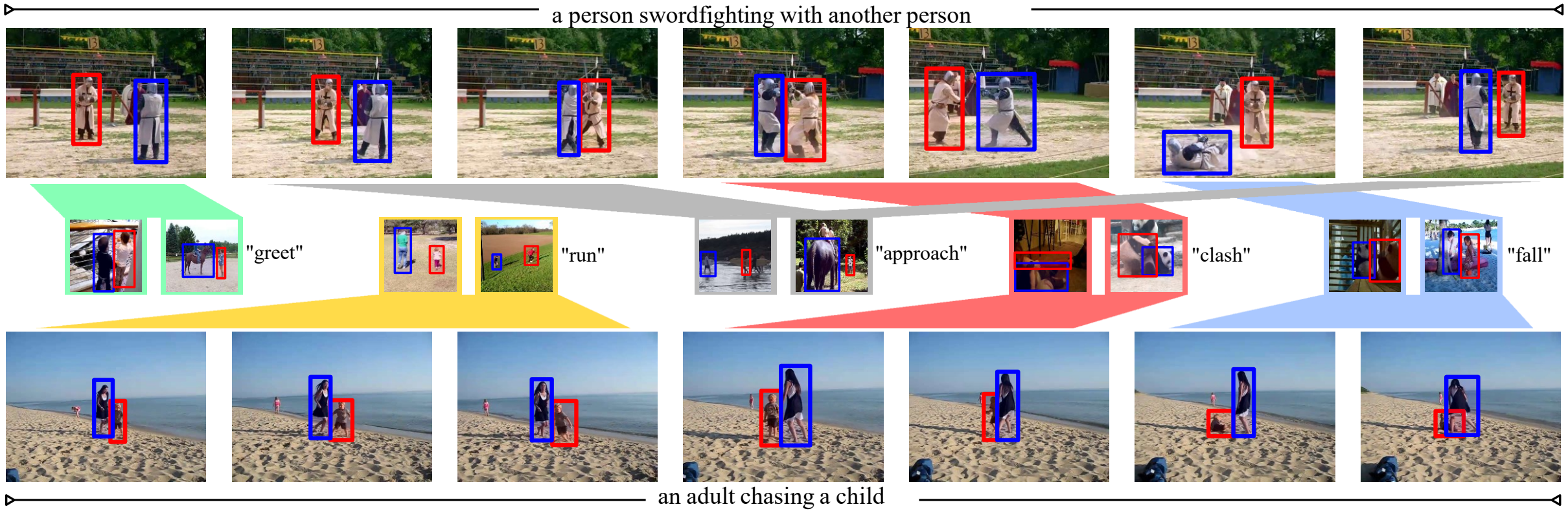

## Encoding

- Layer normalize features plus linear layer
- Learn set of *K* interaction primitives
- Encode tubelet pair as primitive combination

