Hierarchical Vision-Language Alignment for Video Captioning

Junchao Zhang and Yuxin Peng

Institute of Computer Science and Technology

Peking University

Background

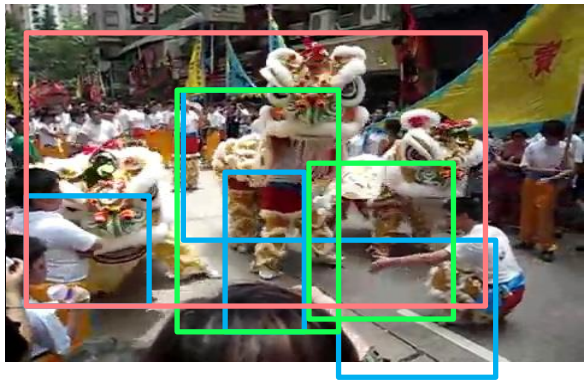
- **Video Captioning**
  - Aims to ***generate natural language sentence*** to describe the input video content
  - Interdisciplinary research across **Computer Vision** and **Natural Language Processing**

Example sentence:

Men in dragon costumes are dancing in the street.

Challenges

- **Challenges are from 2 aspects**
  - **Complex video content understanding**
  - **Vision to language generation**



Object

Visual Relation

Salient Region

...

Vision

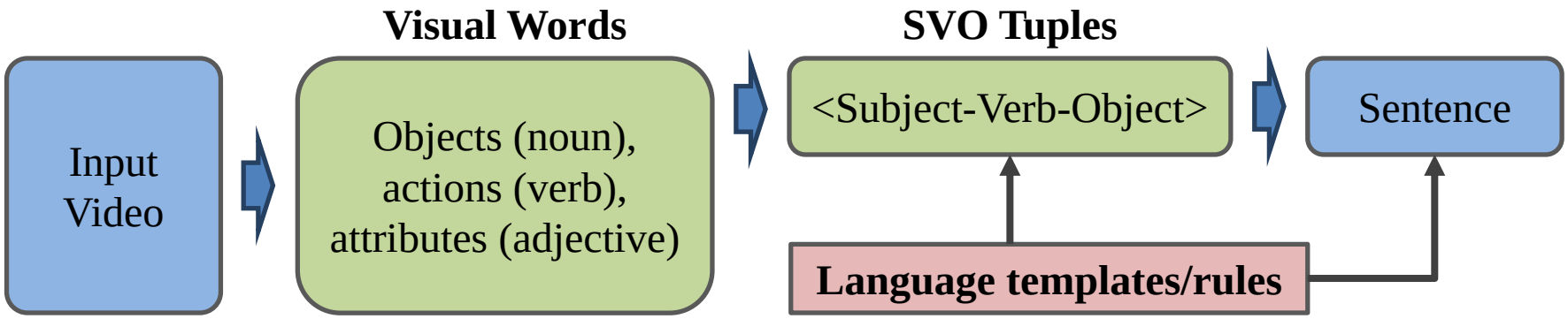
Language

Seek the alignment

Men in dragon costumes are dancing in the street.

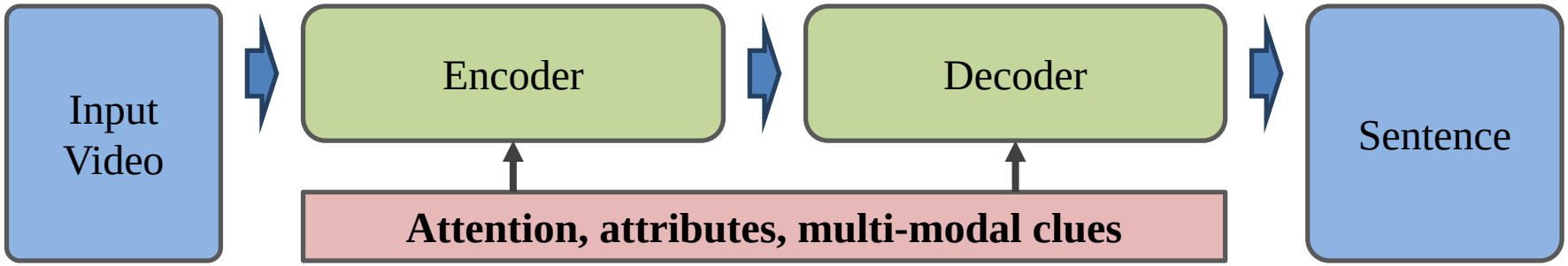
Existing Works

- **Template based language models**
  - Predicting visual words in videos, then relying on ***pre-defined language templates and rules*** to form sentence
  - **Limitation**: The sentences lack of diversities.



