

# The Role of Syntactic Planning in Compositional Image Captioning

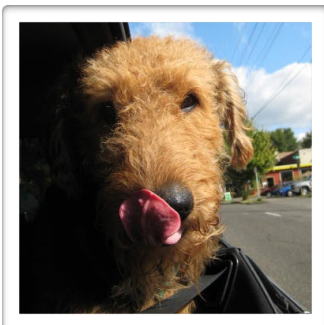
**Emanuele Bugliarello**, Desmond Elliott  
University of Copenhagen

EACL 2021



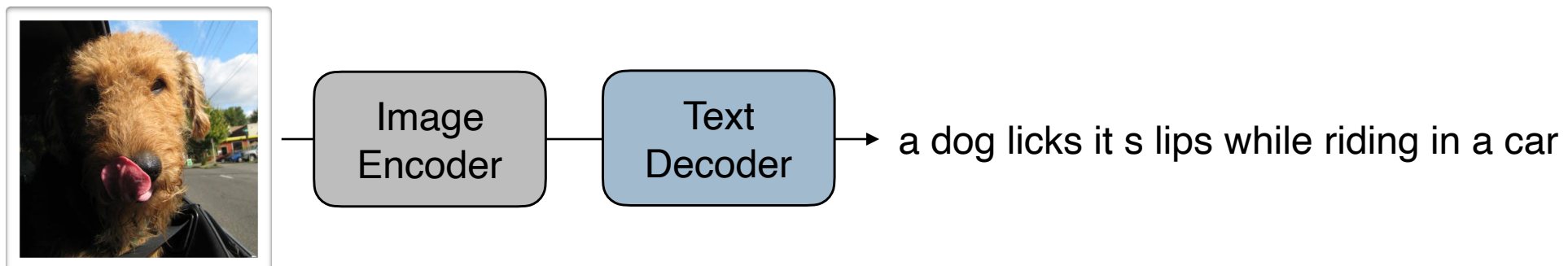
UNIVERSITY OF  
COPENHAGEN

# Image Captioning

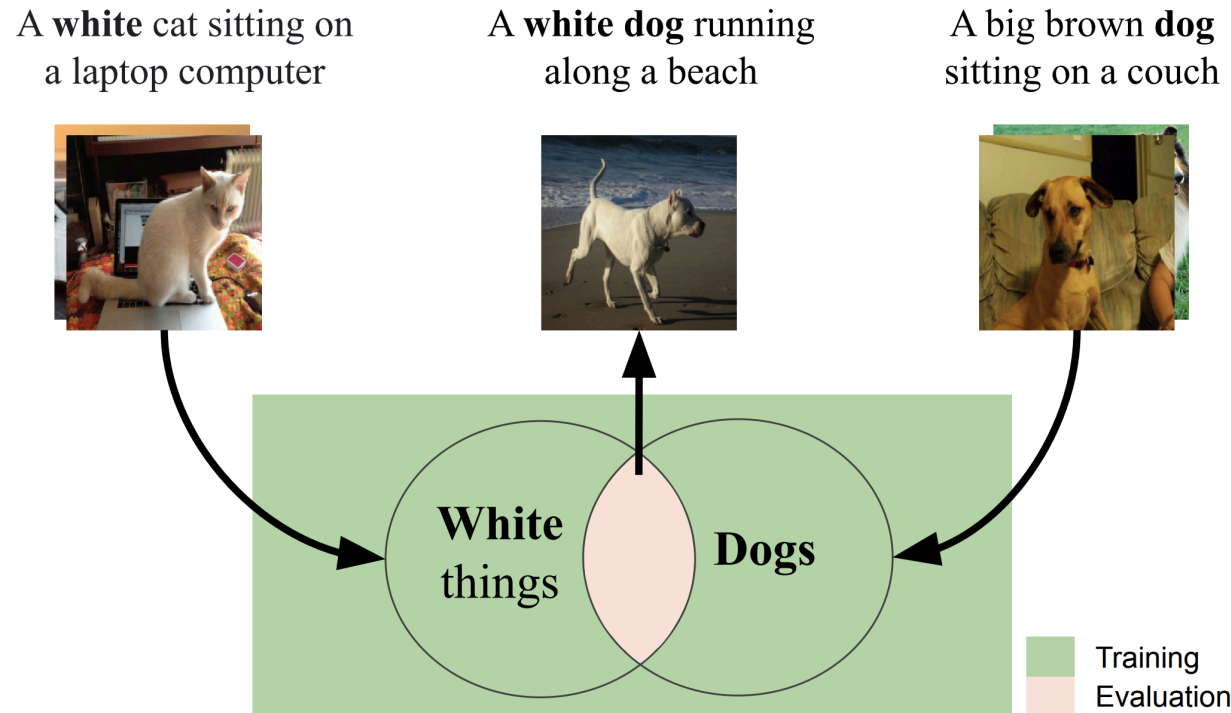


→ a dog licks it s lips while riding in a car

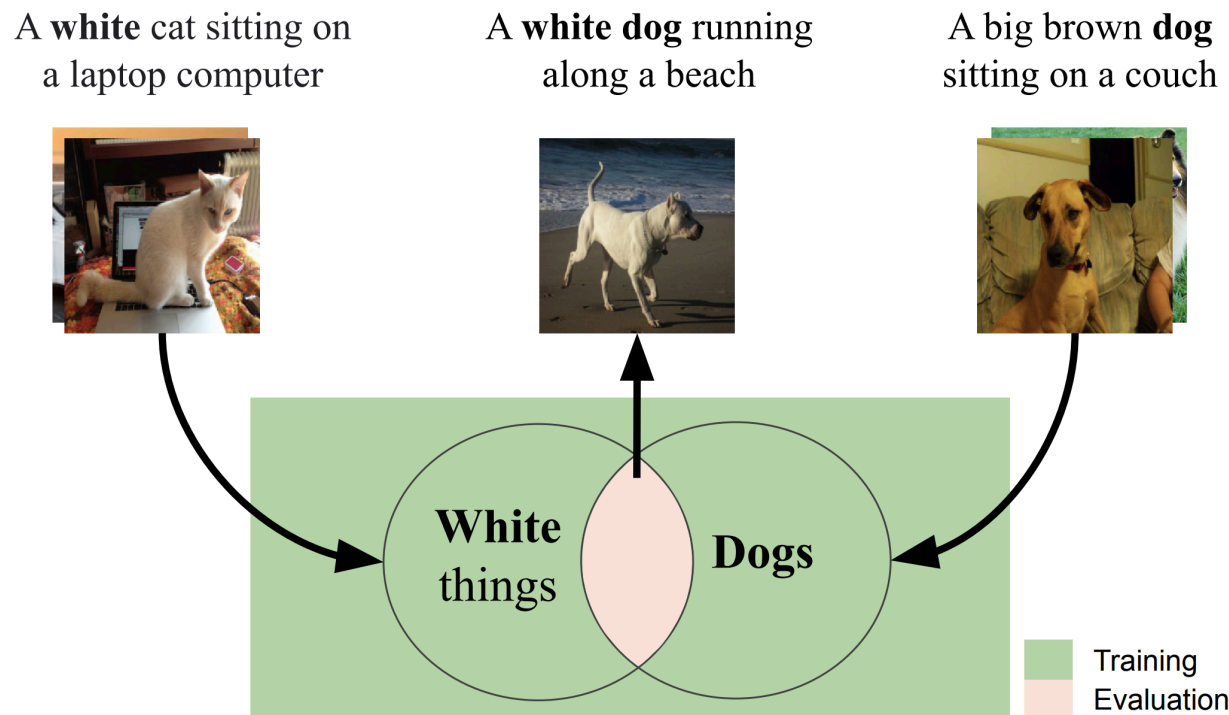
# Image Captioning



# Compositional Image Captioning (Nikolaus et al., 2019)

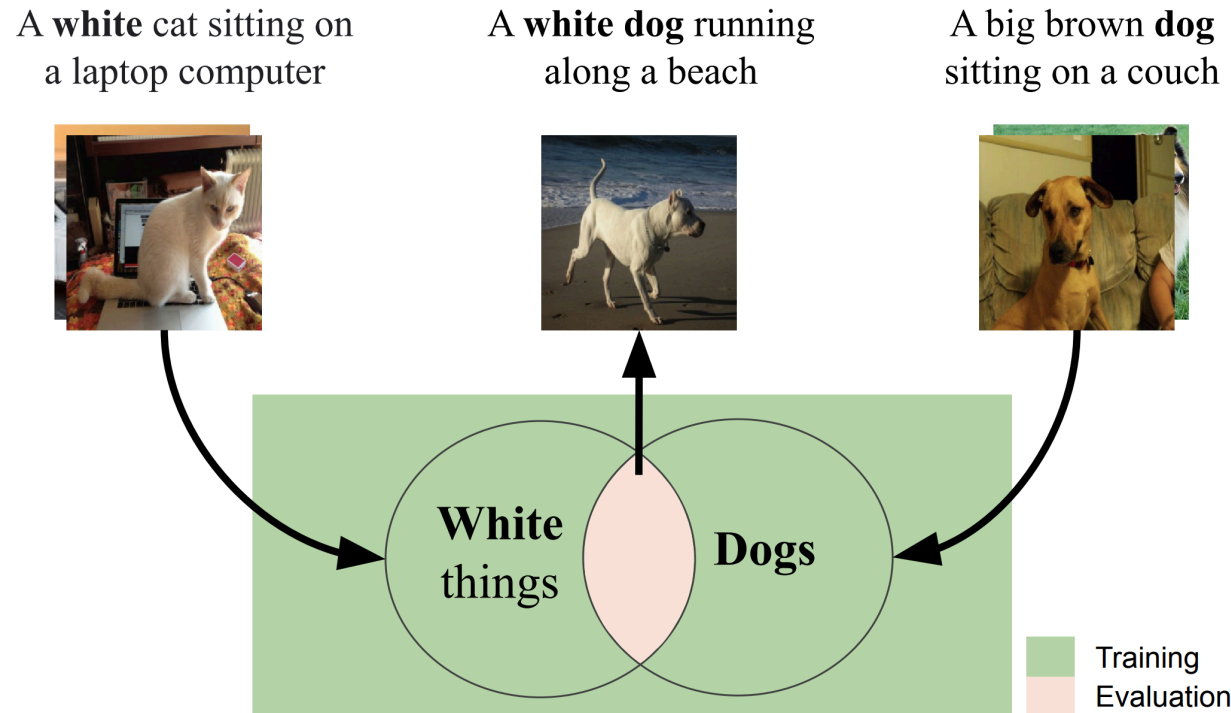


# Compositional Image Captioning (Nikolaus et al., 2019)



RNN-based captioning models do not compositionally generalise

