

# Early Embedding and Late Reranking for Video Captioning

Jianfeng Dong<sup>1</sup>, **Xirong Li**<sup>2</sup>, Weiyu Lan<sup>2</sup>, Yujia Huo<sup>2</sup>, Cees G. M. Snoek<sup>3</sup>

Zhejiang University<sup>1</sup>

Renmin University of China<sup>2</sup>


University of Amsterdam<sup>3</sup>

# Live demo

<http://lixirong.net/demo/vtt>

## Video-to-Text

Upload Video



0:00 / 0:13

a fashion model is walking  
down a runway  
Tags: model, runway, walking,  
woman

How is the generated sentence?

- 👍 good
- 👉 just so so
- 👎 bad

How would you describe this video?

...

提交

# Re-use Video Tags for Captioning

## Predicted tags

## Generated caption



track  
race  
field  
woman

a group of people are running in a  
**race track**



soccer  
player  
game  
playing

a **soccer player** is **playing** a goal on a  
soccer field

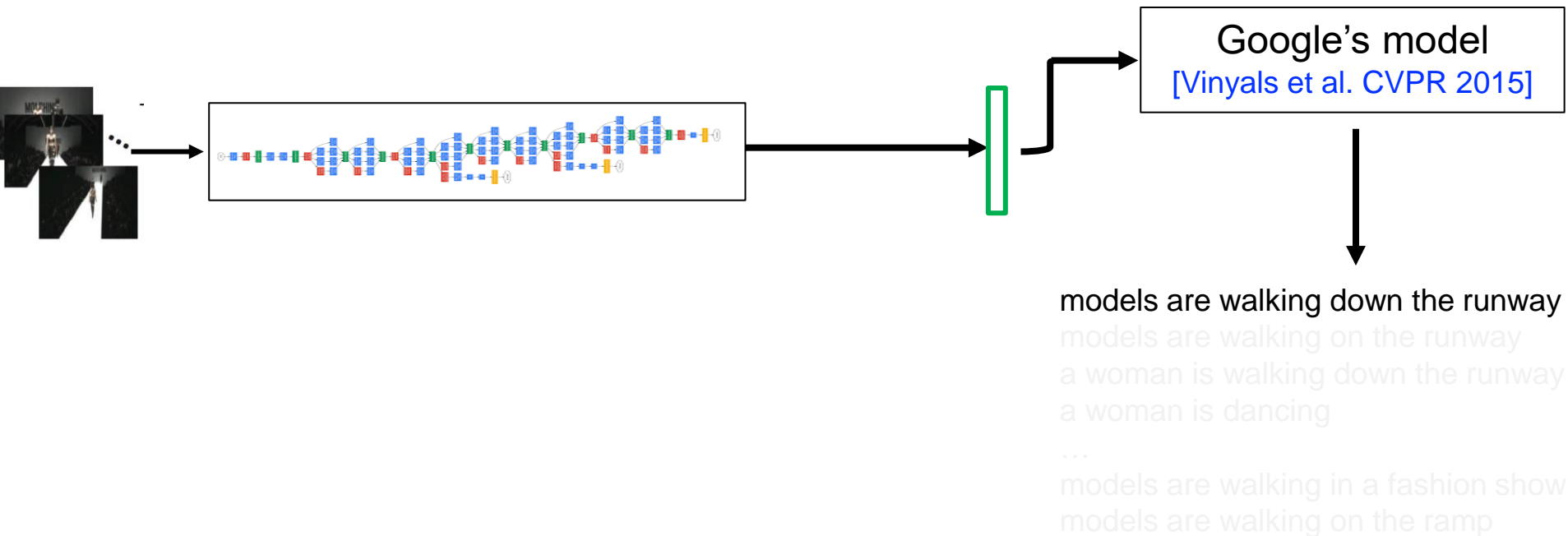


dance  
people  
woman  
dancing

**people** are **dancing** on a stage

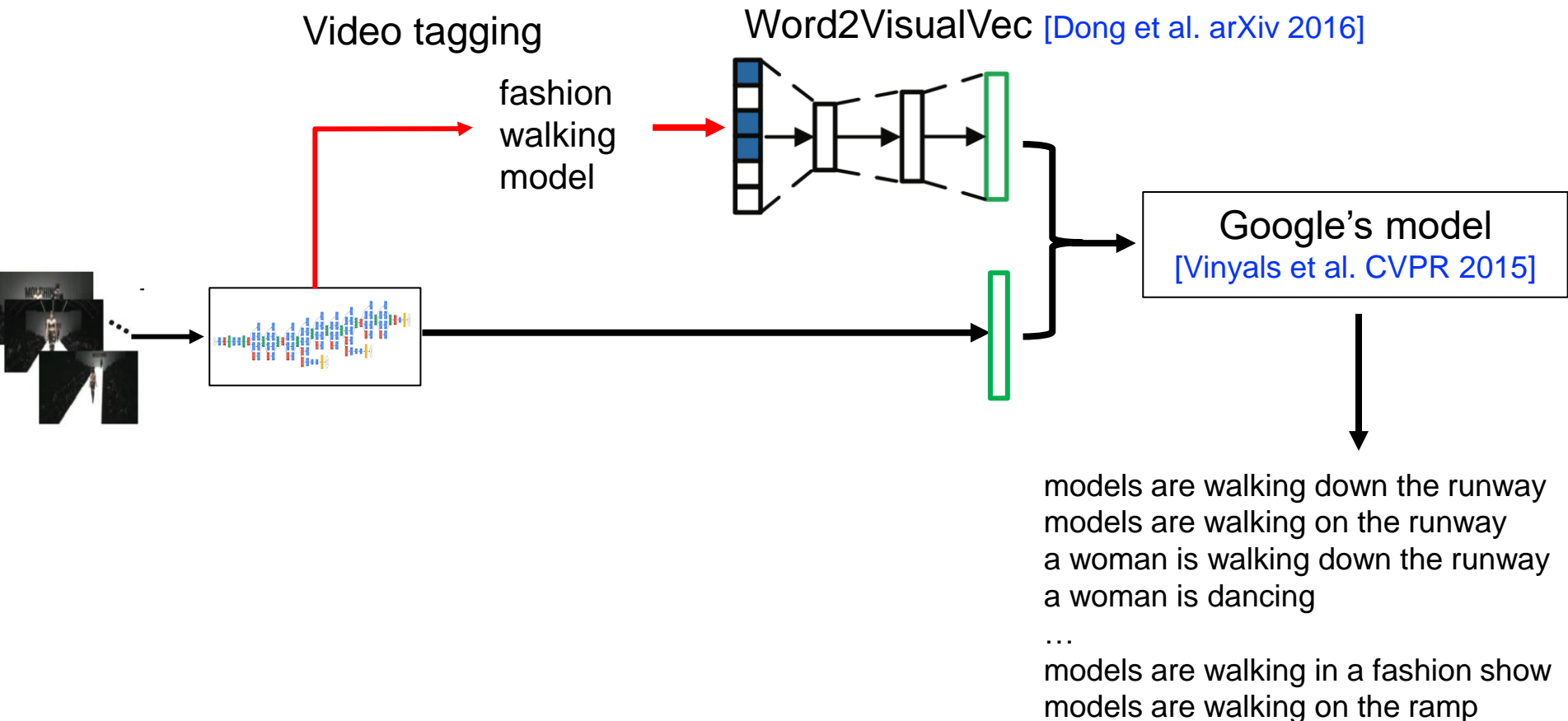
# Proposed Video Captioning System

Google's model for sentence generation