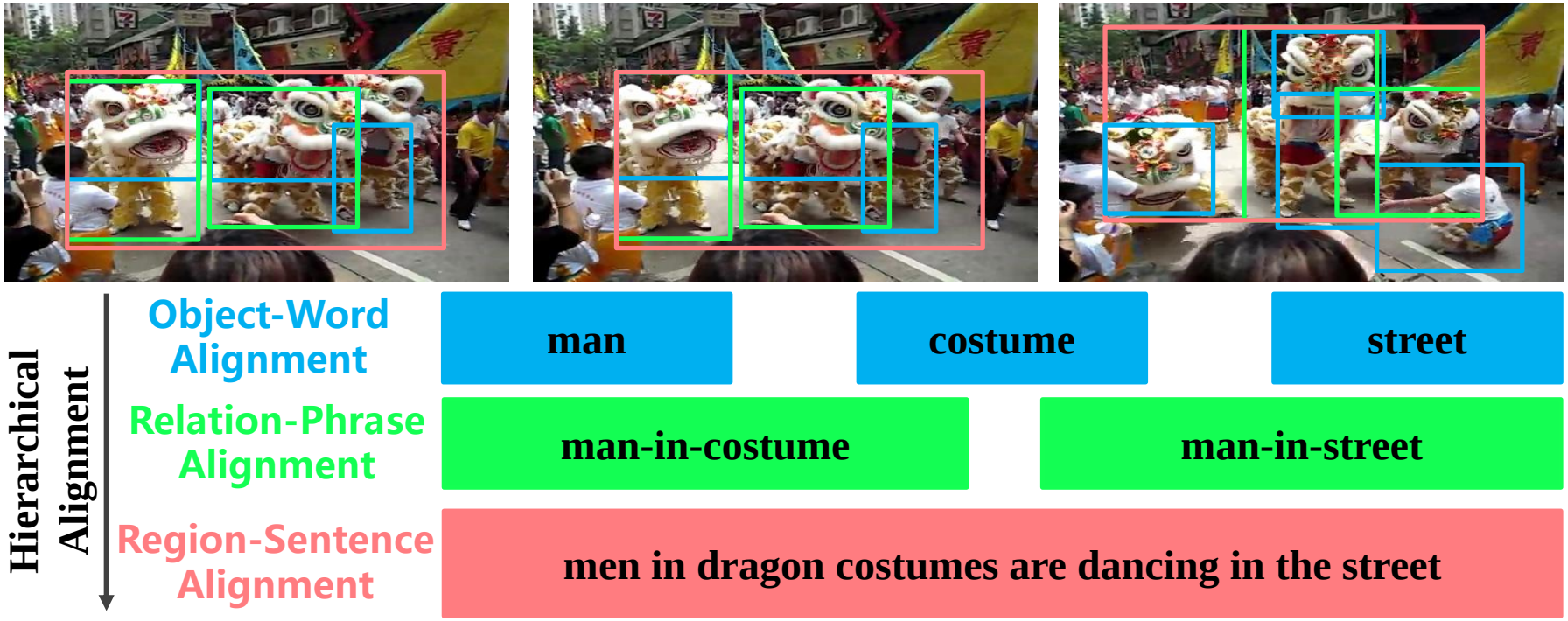
Existing Works

- **Template based language models**
  - Predicting visual words from videos, then relying on ***pre-defined language templates and rules*** to form sentence
  - **Limitation**: The generated sentences lack of diversities.
- **Sequence learning based models**
  - Adopt encoder-decoder framework with ***visual attention***, ***attributes***, etc.
  - **Limitation**: How to ***align*** the visual content to language components remains to be resolved.