# Compositional Image Captioning (Nikolaus et al., 2019)

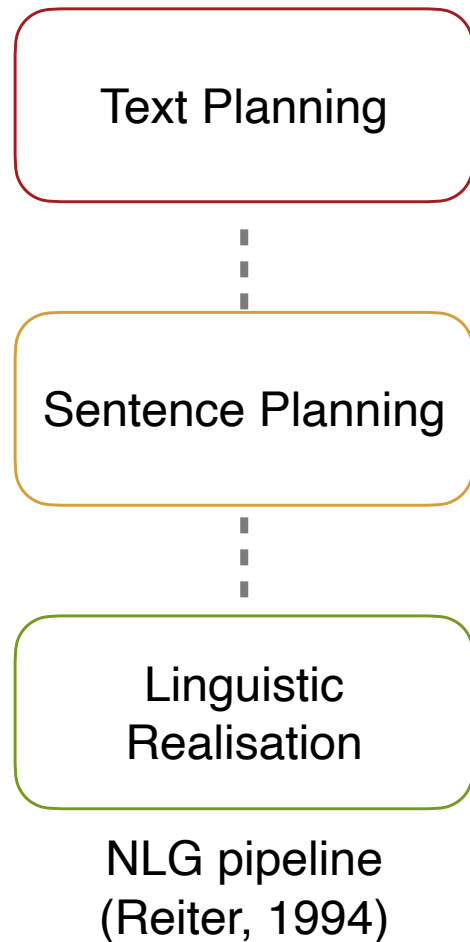


RNN-based captioning models do not compositionally generalise

Due to the text decoder

# Syntactic Planning for Compositional Generalisation

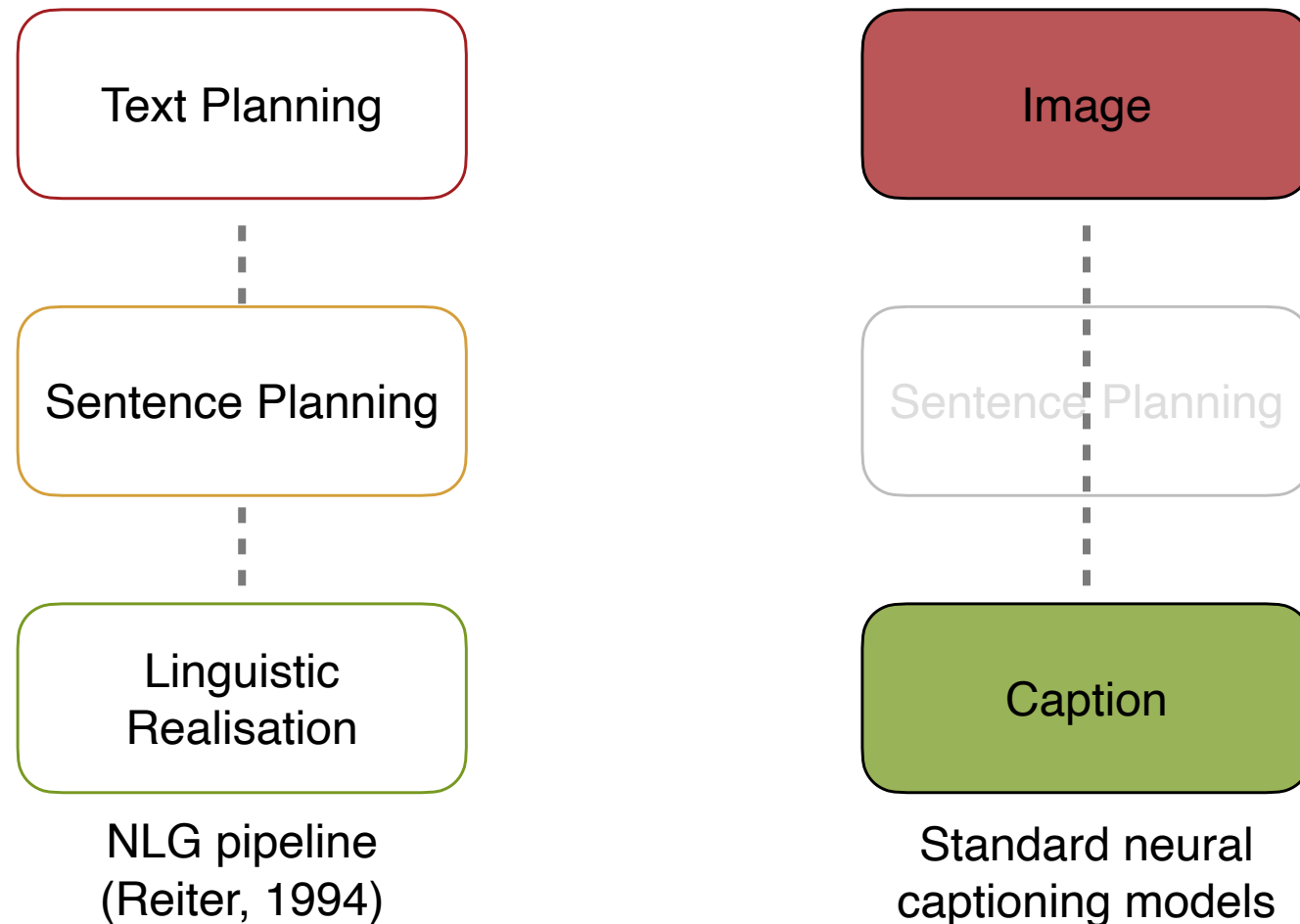
# Syntactic Planning for Compositional Generalisation



Reiter. *Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?* INLG 1994.

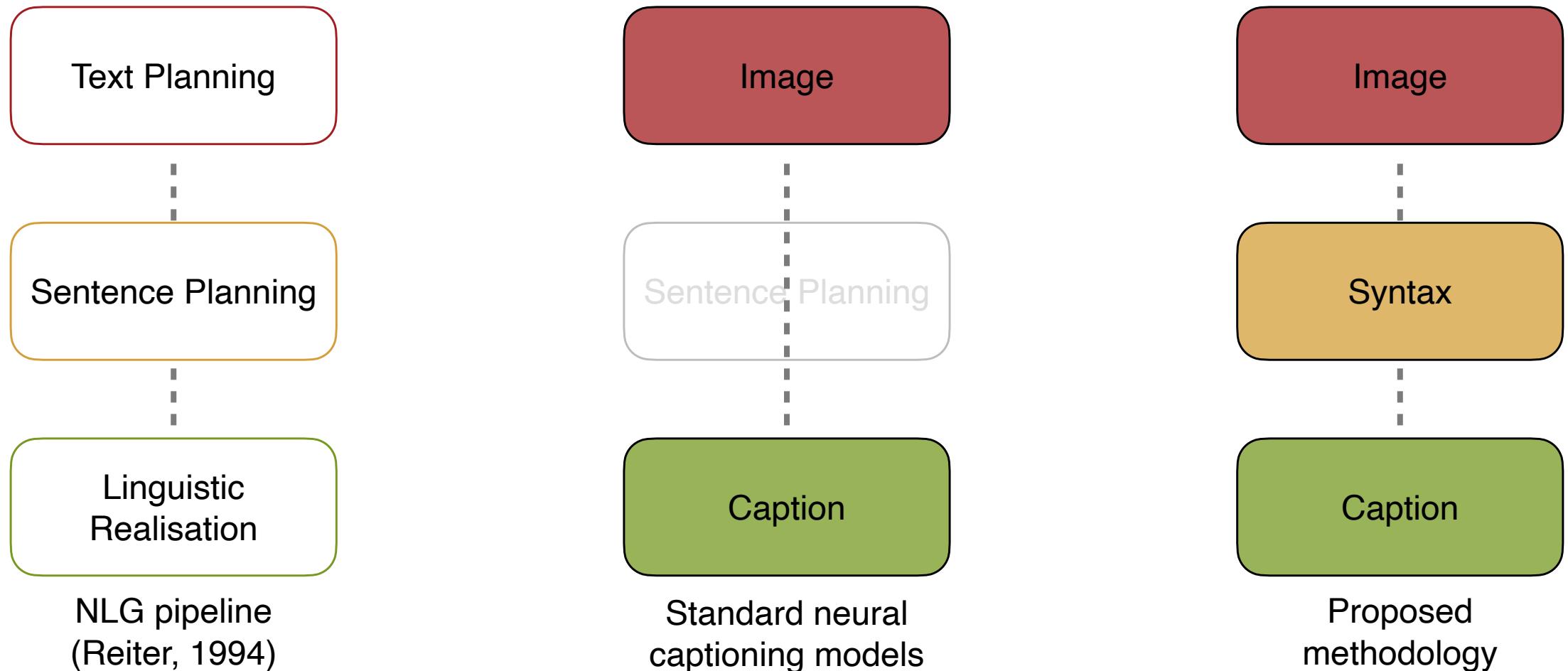


# Syntactic Planning for Compositional Generalisation



Reiter. *Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?* INLG 1994.

# Syntactic Planning for Compositional Generalisation



Reiter. *Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?* INLG 1994.

