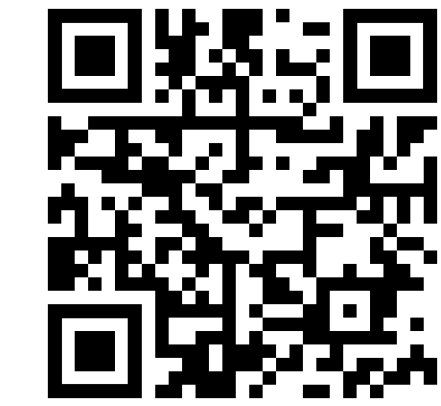


# The Role of Syntactic Planning in Compositional Image Captioning

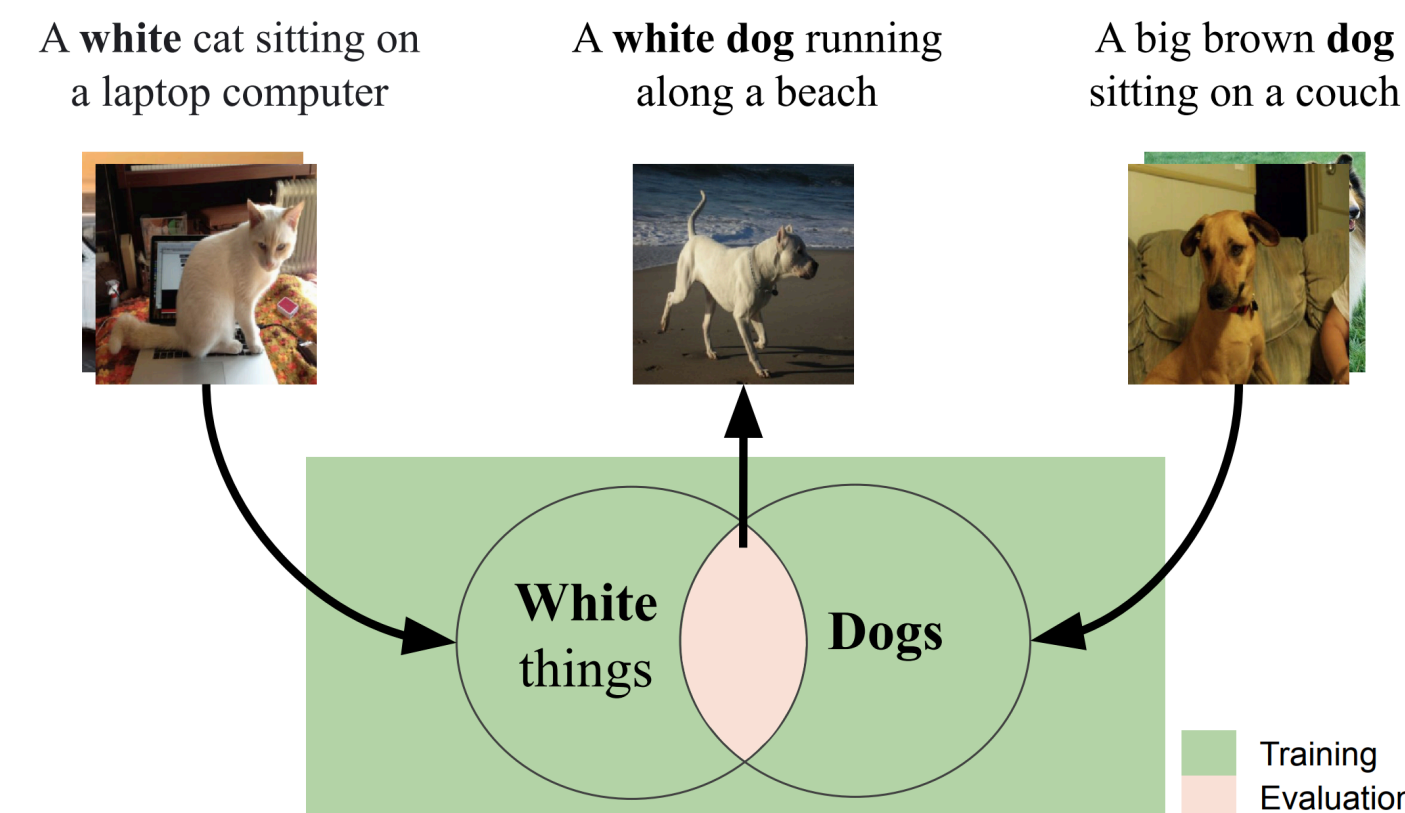
Emanuele Bugliarello and Desmond Elliott

