CS6501: Deep Learning for Visual Recognition

Seq2Seq Model & Text-to-Image Synthesis

Presenter: Fuwen Tan

Today's Class

- Mini-batch training of the RNN model
  - Special "End-of-Sequence" token: <end>
  - Padding
- Sequence-to-sequence model
  - Neural Machine Translation[1]
- Text-to-Image Synthesis[2]

[1] Effective Approaches to Attention-based Neural Machine Translation. Thang Luong, Hieu Pham, and Christopher D. Manning. EMNLP 2015
[2] Text2Scene: Generating Compositional Scenes from Textual Descriptions. Fuwen Tan, Song Feng, Vicente Ordonez. CVPR 2019.

A RNN model will never end

“Hello”, “world”, “!”, “!”, “!”, “!”, “!”, …

Unless: set the maximum length before hand

Sample 1	*“hello”*	*“world”*	*“java”*	*“is”*	*“better”*
Sample 2	*“hello”*	*“hoos”*	*“I”*	*“like”*	*“python”*
Sample 3	*“one”*	*“plus”*	*“eight”*	*“equals”*	*“to”*

I want sentences of 5 words

Or: learn to predict the END.

Ground-truth: *“Hello”, “world”, “<end>”*

Training: learn to generate the ground-truth sequence with *“<end>”*.
Testing: generate the sequence until an *“<end>”* is predicted.

Computing loss: what if #ground-truth != #prediction

Ground-truth	*“hello”*	*“world”*	*“<end>”*		
Prediction 1	*“hello”*	*“<end>”*	*“foo”*		
Prediction 2	*“hello”*	*“how”*	*“are”*	*“you”*	*“<end>”*

$loss_1$ $loss_2$ $loss_3$

Mini-batch training: padding

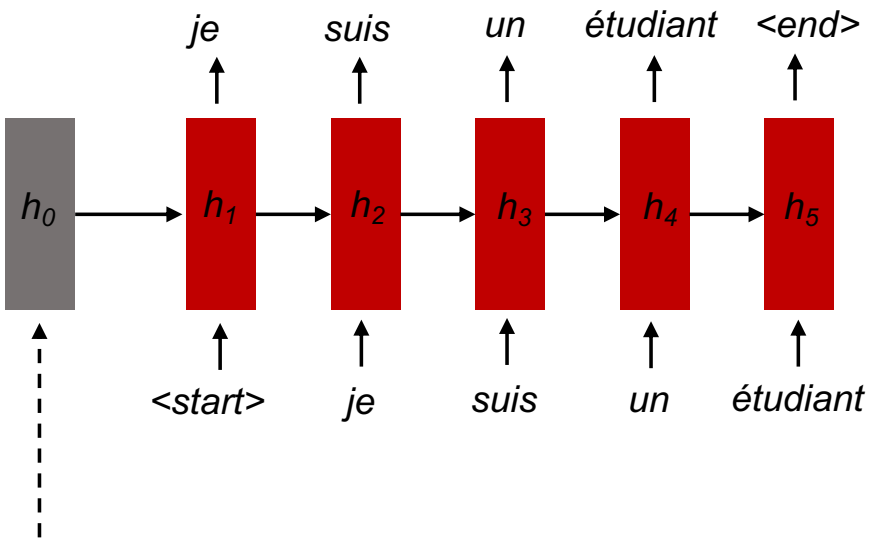
Sample 1	*“hello”*	*“how”*	*“are”*	*“you”*	*“today”*	*“<end>”*
Sample 2	*“a”*	*“dog”*	*“is”*	*“driving”*	*“<end>”*	***“<pad>”***
Sample 3	*“hello”*	*“world”*	*“<end>”*	***“<pad>”***	***“<pad>”***	***“<pad>”***

Mini-batch training: padding

Sample 1	*“hello”*	*“how”*	*“are”*	*“you”*	*“today”*	*“<end>”*
Sample 2	*“a”*	*“dog”*	*“is”*	*“driving”*	*“<end>”*	***“<pad>”***
Sample 3	*“hello”*	*“world”*	*“<end>”*	***“<pad>”***	***“<pad>”***	***“<pad>”***