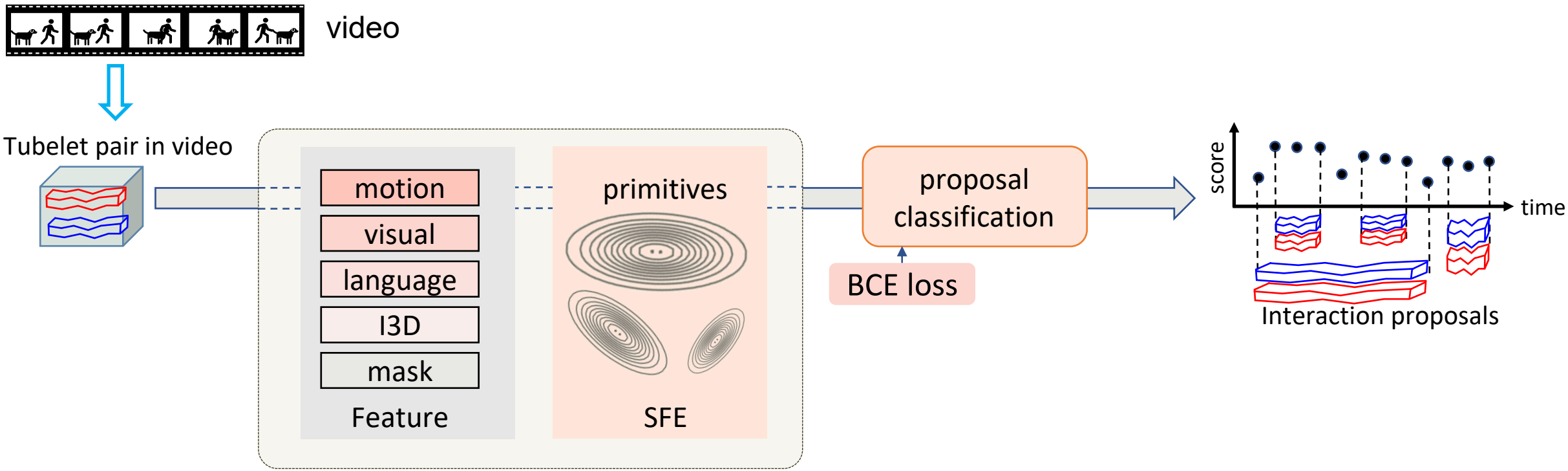


# Intuition behind encoding

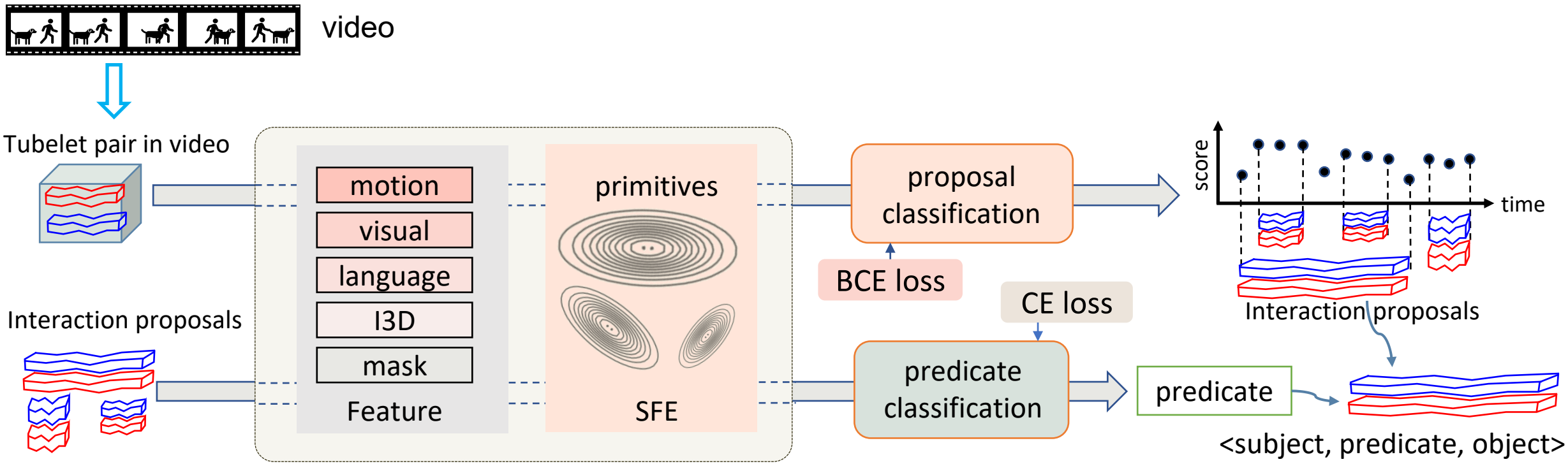
We encode compositions of interaction primitives over tubelet pairs.
Data-driven primitives may correspond to interactions like ``greet'', ``clash'' and ``fall''.



# Social fabric – determine good proposals



# Social fabric




**Stage 2: Predicate prediction**

Encode each proposal (pair of tubelets)