EACL 2021

UNIVERSITY OF  
COPENHAGEN

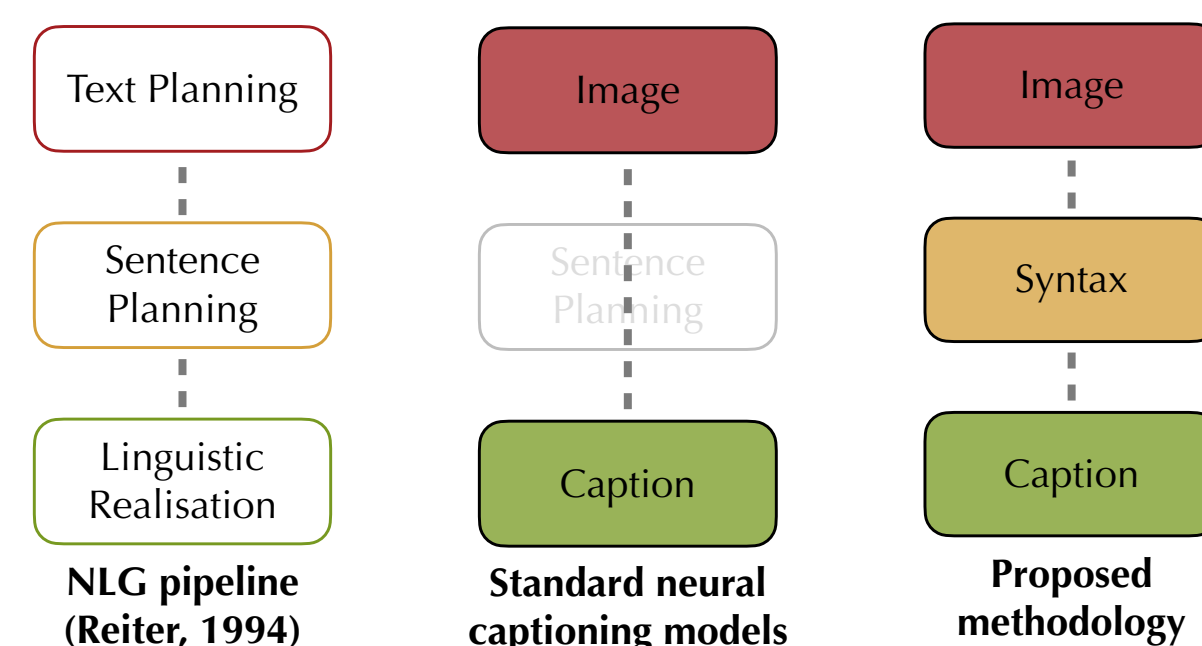


## Compositional Image Captioning



- **Goal:** caption images that have *unseen compositions* of *known concepts*
- **Nikolaus et al. (2019)** showed that RNN-based captioning models **do not compositionally generalise** due to the text decoder

## Syntactic Planning



- **Idea:** Introduce a sentence planning step where the model plans the syntactic structure of the caption
- **Hypothesis:** a model assigns higher probability to the unseen phrase “white dog” if it explicitly expects to generate JJ NN

## Syntactic Granularity & Planning Approaches

- **Syntactic tags:** From chunking to part-of-speech to dependency labels to CCG tags (plus a synthetic IDLE tag for each word)