# Syntactic Granularity & Planning Approaches

# Syntactic Granularity & Planning Approaches

CHUNK

POS

DEP

CCG

# Syntactic Granularity & Planning Approaches

CHUNK

POS

DEP

CCG

IDLE

# Syntactic Granularity & Planning Approaches



Standard  
Decoder

<S> a dog licks it s lips while riding in a car <E>

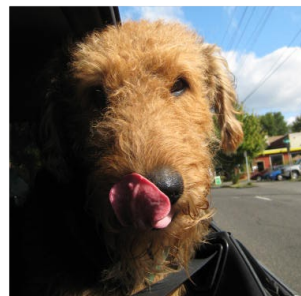


Image  
Encoder

# Syntactic Granularity & Planning Approaches

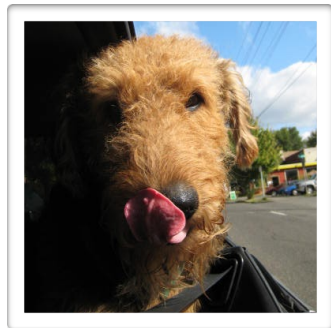


Image  
Encoder

Standard  
Decoder

<S> a dog licks it s lips while riding in a car <E>

Sequential  
Decoder

<S> DET NOUN VERB PRON AUX NOUN SCONJ VERB ADP DET NOUN <T>  
a dog licks it s lips while riding in a car <E>

# Syntactic Granularity & Planning Approaches

CHUNK

POS

DEP

CCG

IDLE

Standard  
Decoder

<S> a dog licks it s lips while riding in a car <E>

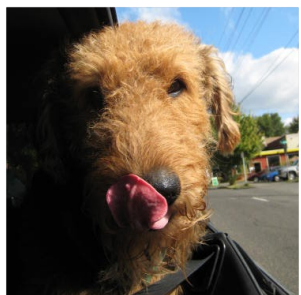
Sequential  
Decoder

<S> DET NOUN VERB PRON AUX NOUN SCONJ VERB ADP DET NOUN <T>  
a dog licks it s lips while riding in a car <E>

Image  
Encoder

Interleave  
Decoder

<S> DET **a** NOUN **dog** VERB **licks** PRON **it** AUX **s** NOUN **lips** SCONJ  
**while** VERB **riding** ADP **in** DET **a** NOUN **car** <E>





# Syntactic Granularity & Planning Approaches

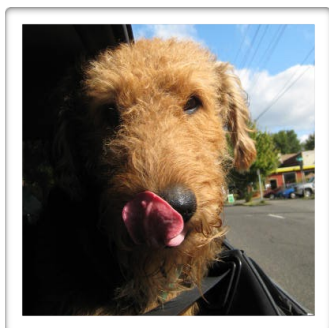
CHUNK

POS

DEP

CCG

IDLE

Image  
EncoderStandard  
Decoder

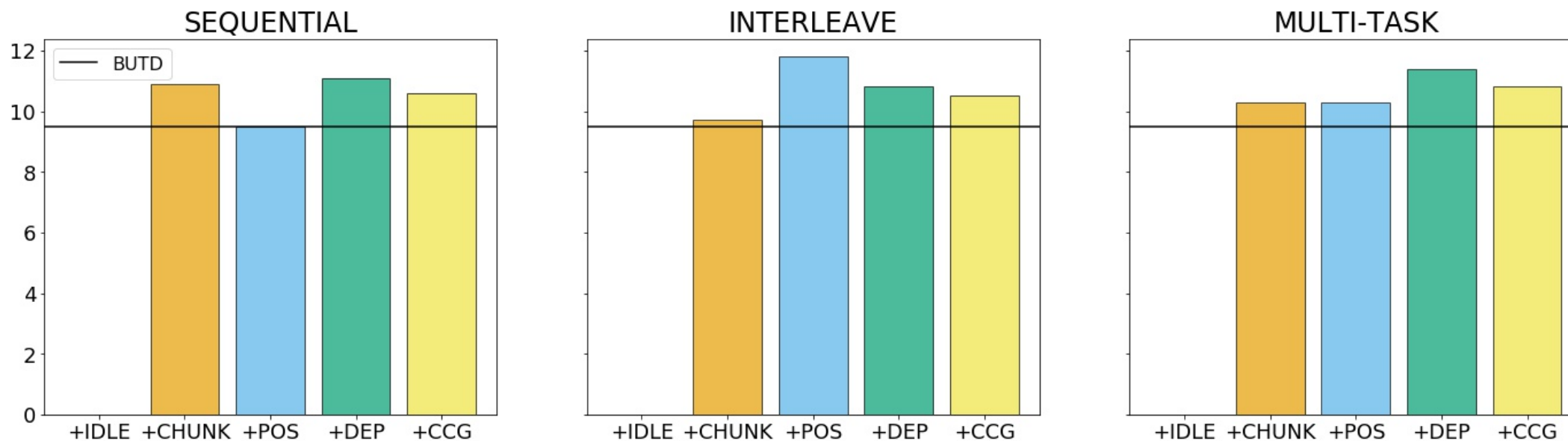
&lt;S&gt; a dog licks it s lips while riding in a car &lt;E&gt;

Sequential  
Decoder<S> DET NOUN VERB PRON AUX NOUN SCONJ VERB ADP DET NOUN <T>  
a dog licks it s lips while riding in a car <E>Interleave  
Decoder<S> DET a NOUN dog VERB licks PRON it AUX s NOUN lips SCONJ  
while VERB riding ADP in DET a NOUN car <E>Multi-task  
Decoder

&lt;S&gt; a dog licks it s lips while riding in a car &lt;E&gt;

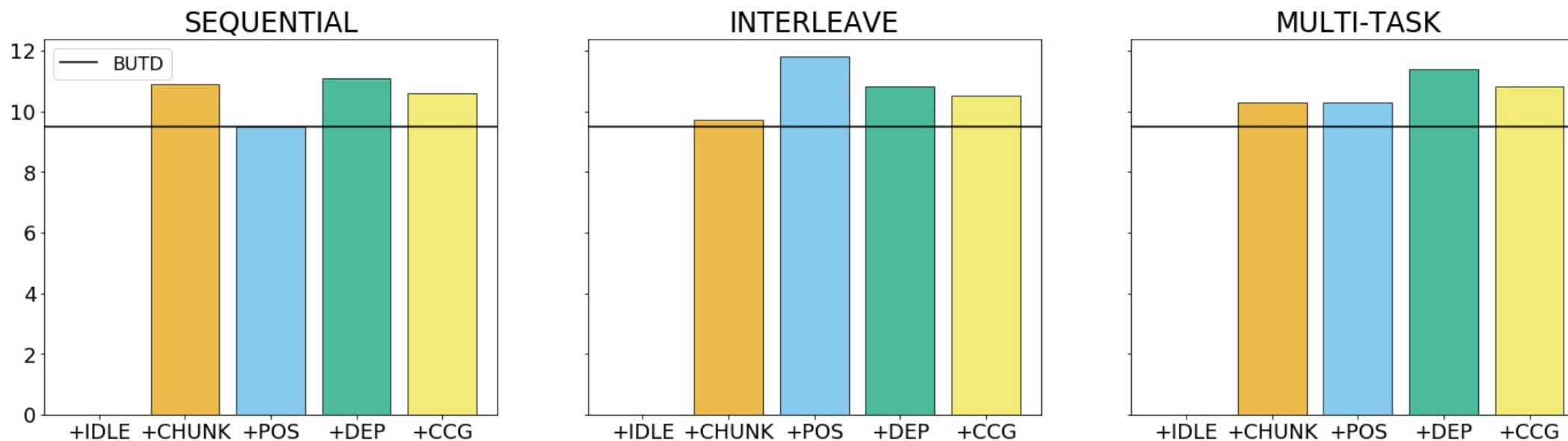
&lt;T&gt; DET NOUN VERB PRON AUX NOUN SCONJ VERB ADP DET NOUN &lt;E&gt;

# Syntax Awareness



Average Recall@5 of unseen concepts by BUTD (Anderson et al., 2018)

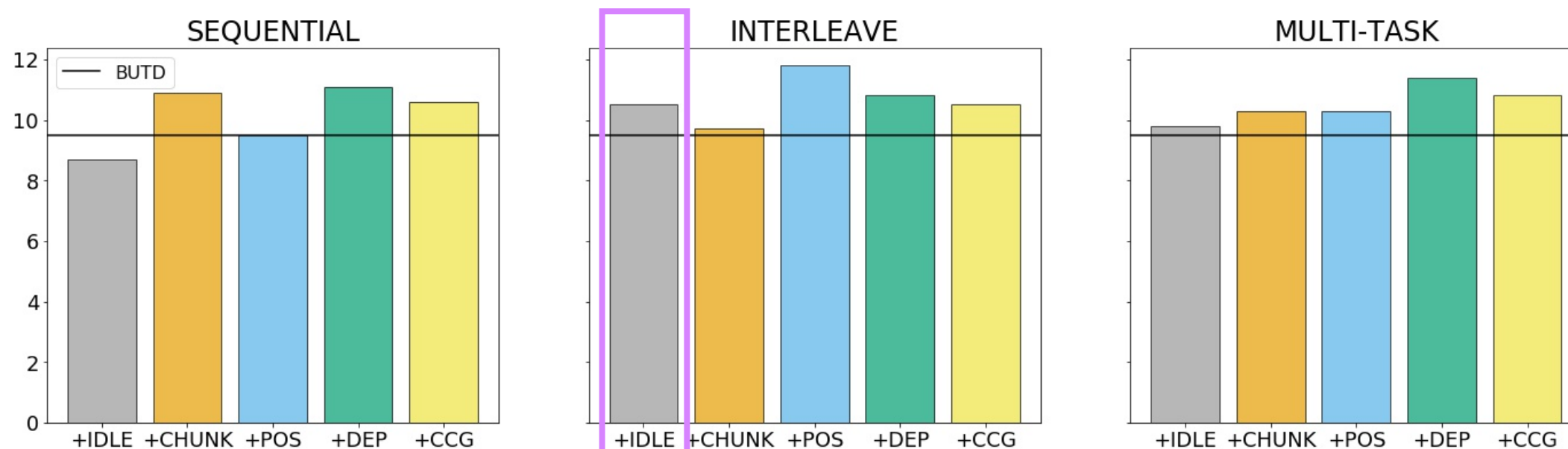
# Syntax Awareness



Average Recall@5 of unseen concepts by BUTD (Anderson et al., 2018)

Syntactic planning helps compositional image captioning

# Syntax Awareness



Average Recall@5 of unseen concepts by BUTD (Anderson et al., 2018)