Motivation

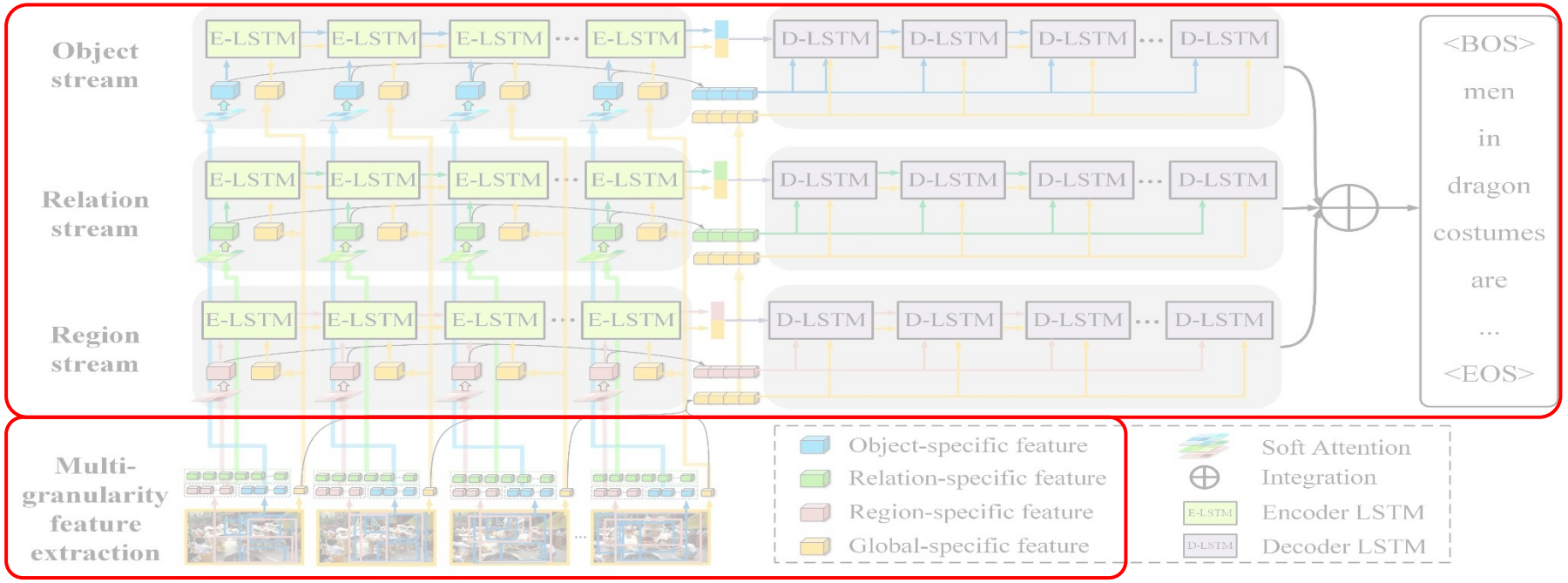
- **Vision-Language Alignments**
  - It's **important** to exploit the **correspondences** between *visual elements* (e.g. object, relation) and *language components* (e.g. word, phrase).