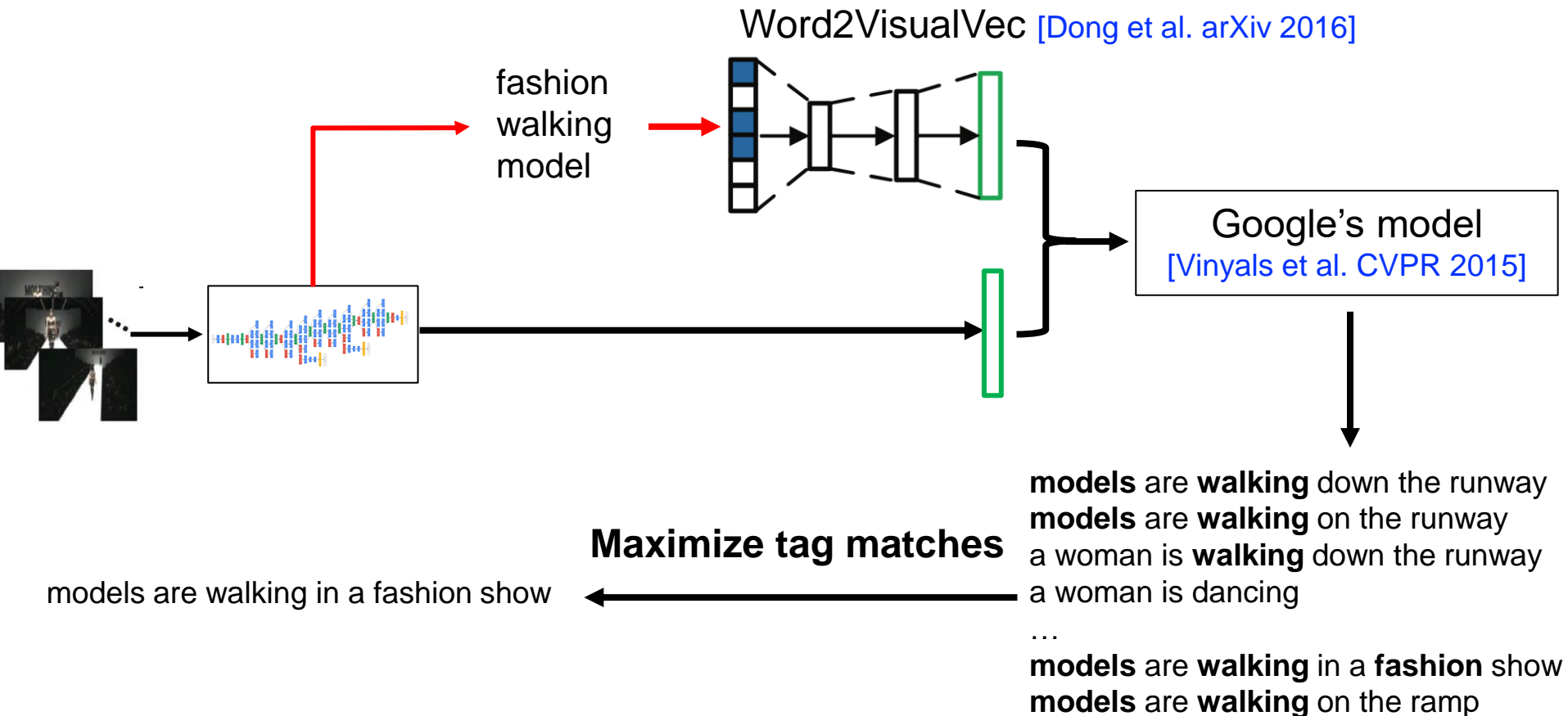
# Proposed Video Captioning System

Better initialization by early embedding



# Proposed Video Captioning System

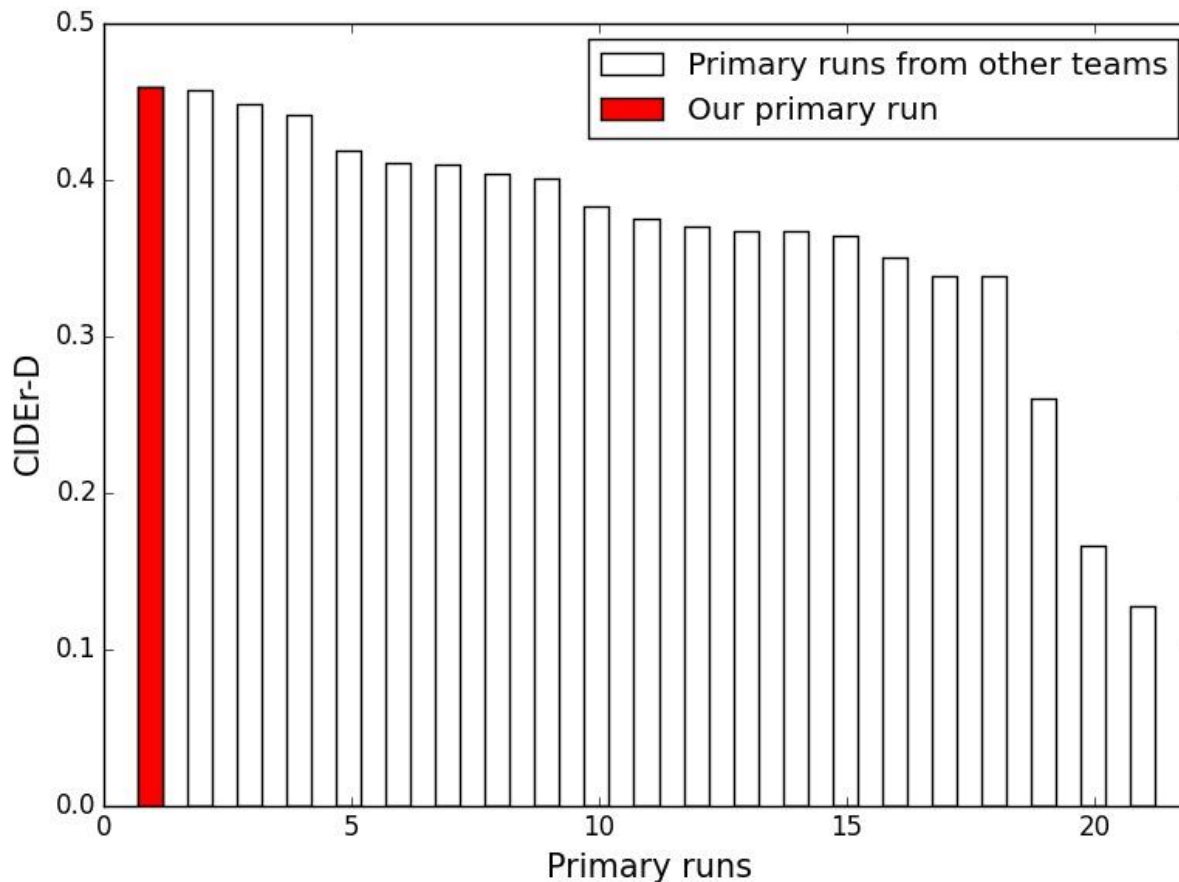
Rerank sentences by matching with video tags



# Official Evaluation

## Best CIDEr-D

measuring human-likeness of generated captions



# Conclusion

Early embedding and Late reranking  
improves LSTM based video captioning

Word2VisualVec plus our winning TRECVID Video-to-Text results  
highlighted in Rising Star Symposium