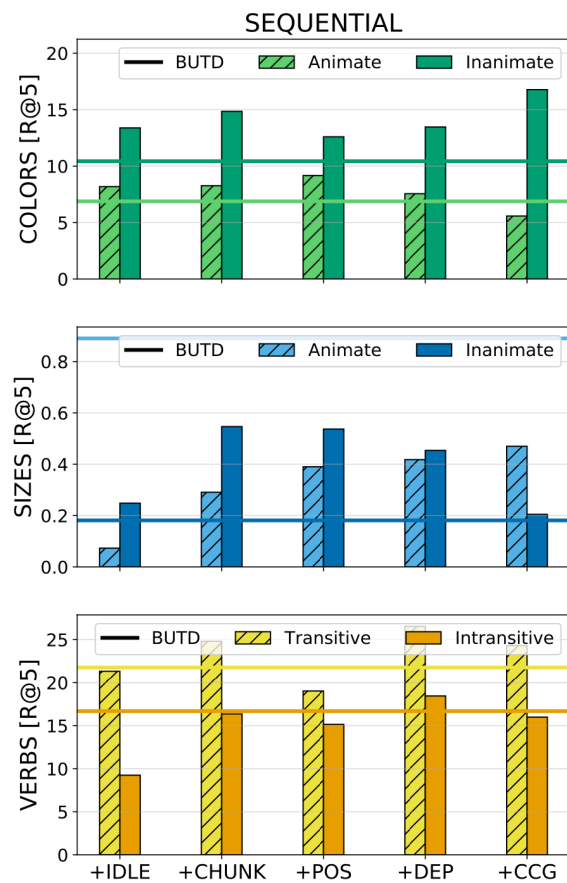
Syntactic planning helps compositional image captioning