Men in dragon costumes are dancing in the street

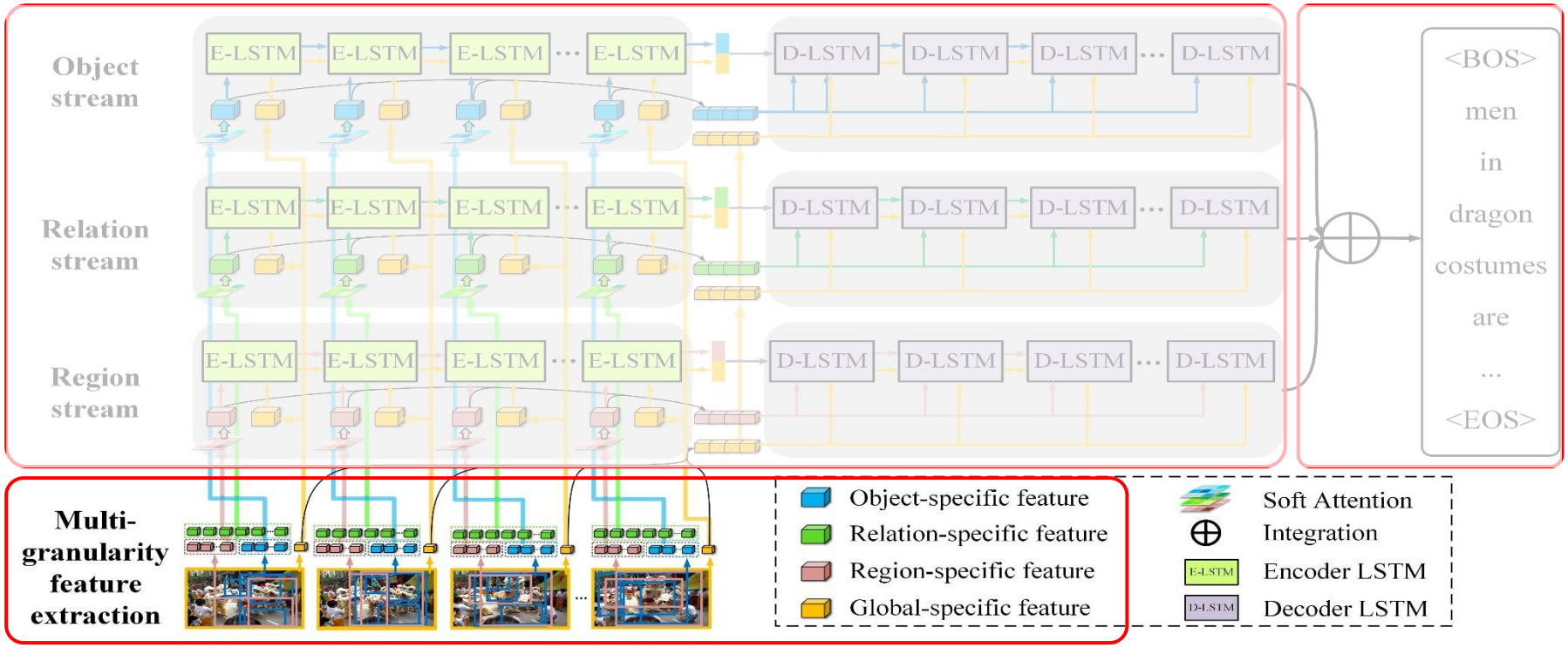
Our Approach: Contributions

- **An Attention Guided Hierarchical Alignment (AGHA) approach**
  - **Hierarchical vision-language alignments** are exploited to provide ***coarse-to-fine guidance*** on video description generation.
  - **Multi-granularity visual features** are exploited to model the complex visual content.



Our Approach: Overview

- **Multi-granularity visual feature extraction**
- **Parallel encoder-decoder streams**
  - Object stream, relation stream and region stream
- **Alignment integration**



Our Approach: Framework (1/5)

- **Multi-granularity visual feature extraction**
  - 4 different granularities of features
  - 4 different specific models as feature extractors

Popular CNN model → **Global** feature

Image Captioning model → **Region-specific** features

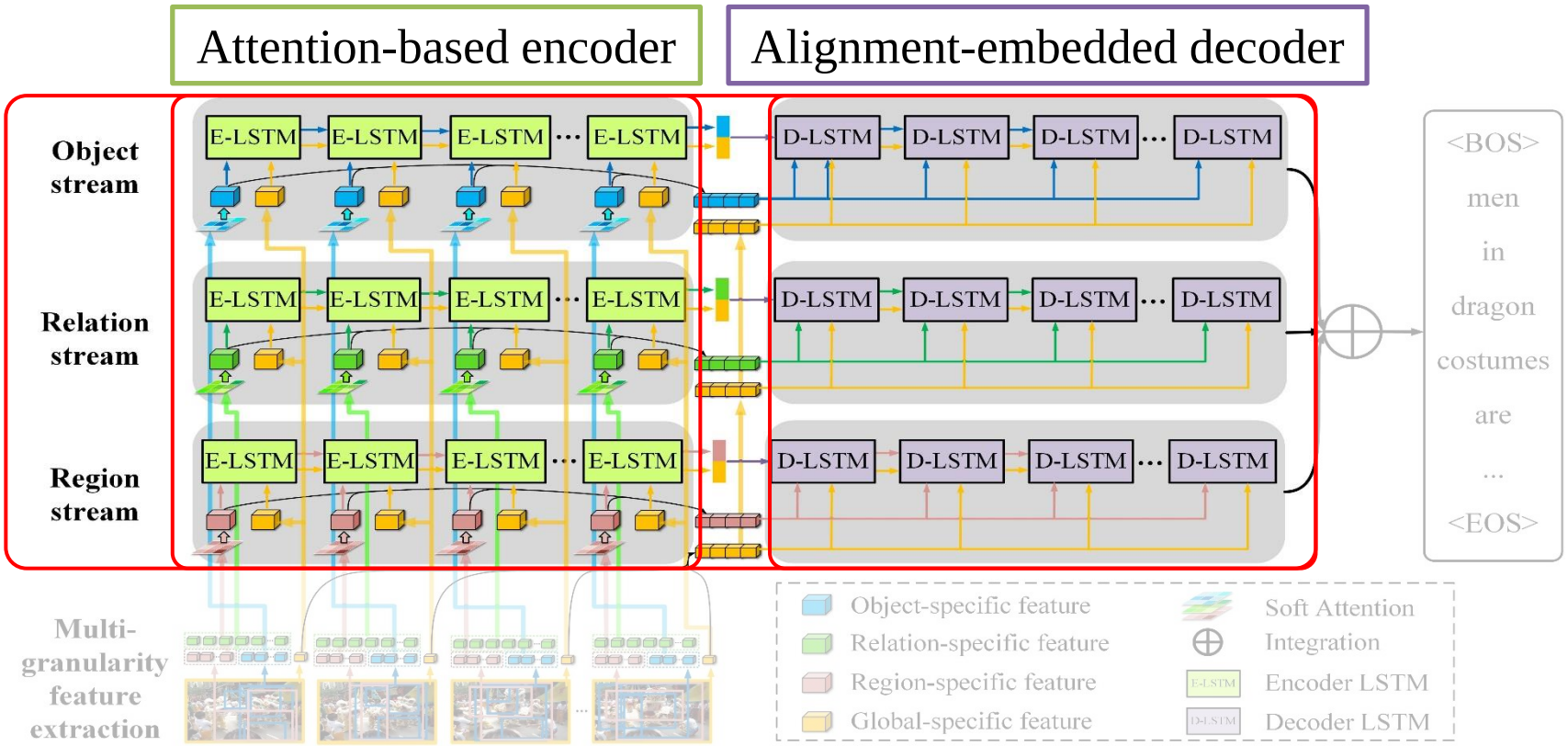
Relation detection model → **Relation-specific** features