Final linear layer obtains predicate score

Multiply with object and subject scores

Shang et al, MM 2017

# Datasets

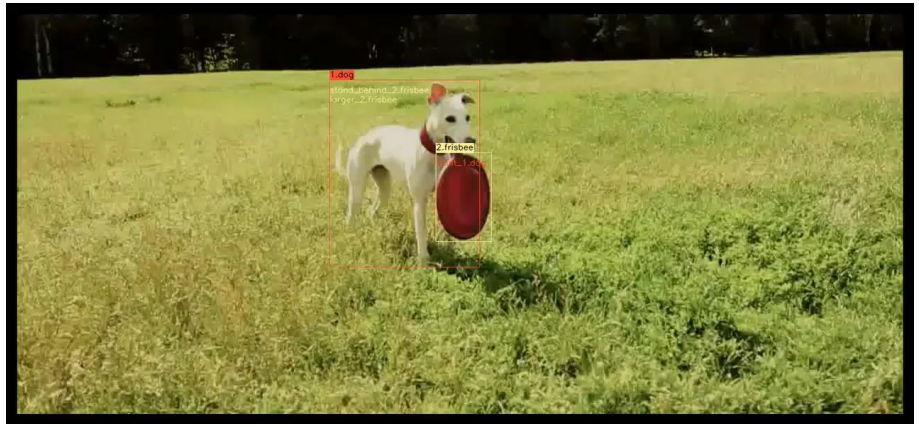
## ImageNet VidVRD dataset

training set: 800 videos

testing set: 200 videos

35 subject/object categories

132 predicate categories



# Ablating the features

| Feature type | | | | | Relation tagging | | | Relation detection | | |
|---|---|---|---|---|---|---|---|---|---|---|
| motion | visual | language | I3D | mask | P@1 | P@5 | P@10 | mAP | R@50 | R@100 |
| ✓ | | | | | 50.97 | 39.57 | 31.58 | 6.14 | 6.74 | 8.70 |
| ✓ | ✓ | | | | 56.89 | 44.76 | 34.07 | 8.93 | 7.38 | 9.22 |
| ✓ | ✓ | ✓ | | | 59.24 | 47.24 | 35.99 | 9.54 | 8.49 | 10.17 |
| ✓ | ✓ | ✓ | ✓ | | 61.52 | 50.05 | 38.48 | 10.04 | 8.94 | 10.69 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **68.86** | **55.16** | **43.40** | **11.21** | **9.99** | **11.94** |

***Tubelet representation effectively captures multi-modal features***

# Ablating the Social Fabric Encoding

| Clusters | 1 | 8 | 32 | 64 | 128 |
|---|---|---|---|---|---|
| mAP | 10.05 | 10.69 | 10.91 | **11.21** | 11.01 |

*64 primitives good trade-off*

| | Relation tagging | Relation detection |
|---|---|---|
| Encoding | P@1 | mAP |
| average pooling | 62.73 | 10.05 |
| transformer | 63.86 | 10.07 |
| NetVLAD | 65.34 | 10.15 |
| NetRVLAD | 66.80 | 10.55 |
| Social Fabric | **68.86** | **11.21** |

*Better than alternatives*

| interaction proposal encoding | predicate prediction encoding | relation detection | | | relation tagging | | |
|---|---|---|---|---|---|---|---|
| | | mAP | R@50 | R@100 | P@1 | P@5 | P@10 |
| | | 6.14 | 6.74 | 8.7 | 50.97 | 39.57 | 31.58 |
| ✓ | | 6.94 | 7.22 | 9.26 | 52.52 | 40.53 | 32.34 |
| ✓ | ✓ | **8.93** | **9.15** | **11.13** | **54.57** | **43.58** | **34.55** |