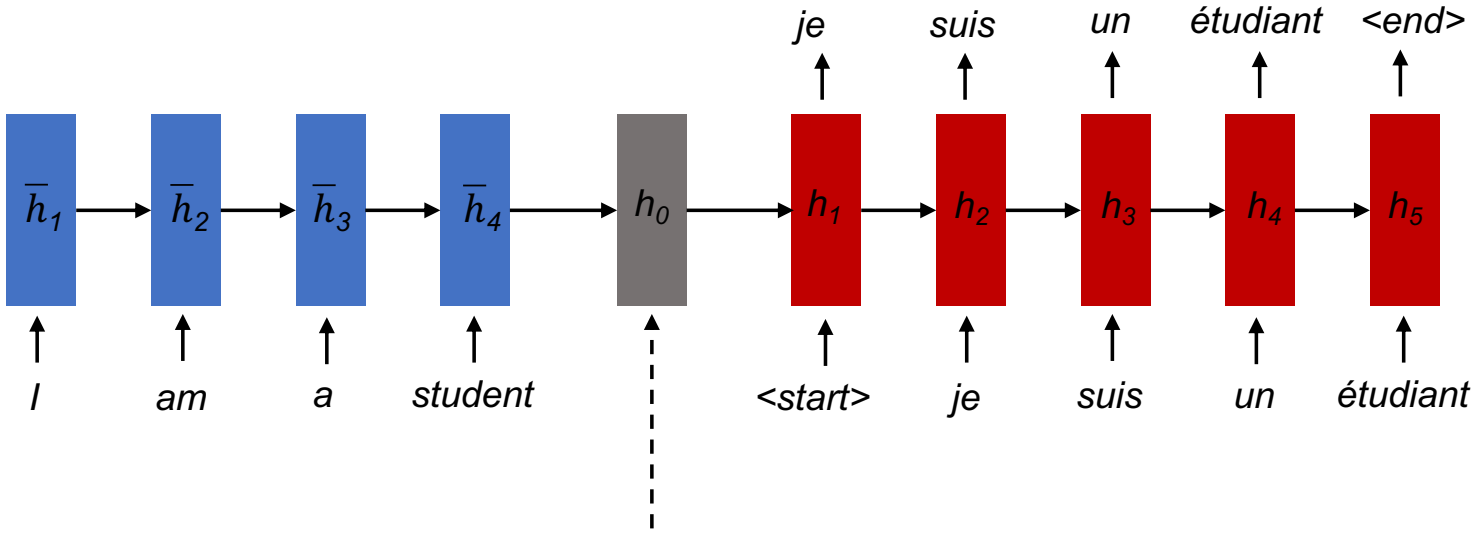
Sample 1	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*
Sample 2	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*	***0.0***
Sample 3	*1.0*	*1.0*	*1.0*	***0.0***	***0.0***	***0.0***

Generating text that makes sense: Language Model



Unconditional: $h_0 = 0$

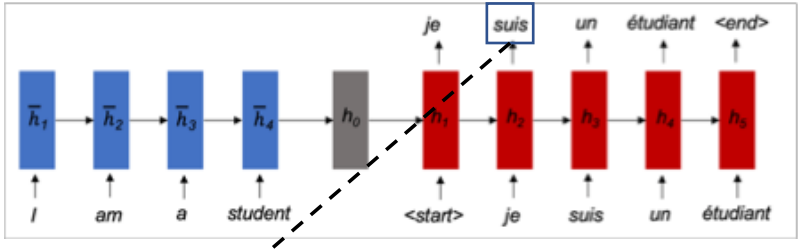
Generating text with a goal: Machine Translation



Conditional: $h_0 = \overline{h}_4$

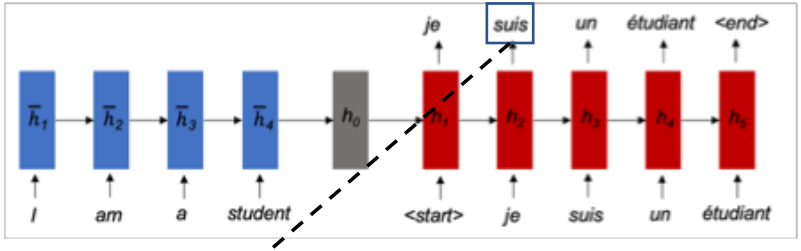
[3] Sequence to Sequence Learning with Neural Networks. Ilya Sutskever, Oriol Vinyals, Quoc V. Le. NeurIPS 2014.