Object detection model → **Object-specific** features

Coarse to fine

Our Approach: Framework (2/5)

- **Parallel encoder-decoder streams**
  - Each stream includes an **attention-based encoder** and an **alignment-embedded decoder**



Our Approach: Framework (3/5)

- **Parallel encoder-decoder streams**
  - **Attention-based encoder**
  - ***Soft attention*** is applied to select useful visual elements
  - LSTM with ***double memories and double states***

Double memories

$$c_t^g = f_t \odot c_{t-1}^g + i_t \odot tanh\left(W_{ch} h_{t-1}^g + W_{cx} g_t\right)$$ ← **Global**

$$c_t^A = f_t \odot c_{t-1}^A + i_t \odot tanh\left(U_{ch} h_{t-1}^A + U_{cx} A_t^\alpha\right)$$ ← **Object/ Relation/ Region**

Double states

$$h_t^g = o_t \odot tanh\left(c_t^g\right)$$ ← **Global**

$$h_t^A = o_t \odot tanh\left(c_t^A\right)$$ ← **Object/ Relation/ Region**

Our Approach: Framework (4/5)

- **Parallel encoder-decoder streams**
  - **Alignment-embedded decoder**
  - LSTM with ***specific alignment information***

Hidden state $h_t = o_t \odot tanh(c_t)$

Memory $c_t = f_t \odot c_{t-1} + i_t \odot tanh(\varphi_c)$

Object stream $\varphi_c = W_{ch}h_{t-1} + W_{cx}x_t + U_c^g g_t^\beta + U_c^{obj} obj_t^\beta$