Directly mapping an image onto words is sub-optimal

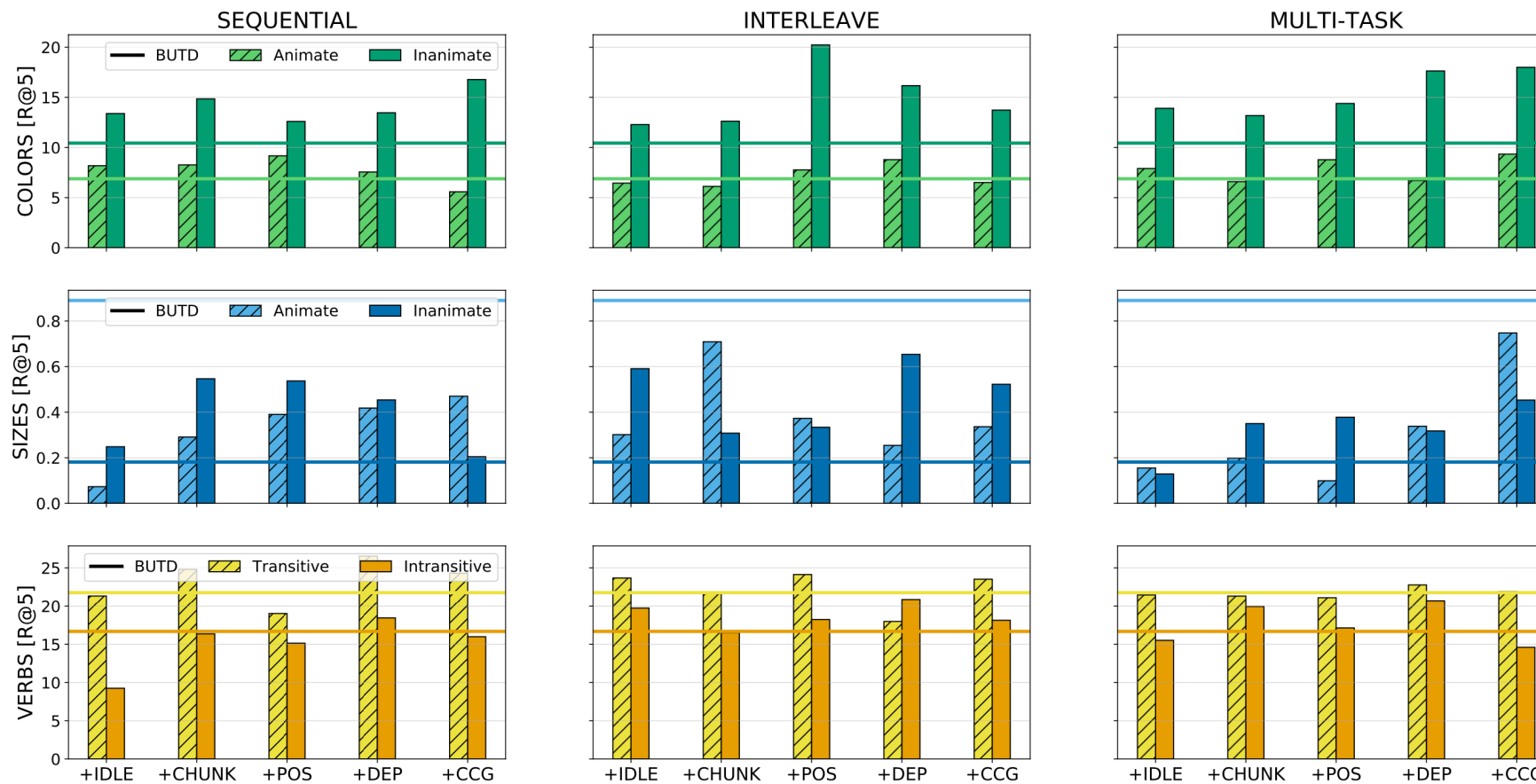


# Generalisation across categories

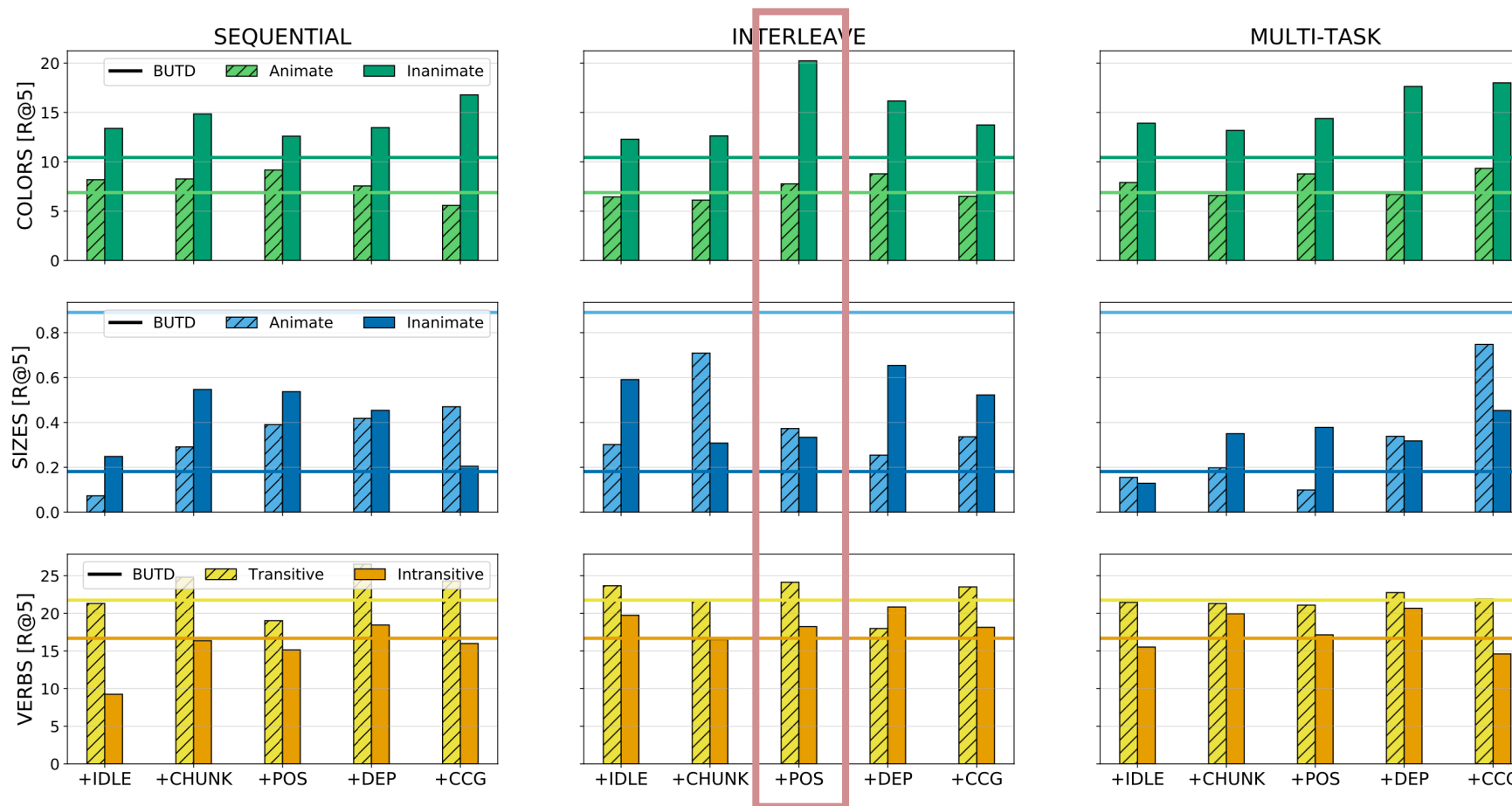
# Generalisation across categories



# Generalisation across categories

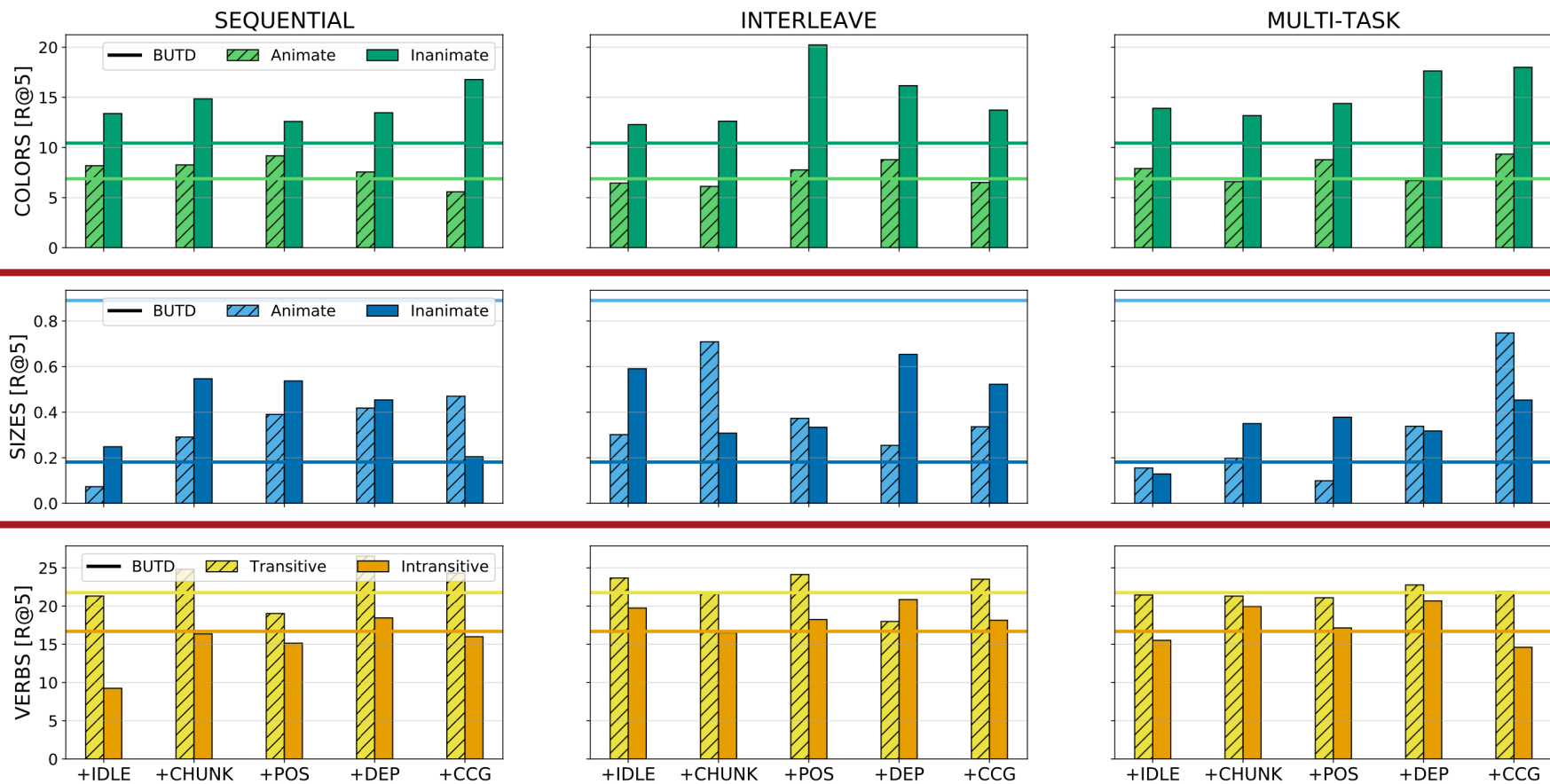


# Generalisation across categories

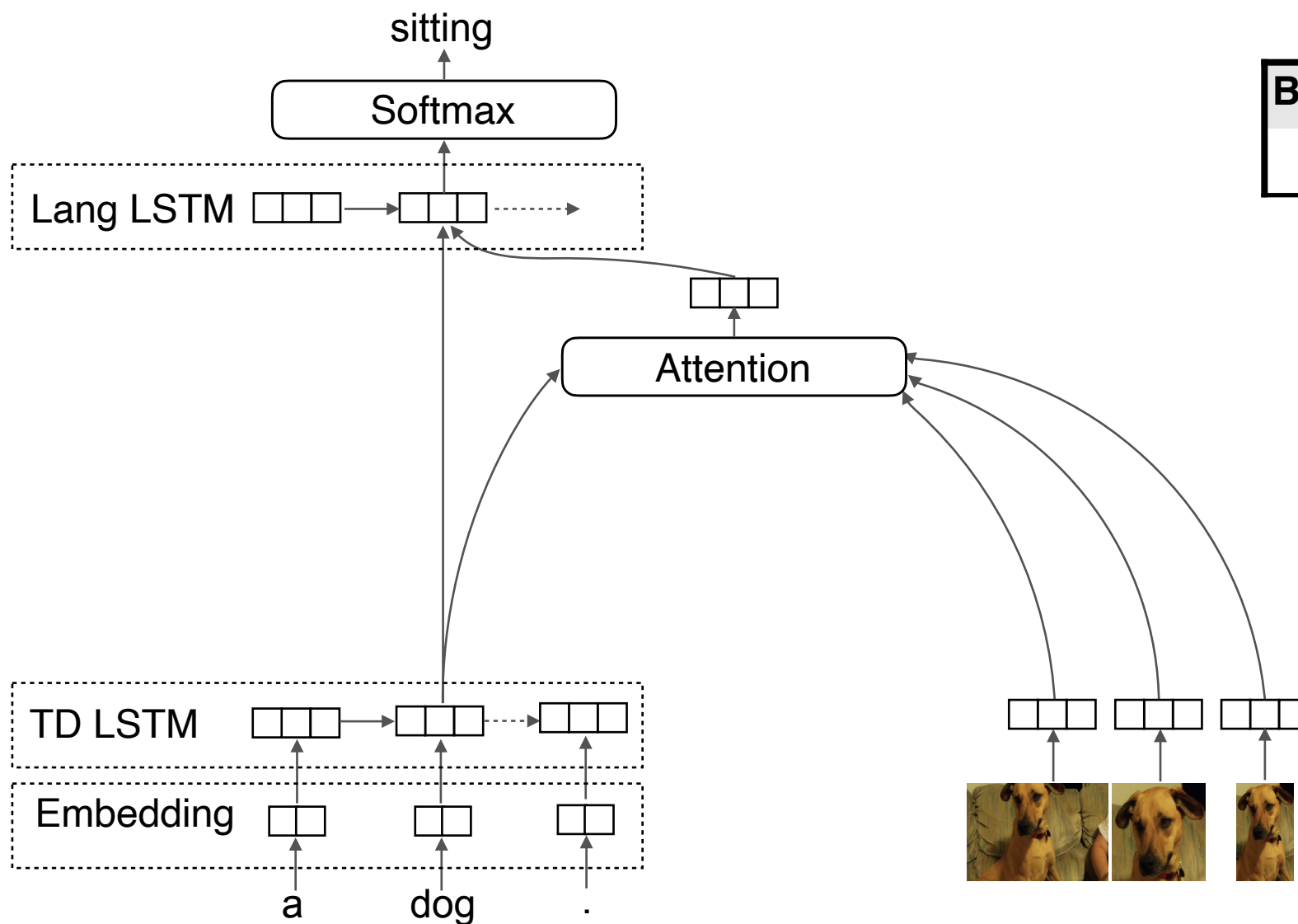