Seq2Seq model



Seq2Seq: “suis” ← $f(h_2) = softmax(W_s h_2)$

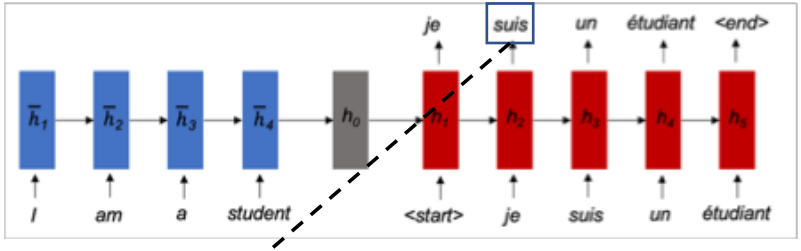
Seq2Seq model with perfect word alignments



Seq2Seq: "suis" ← $f(h_2)$

Ideally: "suis" ← $f'(h_2, \text{“am”})$

Seq2Seq model with perfect word alignments



Seq2Seq: "suis" ← $f(h_2)$

Ideally: "suis" ← $f'(h_2, \text{“am”})$

Or: "suis" ← $f'(h_2, \bar{h}_2)$

Seq2Seq model with attention

Ideally: "suis" $\longleftarrow$ $f'(h_2, \bar{h}_2)$

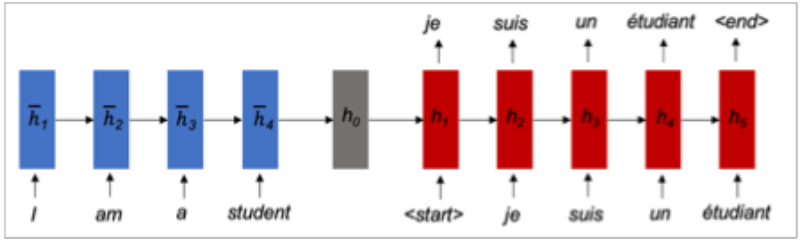
In practice: "suis" $\longleftarrow$ $f'(h_2,\ c_2)$

$$c_2 = \sum_{k=1}^{4} w_{2,k} \bar{h}_k$$

Pray that S: $w_{2,2} = 1, w_{2,k!=2} = 0$ is true

Or train the model such that S is almost true

Seq2Seq model with attention



Key assumption: $\bar{h}_2 \approx h_2 \approx h_0 - \text{“}je\text{”} \approx \bar{h}_4 - \text{“}je\text{”}$

$$w_{2,k} = \frac{exp\ (score(h_2, \bar{h}_k))}{\sum_j exp\ (score(h_2, \bar{h}_j))}$$

$$score\left(h_2, \bar{h}_k\right) = h_2^T W_a \bar{h}_k$$

[1] Effective Approaches to Attention-based Neural Machine Translation. Thang Luong, Hieu Pham, and Christopher D. Manning. EMNLP 2015

Seq2Seq model with attention

"suis" ← $f'(h_2,\ c_2) = softmax(W_s \tanh(W_c[h_2; c_2]))$

[1] Effective Approaches to Attention-based Neural Machine Translation. Thang Luong, Hieu Pham, and Christopher D. Manning. EMNLP 2015