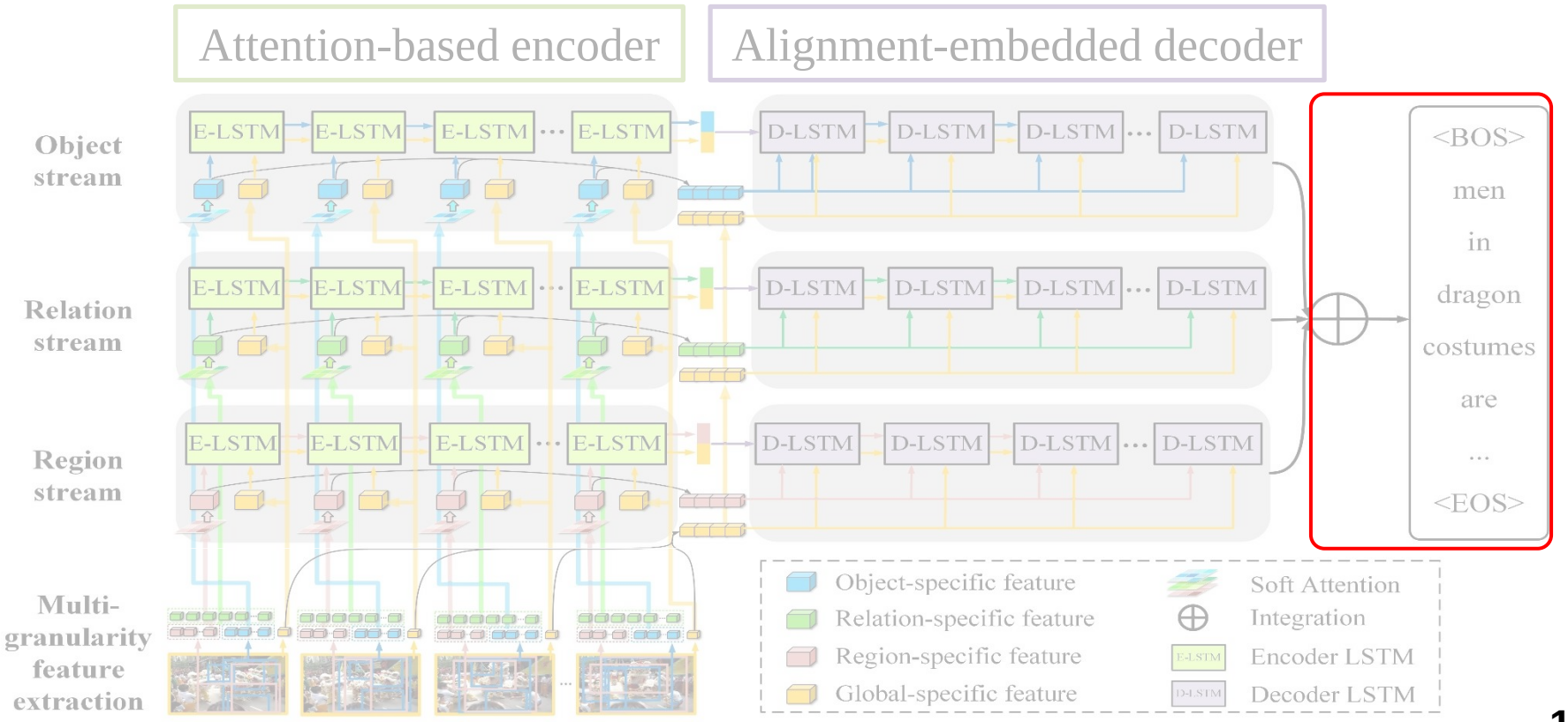
Global information ***Object-word alignment***

Relation stream $\varphi_c = W_{ch}h_{t-1} + W_{cx}x_t + U_c^g g_t^\beta + U_c^{obj} rel_t^\beta$

Object-word, relation-phrase and ***region-sentence alignments*** are exploited respectively to provide guidance on sentence generation

Our Approach: Framework (5/5)

- **Alignment integration**
  - Integration is performed on three streams to exploit ***complementarities*** among different vision-language alignments



Experiment (1/6)

- **Dataset: Microsoft Video Description Corpus (MSVD)**
  - 1,970 video clips from Youtube
  - 8,000 English descriptions, with roughly 40 descriptions per video
  - 1,200 clips for training, 100 clips for validation, and 670 clips for testing
- **Metrics**
  - Totally **6 widely-used metrics** included
  - BLEU@N: BLEU@1~4 (Denoted as B@1~4 for short)
  - METEOR (Denoted as M for short)
  - CIDEr (Denoted as C for short)

Experiment (2/6)