CHUNK    POS    DEP    CCG    IDLE

- **Planning approaches:**

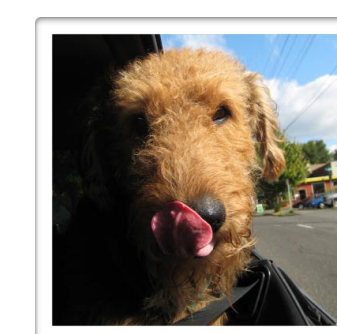
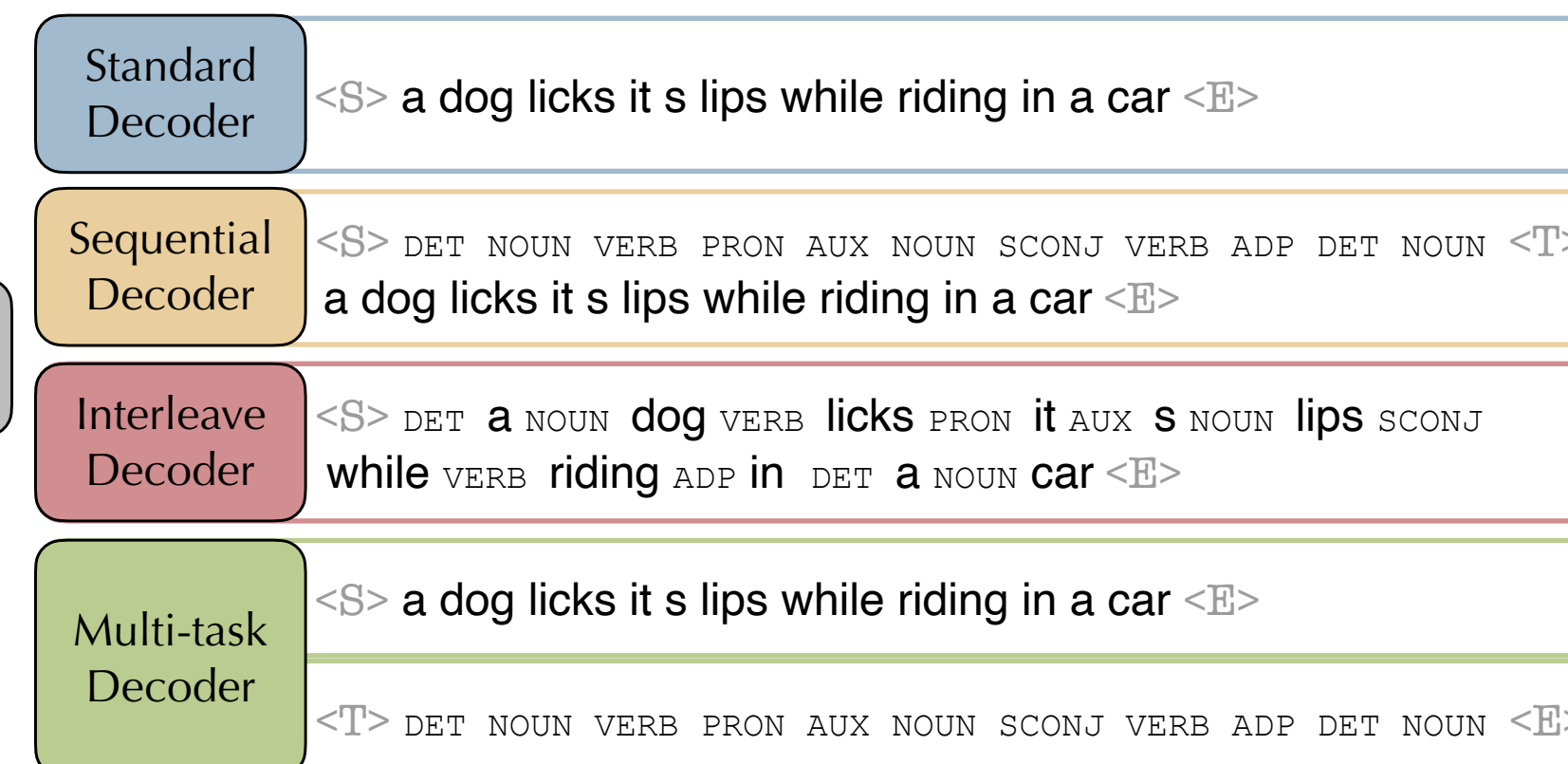
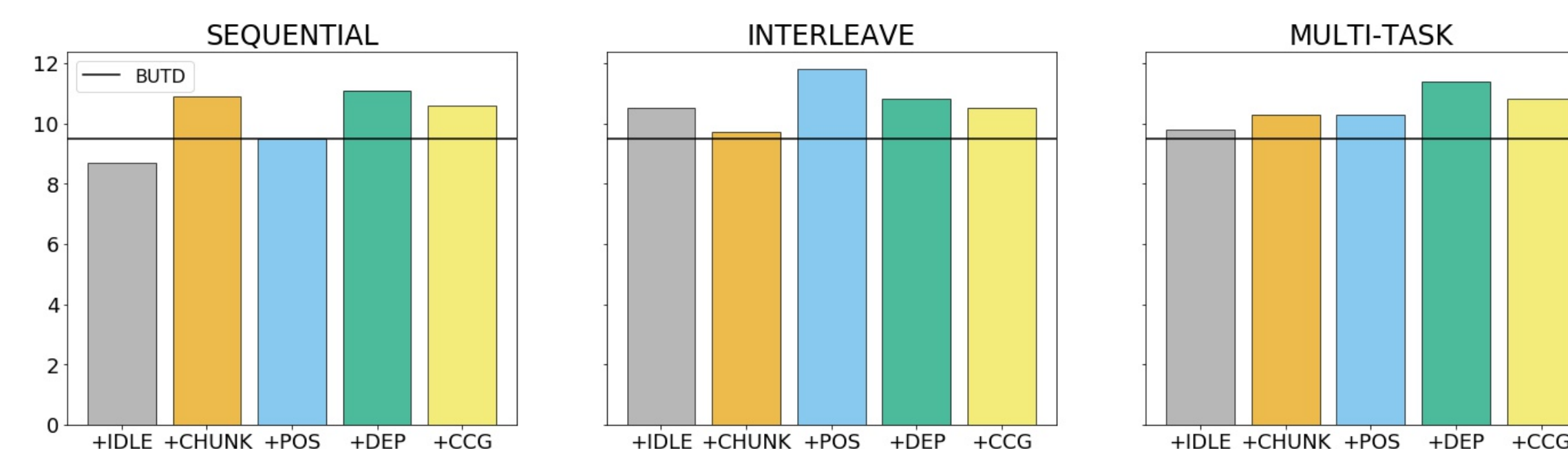