*Contribute to both two stages*

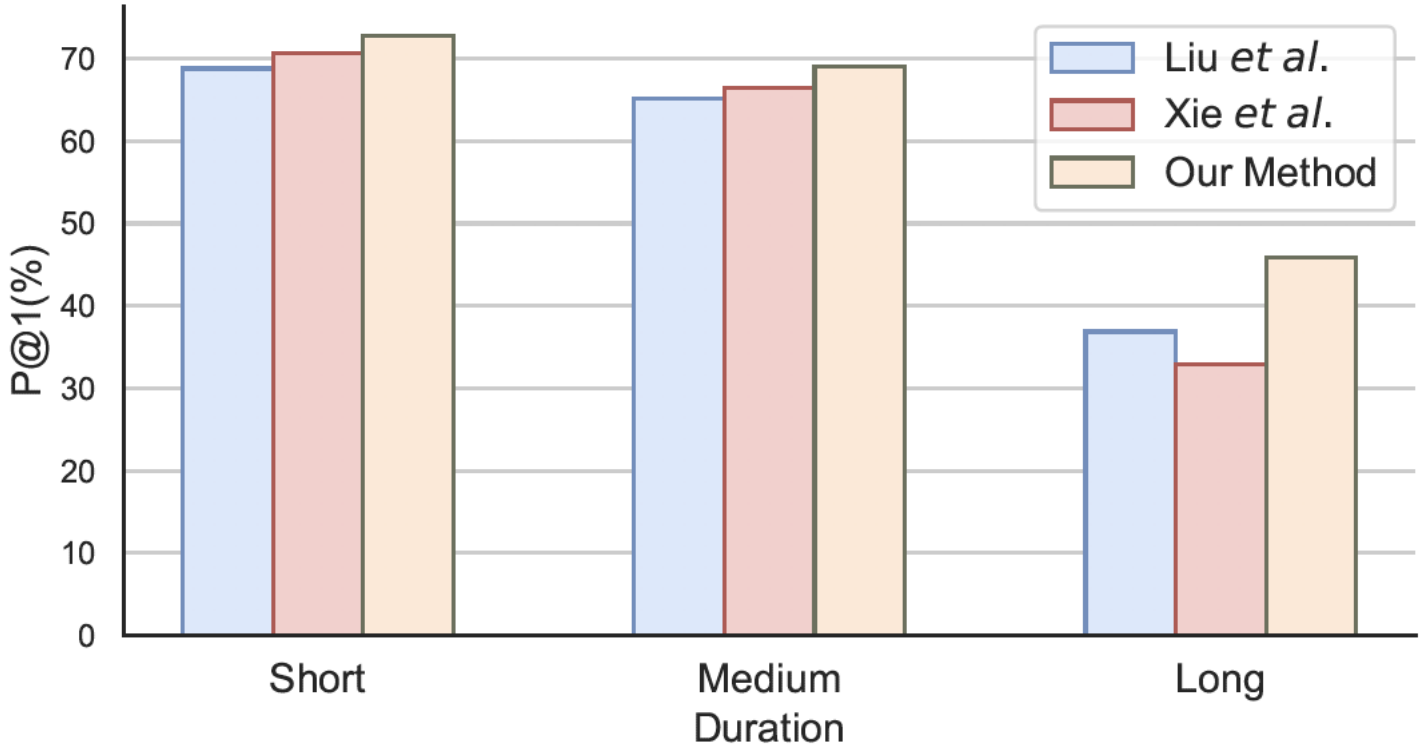
# Compared to SOTA

| | ImageNet-VidVRD | | | | | | VidOR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Relation tagging | | | Relation detection | | | Relation tagging | | Relation detection | | |
| | P@1 | P@5 | P@10 | mAP | R@50 | R@100 | P@1 | P@5 | mAP | R@50 | R@100 |
| Shang *et al.* | 43.00 | 28.90 | 20.80 | 8.58 | 5.54 | 6.37 | - | - | - | - | - |
| Tsai *et al.* | 51.50 | 39.50 | 28.23 | 9.52 | 7.05 | 8.67 | - | - | - | - | - |
| Qian *et al.* | 57.50 | 41.00 | 28.50 | 16.26 | 8.07 | 9.33 | - | - | - | - | - |
| Sun *et al.* | - | - | - | - | - | - | 51.20 | 40.73 | 6.56 | 6.89 | 8.83 |
| Su *et al.* | 57.50 | 41.40 | 29.45 | 19.03 | 9.53 | 10.38 | 50.72 | 41.56 | 6.59 | 6.35 | 8.05 |
| Liu *et al.* | 60.00 | 43.10 | 32.24 | 18.38 | 11.21 | 13.69 | 48.92 | 36.78 | 6.85 | 8.21 | 9.90 |
| Xie *et al.* | - | - | - | - | - | - | 67.43 | - | 9.93 | 9.12 | - |
| ***This paper***, features as Su *et al.* | 57.50 | 43.40 | 31.90 | 19.23 | 12.74 | 16.19 | 54.57 | 43.58 | 8.93 | 9.15 | 11.13 |
| ***This paper***, features as Liu *et al.* | 61.00 | 47.50 | 36.60 | 19.77 | 12.91 | 16.32 | 55.40 | 45.74 | 9.13 | 9.36 | 11.30 |
| ***This paper***, features as Xie *et al.* | - | - | - | - | - | - | 68.62 | 53.34 | 11.05 | 9.91 | 11.89 |
| ***This paper***, our features | **62.50** | **49.20** | **38.45** | **20.08** | **13.73** | **16.88** | **68.86** | **55.16** | **11.21** | **9.99** | **11.94** |