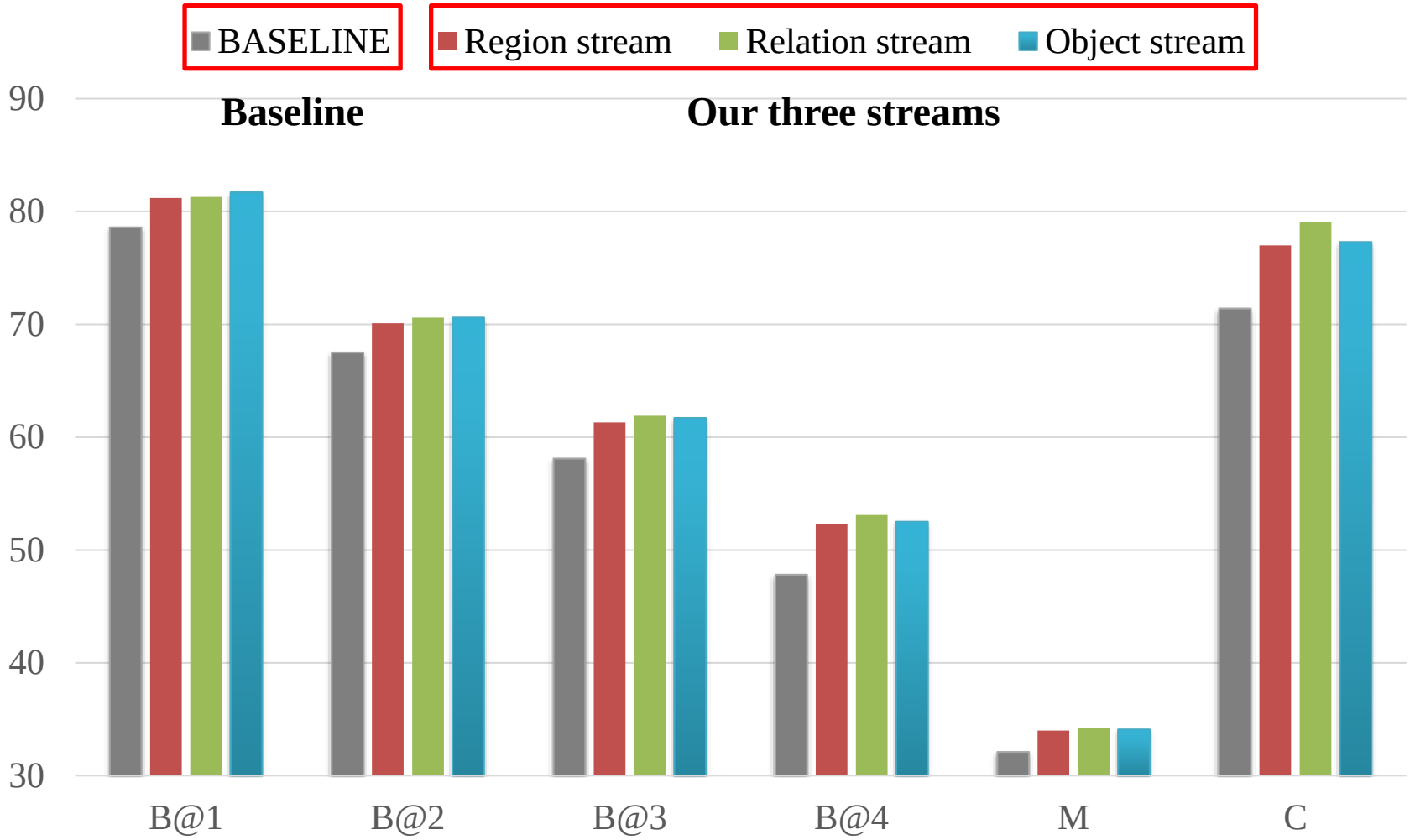
- **Comparisons with 10 state-of-the-art methods**

Methods	B@1	B@2	B@3	B@4	M	C
Our AGHA	**83.1**	**73.0**	**64.3**	**55.1**	**35.3**	**83.3**
RecNet [CVPR 2018]	-	-	-	52.3	34.1	80.3
MCNN+MCF [IJCAI 2018]	-	-	-	46.5	33.7	75.5
M&M-TGM [ACM MM 2017]	-	-	-	48.8	34.4	80.5
MA-LSTM [ACM MM 2017]	82.3	71.1	61.8	52.3	33.6	70.4
DMRM [ACM MM 2017]	-	-	-	51.1	33.6	74.8
LSTM-TSA [CVPR 2017]	82.8	72.0	62.8	52.8	33.5	74.0
TDDF [CVPR 2017]	-	-	-	45.8	33.3	73.0
mGRU [CVPR 2016]	82.5	72.2	63.3	53.8	34.5	81.2
HRNE [CVPR 2016]	79.2	66.3	55.1	43.8	33.1	-
h-RNN [CVPR 2016]	81.5	70.4	60.4	49.9	32.6	65.8

"-" means the corresponding result is not provided in the paper of compared method.