Perform much better for long sequences

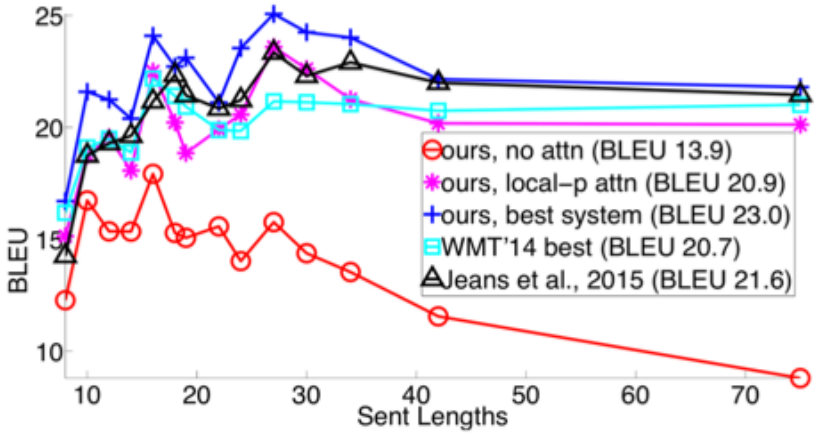
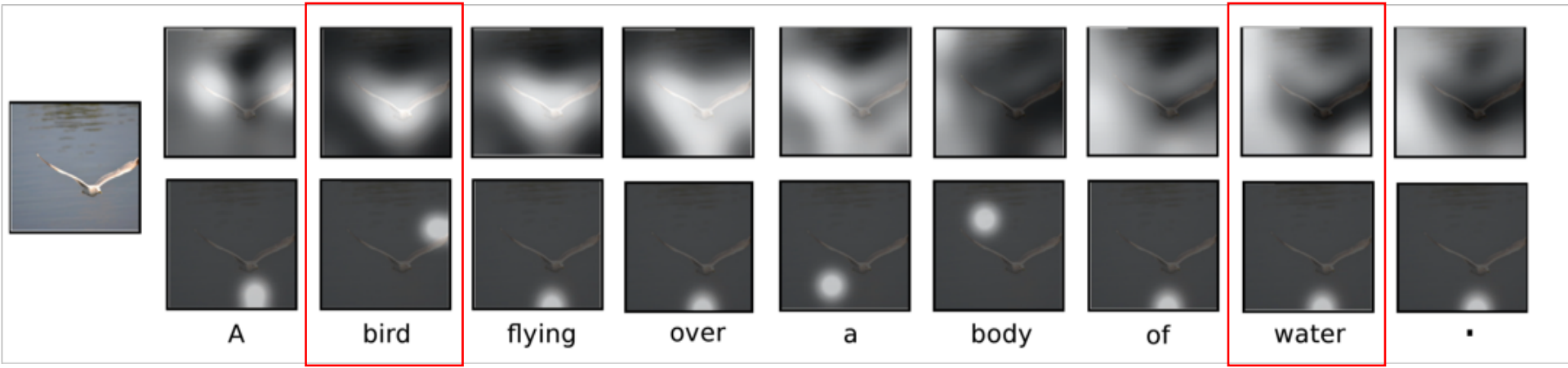


Figure 6: **Length Analysis** – translation qualities of different systems as sentences become longer.

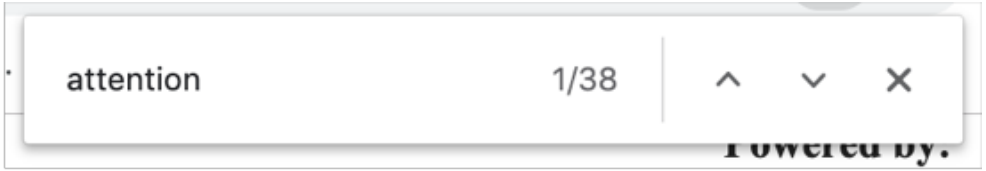
[1] Effective Approaches to Attention-based Neural Machine Translation. Thang Luong, Hieu Pham, and Christopher D. Manning. EMNLP 2015

Also very helpful in image captioning



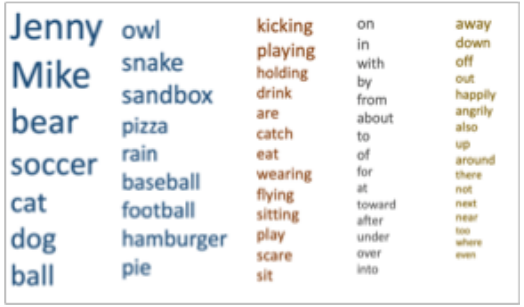
[4] Show, attend and tell: neural image caption generation with visual attention. Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. ICML 2015