Image Encoder



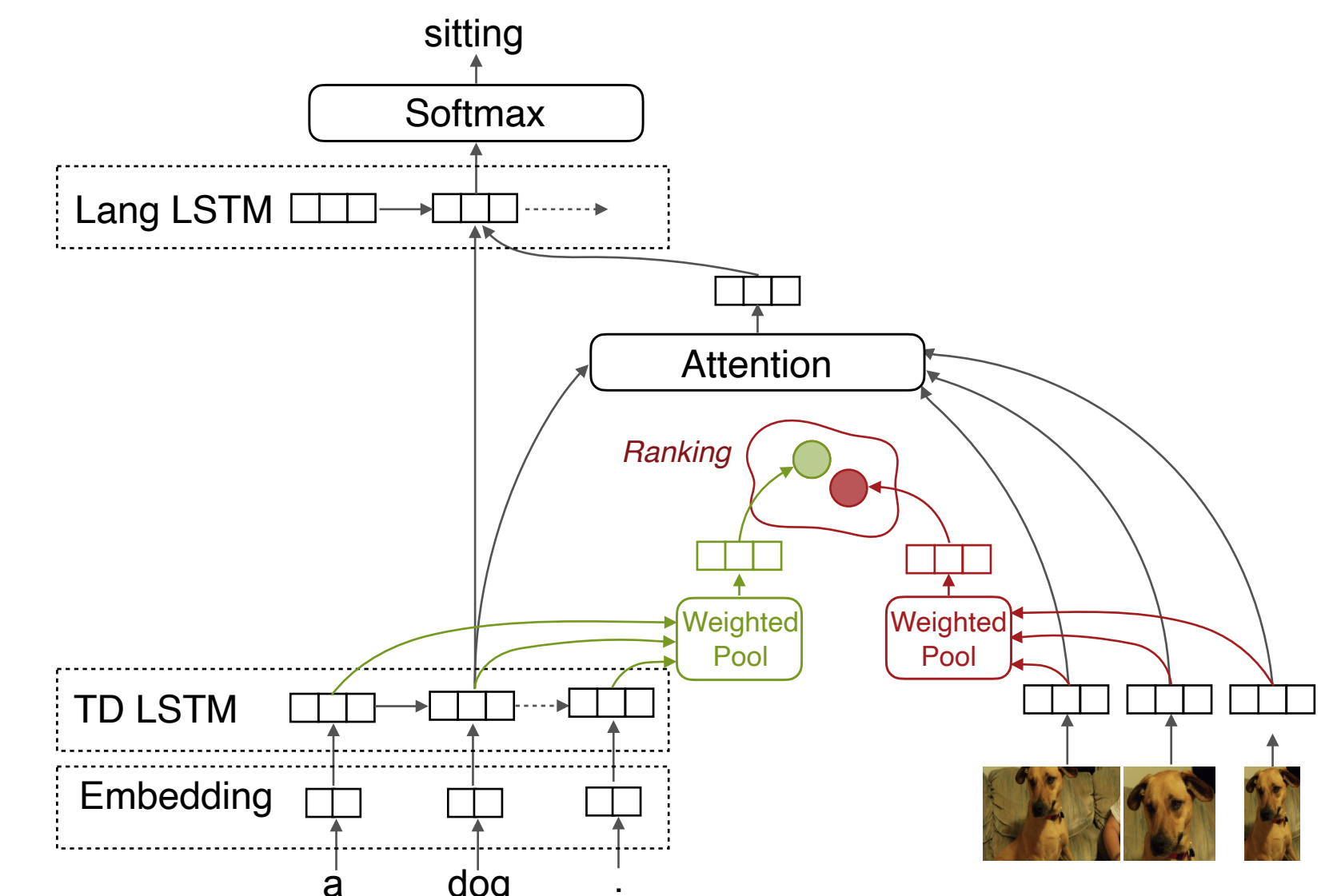
## Syntax Awareness



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

- Directly mapping an image onto words is sub-optimal
- Breaking bi-gram sequences with **IDLE** is useful
- **Syntactic planning improves compositional image captioning**

## BUTR<sub>weight</sub>: Adaptive Re-ranking

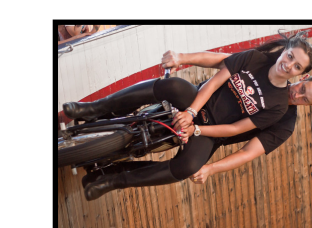


## Results & Examples

	BUTD	BUTR	BUTR <sub>weight</sub>	M2
Standard	9.5	15.0	14.9	10.6
Interleave POS	11.8	12.0	<b>16.4</b>	13.2

Average Recall@5 of unseen concepts

- Interleaving POS tags with words is model-agnostic
- Transformers do not compositionally generalise



**BUTD** there is a woman that is on the floor  
**BUTD+POS** a woman riding a bike on a wooden floor



**BUTD** a woman with a child sitting on a bench  
**BUTD+POS** a girl that is standing on a skateboard