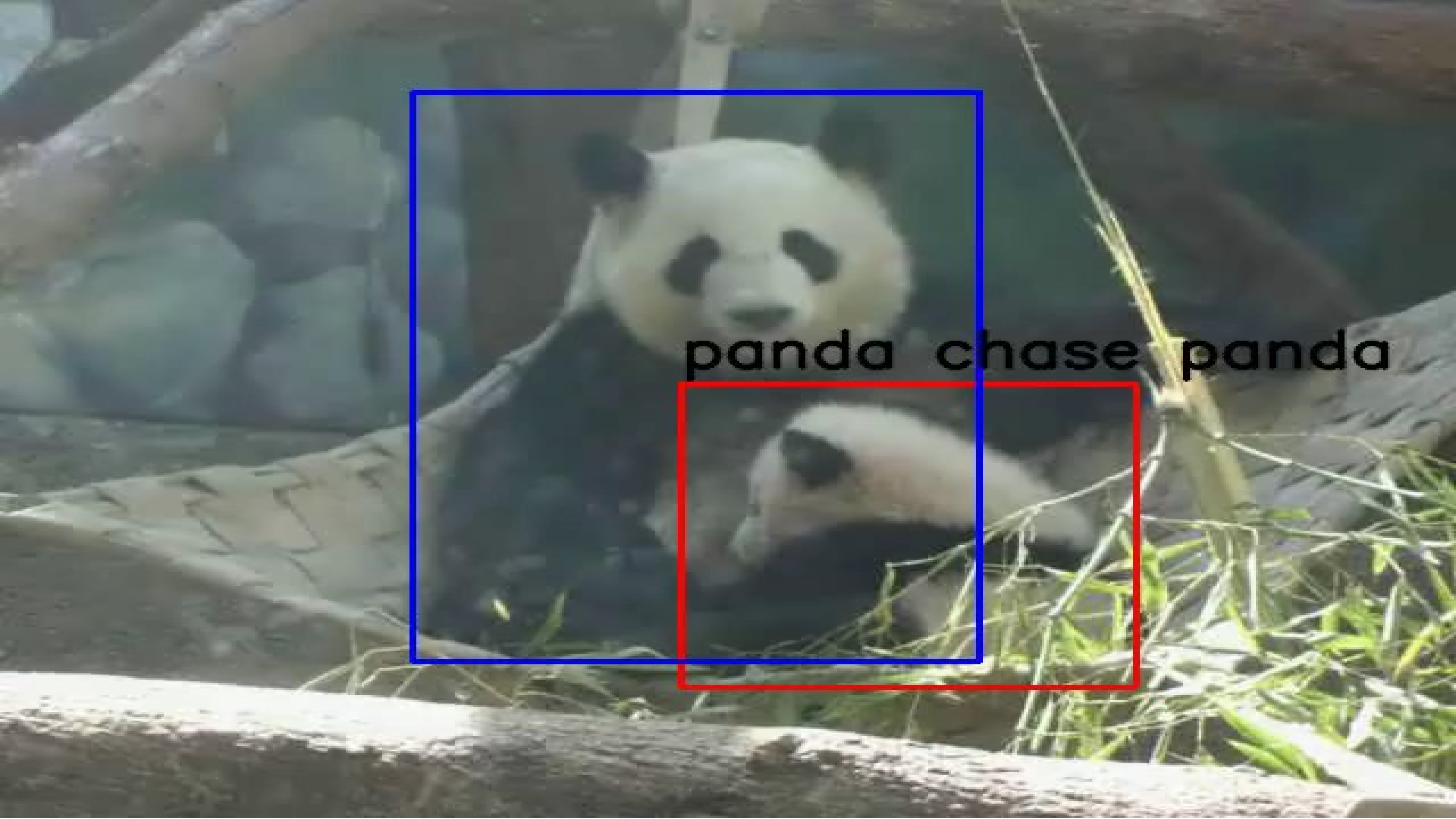
***We outperform all snippet-methods as well as Liu et al. using their features***