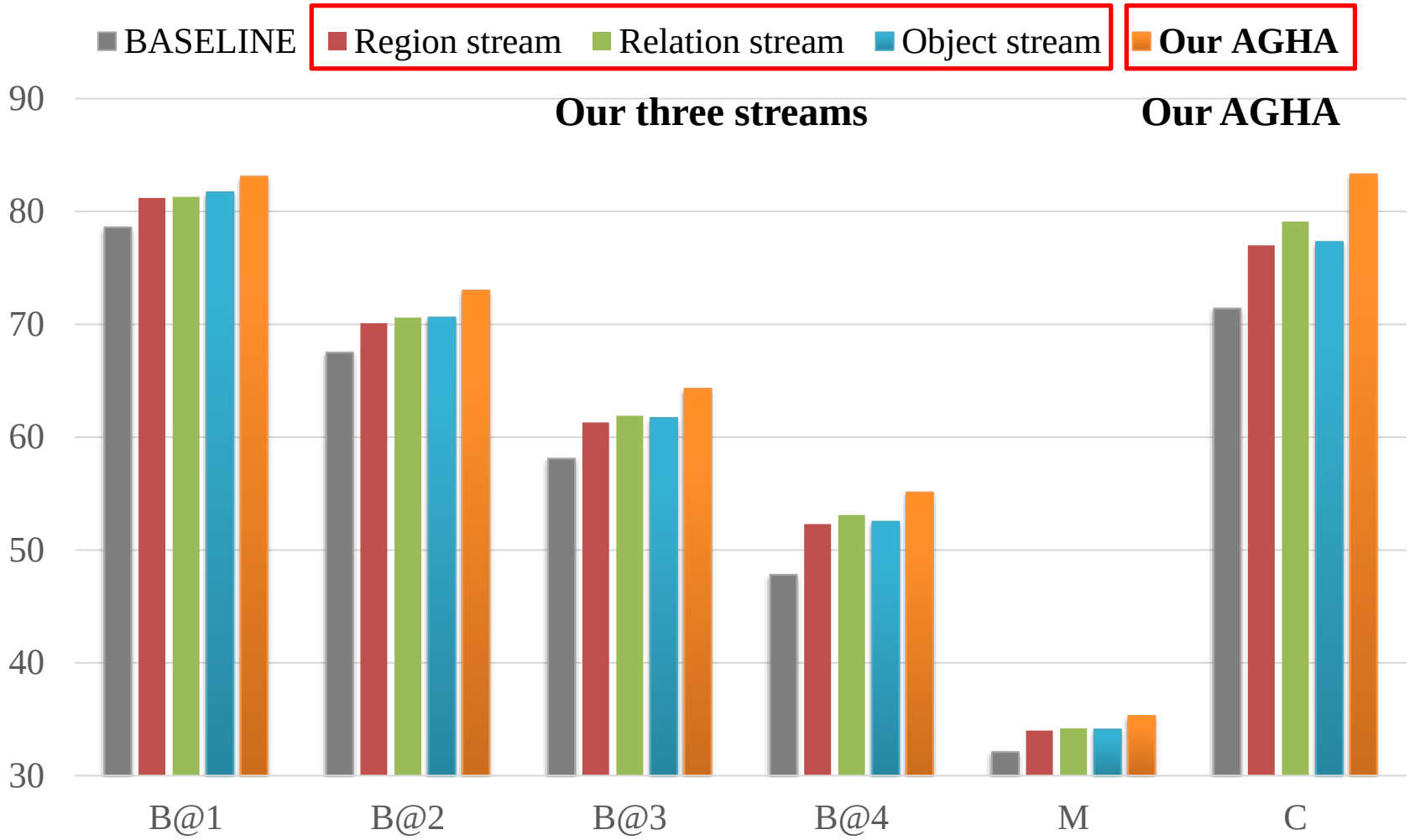
Experiment (3/6)

- **Ablation study**



Experiment (4/6)

- **Ablation study**



Experiment (5/6)

- **Qualitative Analysis**
  - Green indicates correct words, while red indicates the wrong words

DMRM (ACM MM 2017):
a man is running

Our AGHA:
a man is playing football

DMRM (ACM MM 2017):
a woman is cracking eggs

AGHA:
a woman is mixing ingredients in a bowl

Experiment (6/6)

- **Qualitative Analysis**
  - Green indicates correct words, while red indicates the wrong

DMRM (ACM MM 2017):
a man is talking to another man

Our AGHA:
a group of people are playing music

DMRM (ACM MM 2017):
people are dancing

AGHA:
a group of men are fighting

Conclusion

- **Conclusion**
  - We propose an attention guided hierarchical alignment (AGHA) approach for video captioning.
  - It exploits hierarchical vision-language alignments and mines their complementarities to capture the ***semantic correspondences*** between visual content and language sentence.
  - It also explores multi-granularity visual features to capture ***coarse-to-fine visual information*** for complex video content understanding.
- **Future work**
  - To explore the ***interactions*** among different vision-language alignments
  - To employ ***one-shot or few-shot learning*** to use less training data

Thank you!

MMM 2019

Lab Homepage

Github Homepage

Multimedia Information Processing Lab (MIPL)

http://www.icst.pku.edu.cn/mipl/