# Generalisation across categories

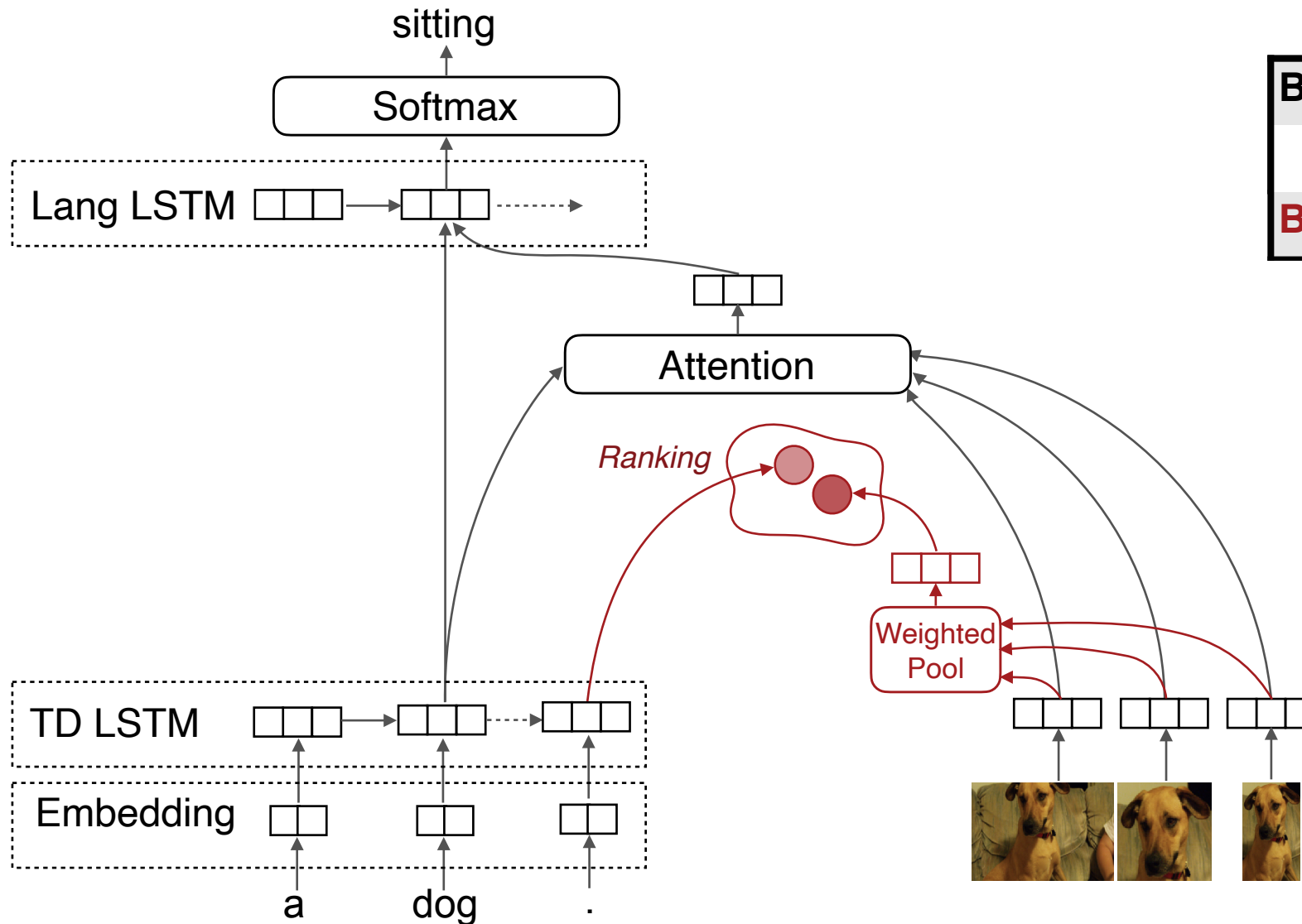


# Results: BUTD & BUTR



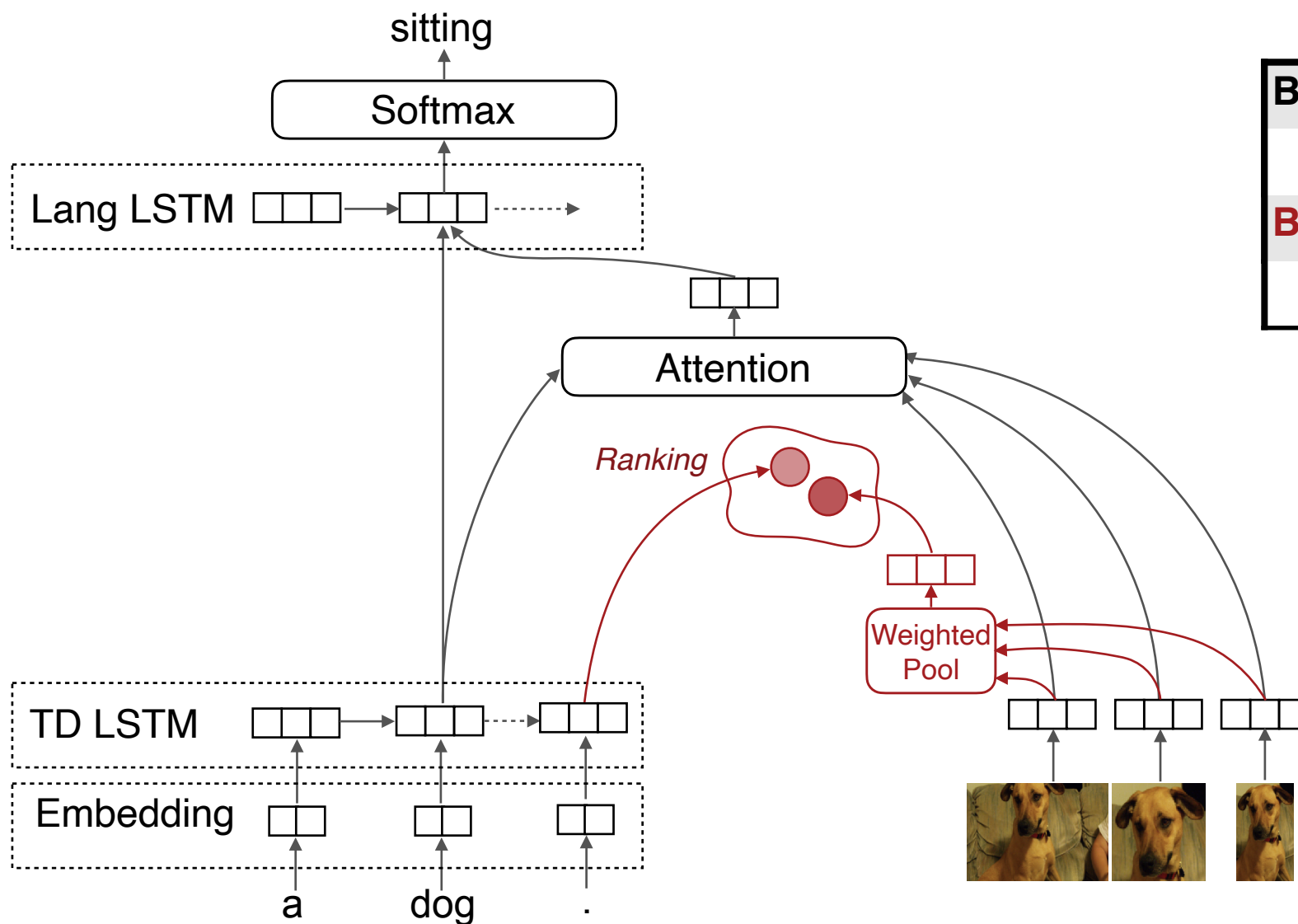
<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8

# Results: BUTD & BUTR



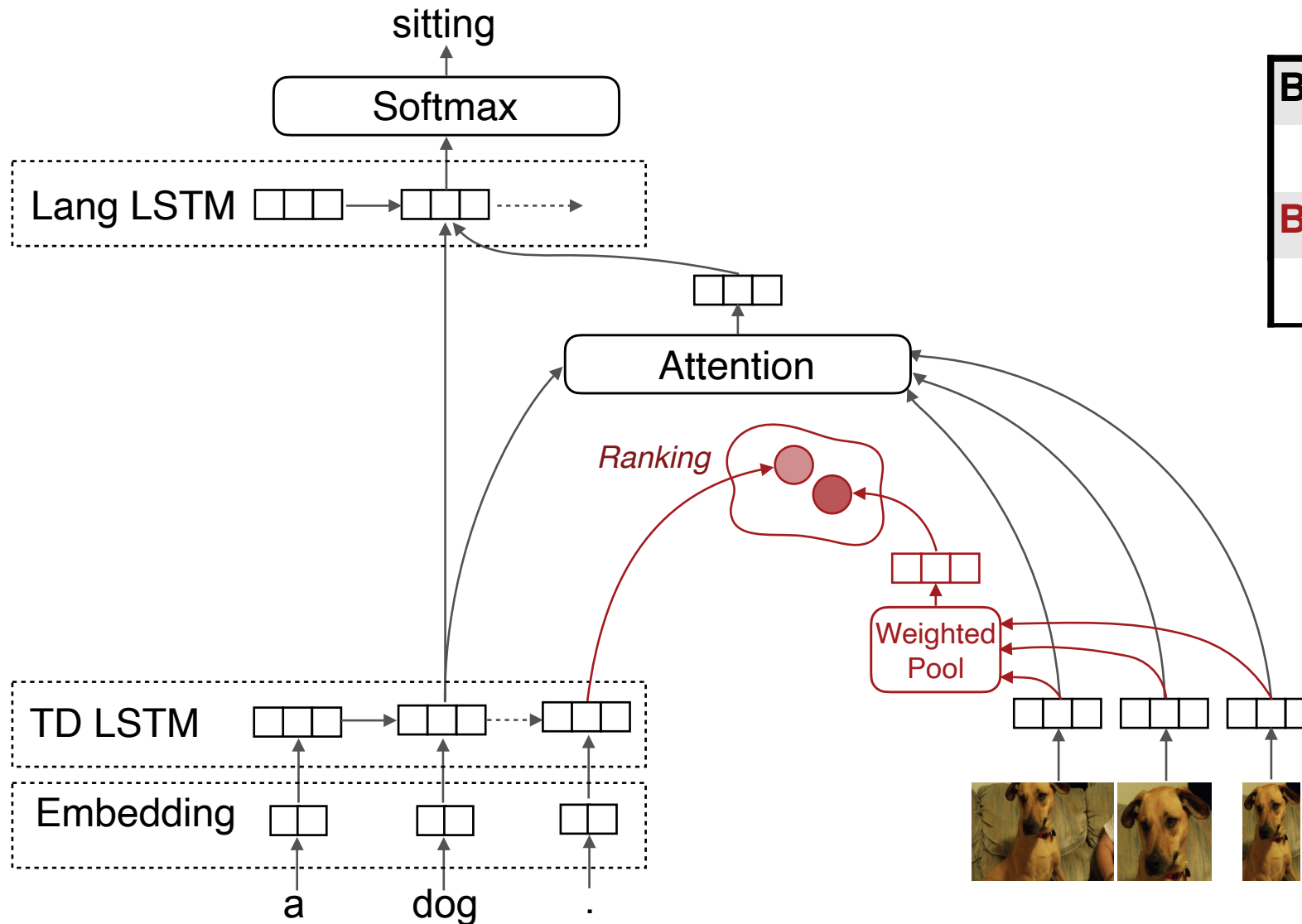
<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0