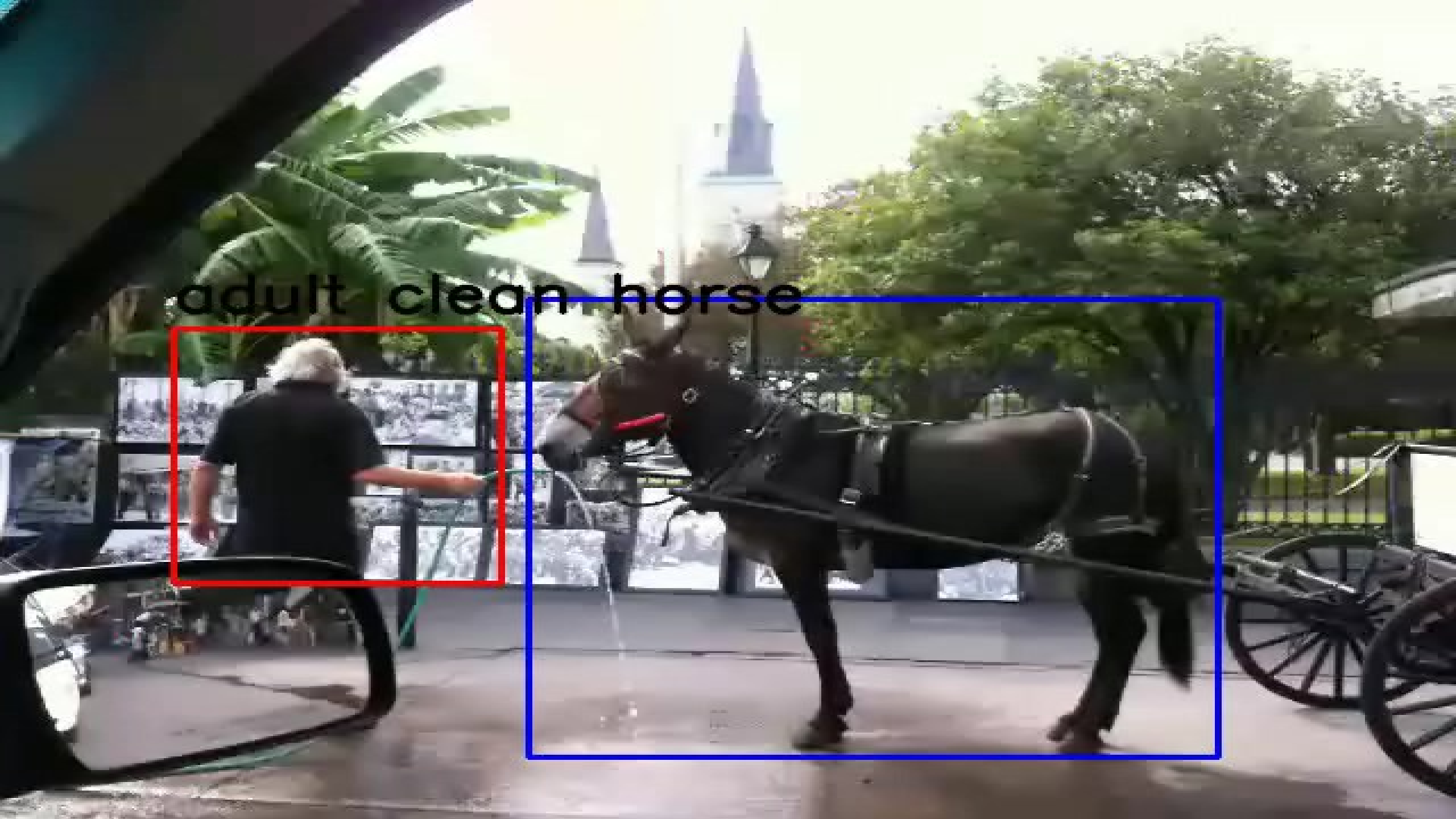
***Further improvements with our video-language feature set***

# Comparison along relation duration on VidOR




***The longer the video relation, the more our performance improves.***


panda chase panda


adult
clean horse


chicken in_front_of child

# Conclusions

- We propose a two-stage method for video relation detection.
- We encode subjects and objects as interactions from the start.
- The social fabric encoding captures shared interaction primitives.
- Experiments show the effectiveness of our method.

Email: s.chen3@uva.nl

https://github.com/shanshuo/Social-Fabric