# Results: BUTD & BUTR



<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0

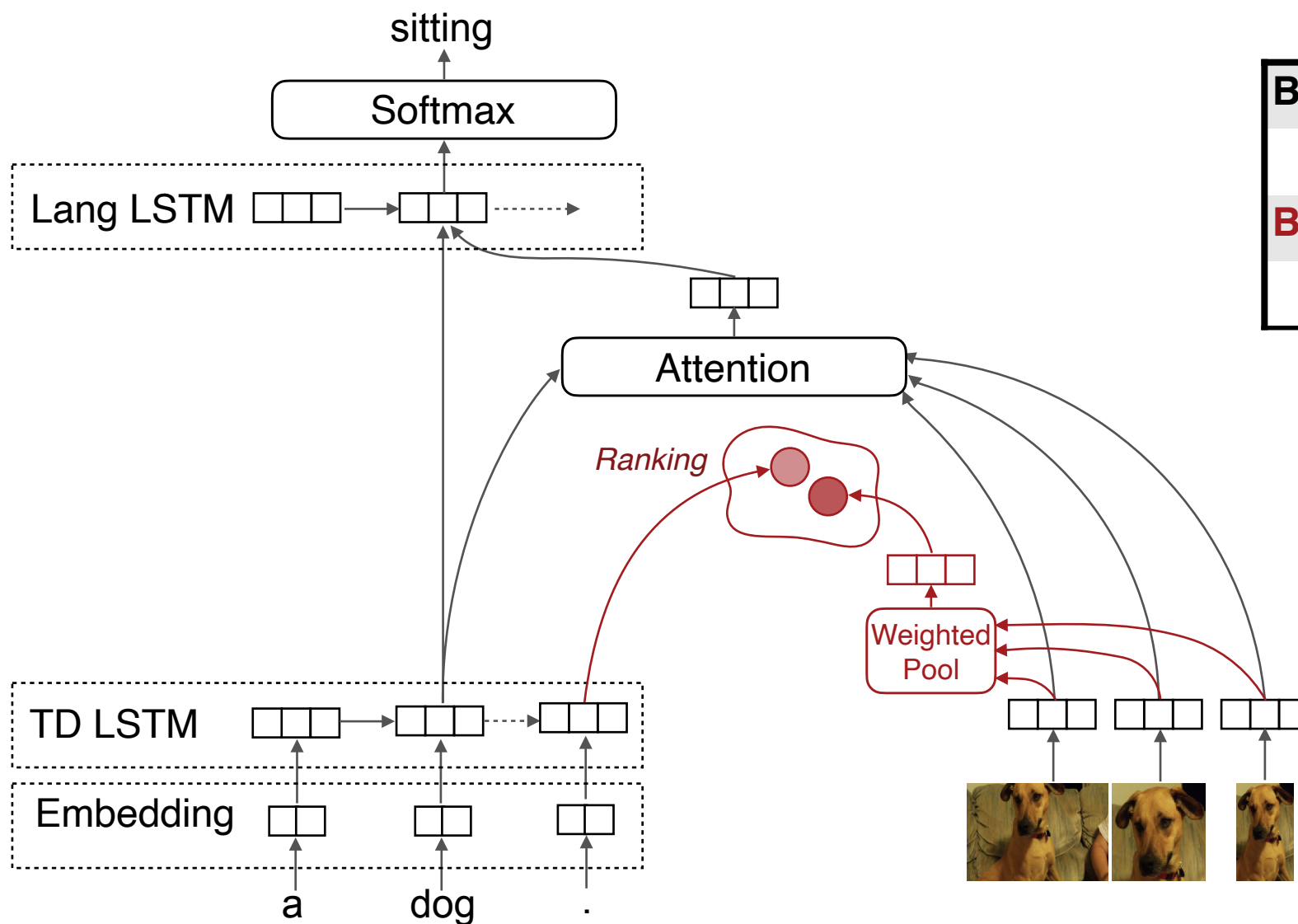
# Results: BUTD & BUTR



<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0

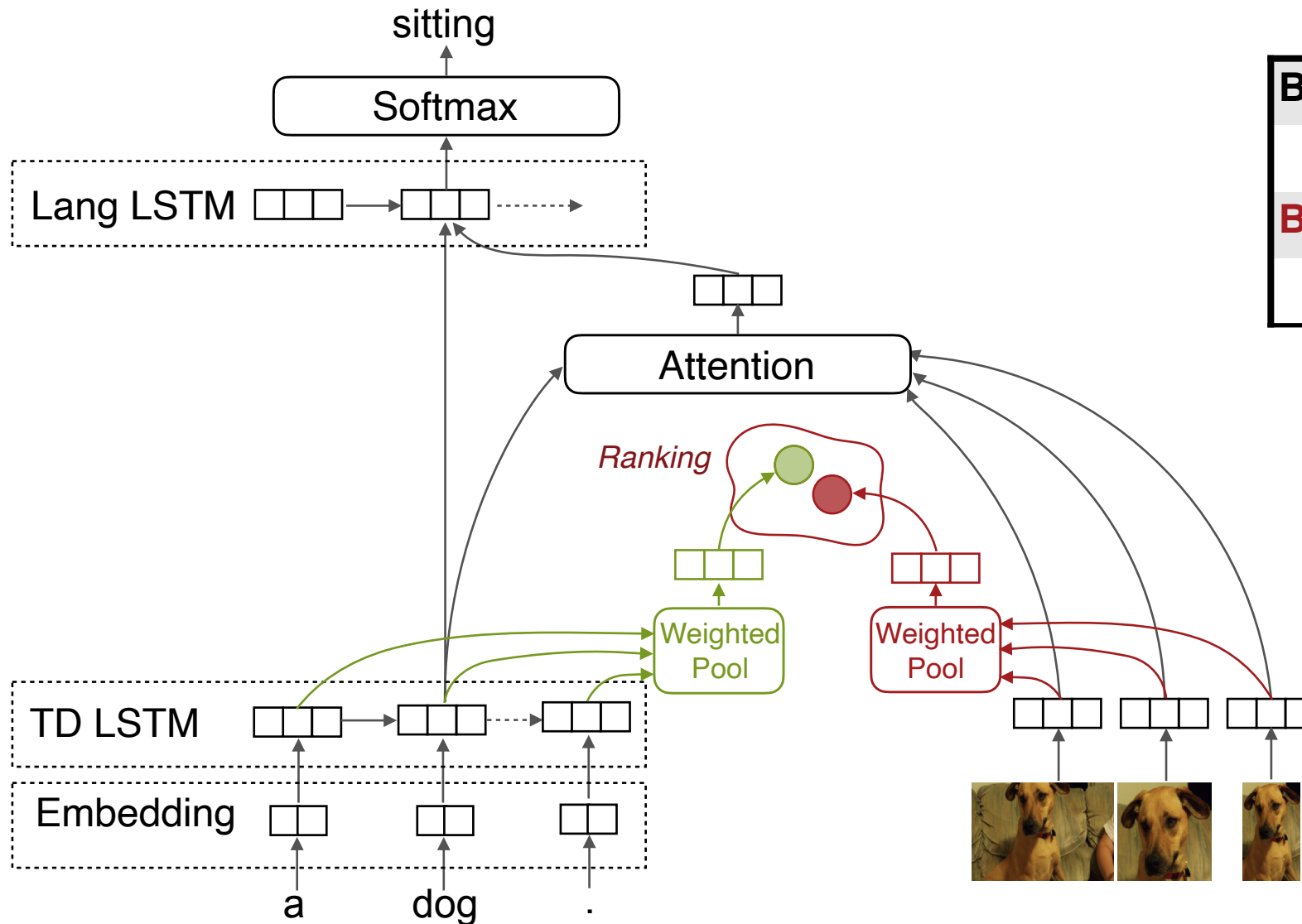
**Syntax hurts retrieval**

# Results: BUTRweight



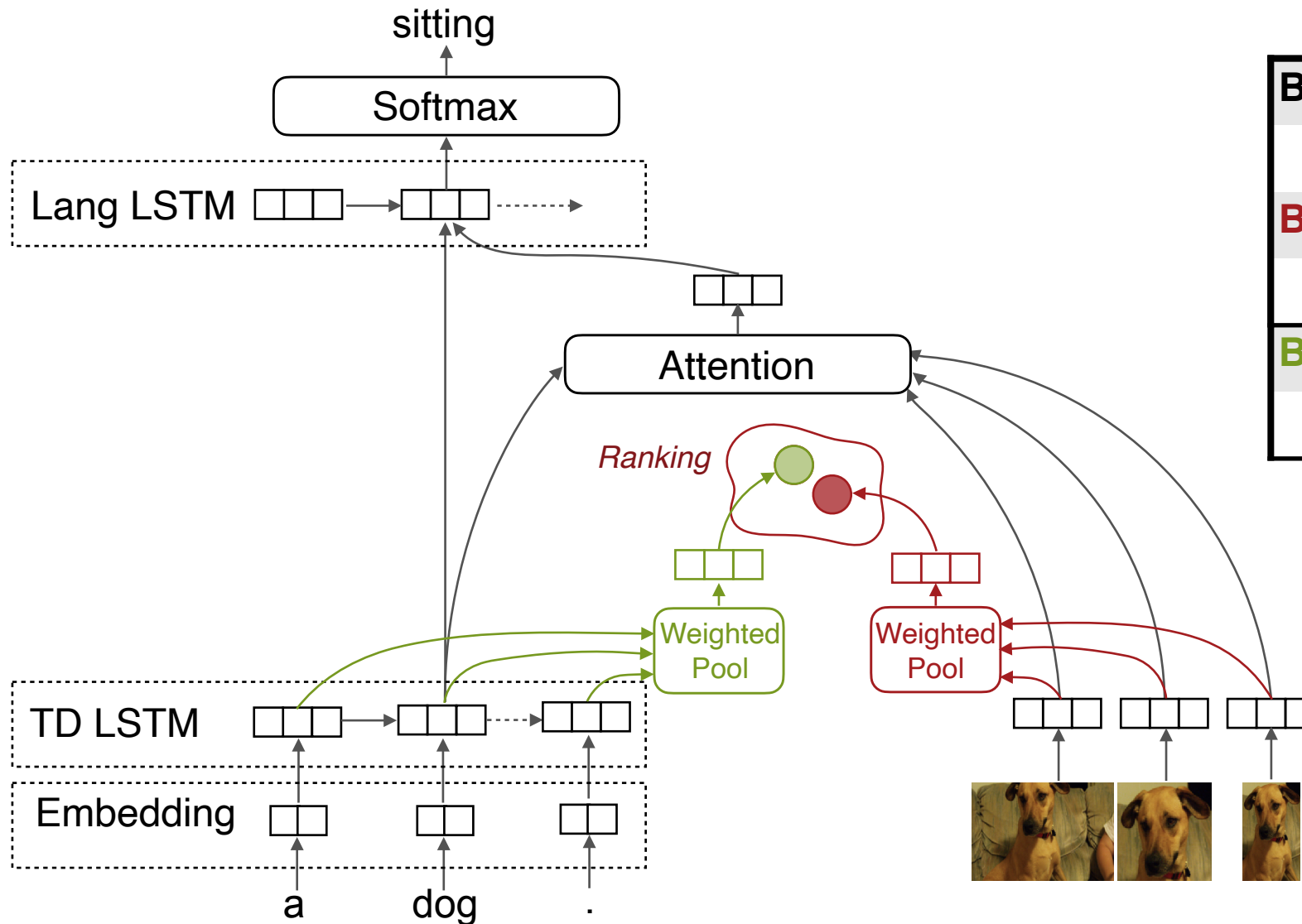
<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0

# Results: BUTRweight



<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0

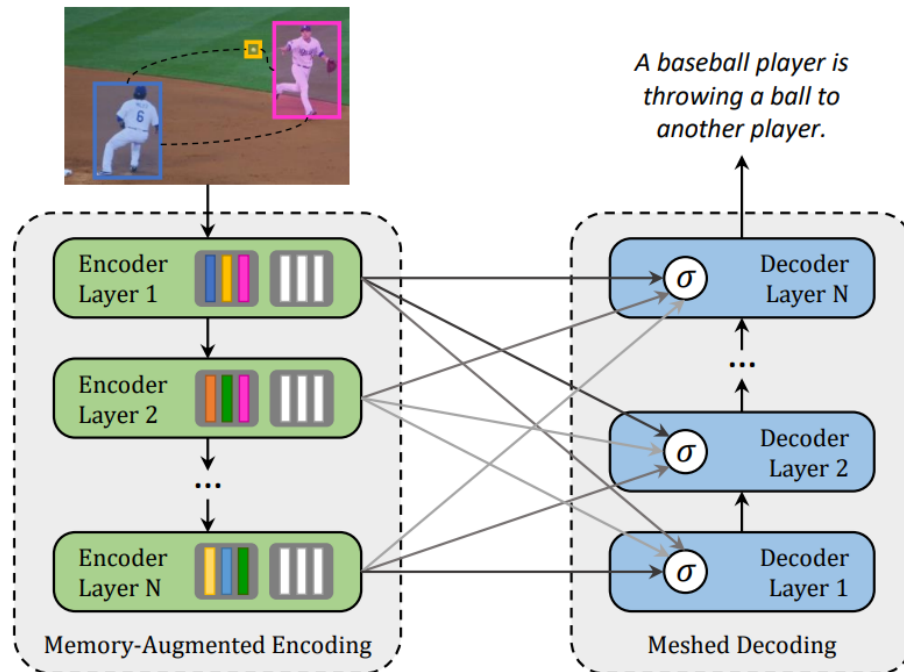
# Results: BUTRweight



<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0
<b>BUTRweight (ours)</b>	14.9
<b>+POS</b>	<b>16.4</b>

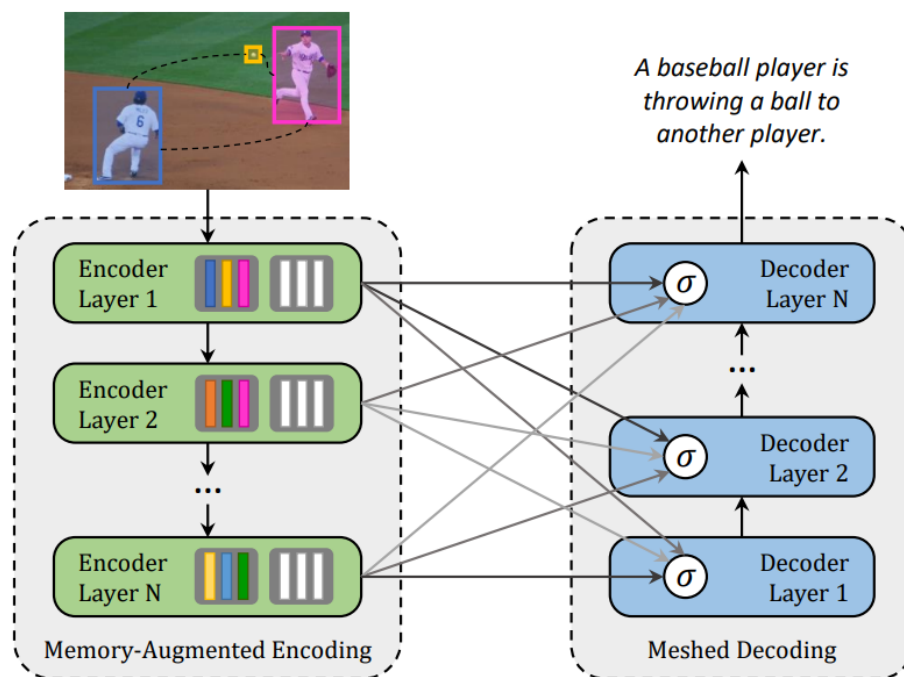


# Results: M2 Transformer (Cornia et al., 2020)



<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0
<b>BUTRweight (ours)</b>	14.9
<b>+POS</b>	<b>16.4</b>

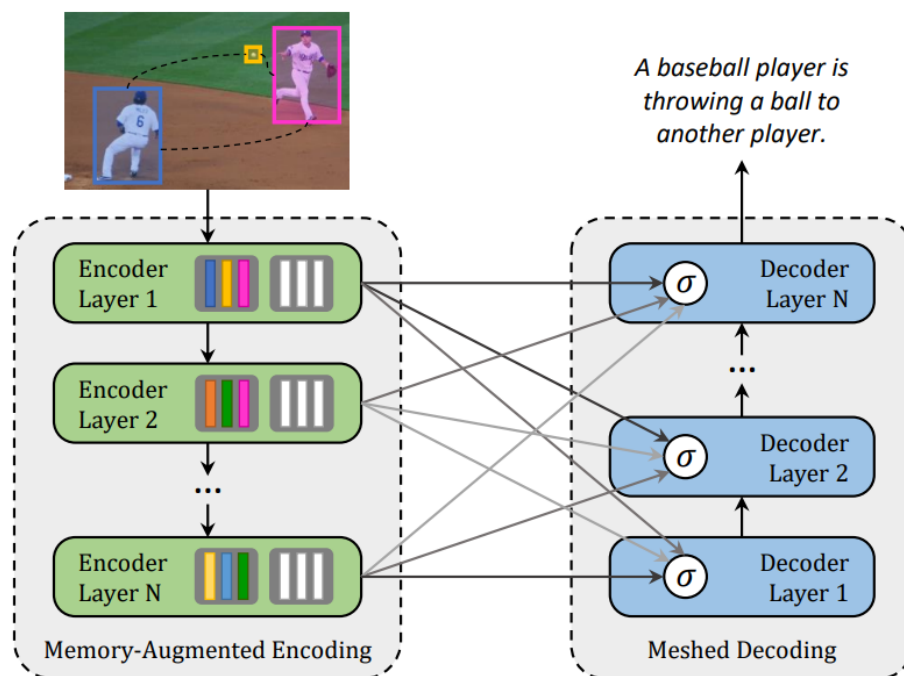
# Results: M2 Transformer (Cornia et al., 2020)



<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0
<b>BUTRweight (ours)</b>	14.9
<b>+POS</b>	<b>16.4</b>
<b>M2 TRM (Cornia+, 2020)</b>	10.6

**Transformers do not compositionally generalise**

# Results: M2 Transformer (Cornia et al., 2020)



<b>BUTD (Anderson+, 2018)</b>	9.5
<b>+POS</b>	11.8
<b>BUTR (Nikolaus+, 2019)</b>	15.0
<b>+POS</b>	12.0
<b>BUTRweight (ours)</b>	14.9
<b>+POS</b>	<b>16.4</b>
<b>M2 TRM (Cornia+, 2020)</b>	10.6
<b>+POS</b>	13.2

Transformers do not compositionally generalise

Interleaving POS tags with words is model-agnostic

# Qualitative Analysis: ride-woman

**BUTD**

**BUTD+POS**

# Qualitative Analysis: ride-woman



**BUTD**

there is a woman that is  
on the floor

**BUTD+POS**

a woman riding a bike on  
a wooden floor

# Qualitative Analysis: ride-woman



**BUTD**

there is a woman that is  
on the floor

**BUTD+POS**

a woman riding a bike on  
a wooden floor



a woman with a child  
sitting on a bench

a **girl that is standing**  
on a skateboard

# Qualitative Analysis: ride-woman

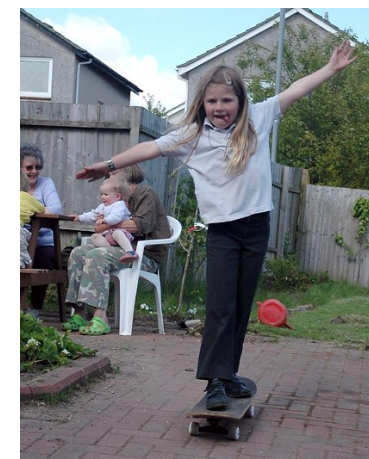


**BUTD**

there is a woman that is  
on the floor

**BUTD+POS**

a woman riding a bike on  
a wooden floor



a woman with a child  
sitting on a bench

a **girl that is standing**  
on a skateboard

**Higher quality** captions with more unseen concept pairs

# Conclusion



# Conclusion

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning

# Conclusion

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning
- Syntactic planning *consistently* improves compositional generalisation

# Conclusion

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning
- Syntactic planning *consistently* improves compositional generalisation
  - With no loss for standard metrics (see paper)

# Conclusion

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning
- Syntactic planning *consistently* improves compositional generalisation
  - With no loss for standard metrics (see paper)
  - In both RNN- and Transformer-based models

# Conclusion

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning
- Syntactic planning *consistently* improves compositional generalisation
  - With no loss for standard metrics (see paper)
  - In both RNN- and Transformer-based models
- Future work

# Conclusion

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning
- Syntactic planning *consistently* improves compositional generalisation
  - With no loss for standard metrics (see paper)
  - In both RNN- and Transformer-based models
- Future work
  - Syntax-aware captioning models

# Conclusion

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning
- Syntactic planning *consistently* improves compositional generalisation
  - With no loss for standard metrics (see paper)
  - In both RNN- and Transformer-based models
- Future work
  - Syntax-aware captioning models
  - Fine-grained syntactic tags (e.g. CCG)

# Conclusion



Thanks!

- Revisited traditional NLG pipeline
  - Syntactic planning stage for image captioning
- Syntactic planning *consistently* improves compositional generalisation
  - With no loss for standard metrics (see paper)
  - In both RNN- and Transformer-based models
- Future work
  - Syntax-aware captioning models
  - Fine-grained syntactic tags (e.g. CCG)

Code, models & data available online at <https://github.com/e-bug/syncap>