ECCV 2018 accepted 776 papers



38 of them with “attention” in their titles

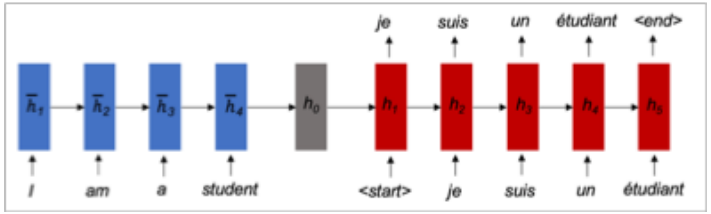
Seq2Seq vs Text-to-Image Synthesis

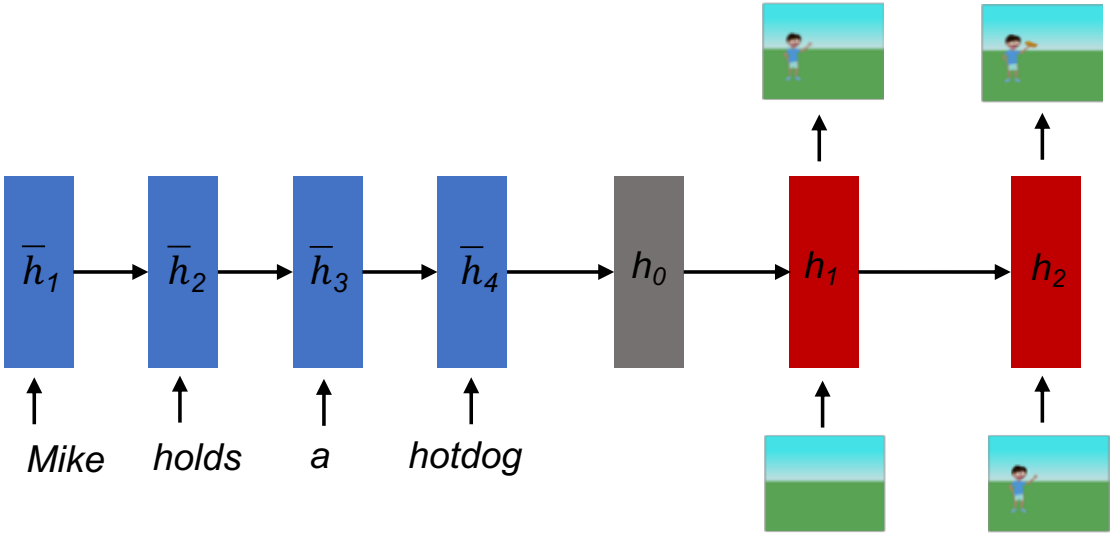


Sentence: composition of words

Image: composition of patches

Can we do this?





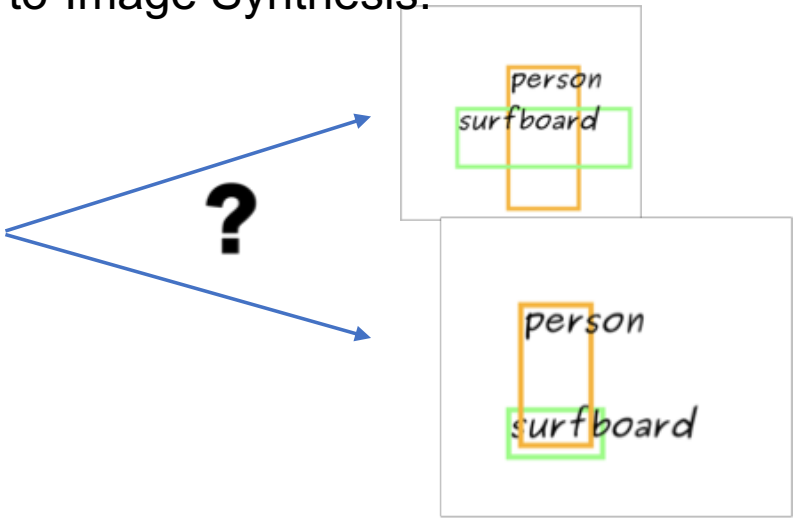
Challenges

Machine Translation:

"I am a student" → *"je suis un étudiant"*

Text-to-Image Synthesis:

"A person is ***holding*** *a surfboard"*



Challenges: in each step

Machine Translation:

student → étudiant

Text-to-Image Synthesis:

*"A person is **holding** a surfboard"*

object category: person, surfboard

location: somewhere in the 2D world

attributes: size, pose, expression, …

Challenges: in each step

Text-to-Image Synthesis:

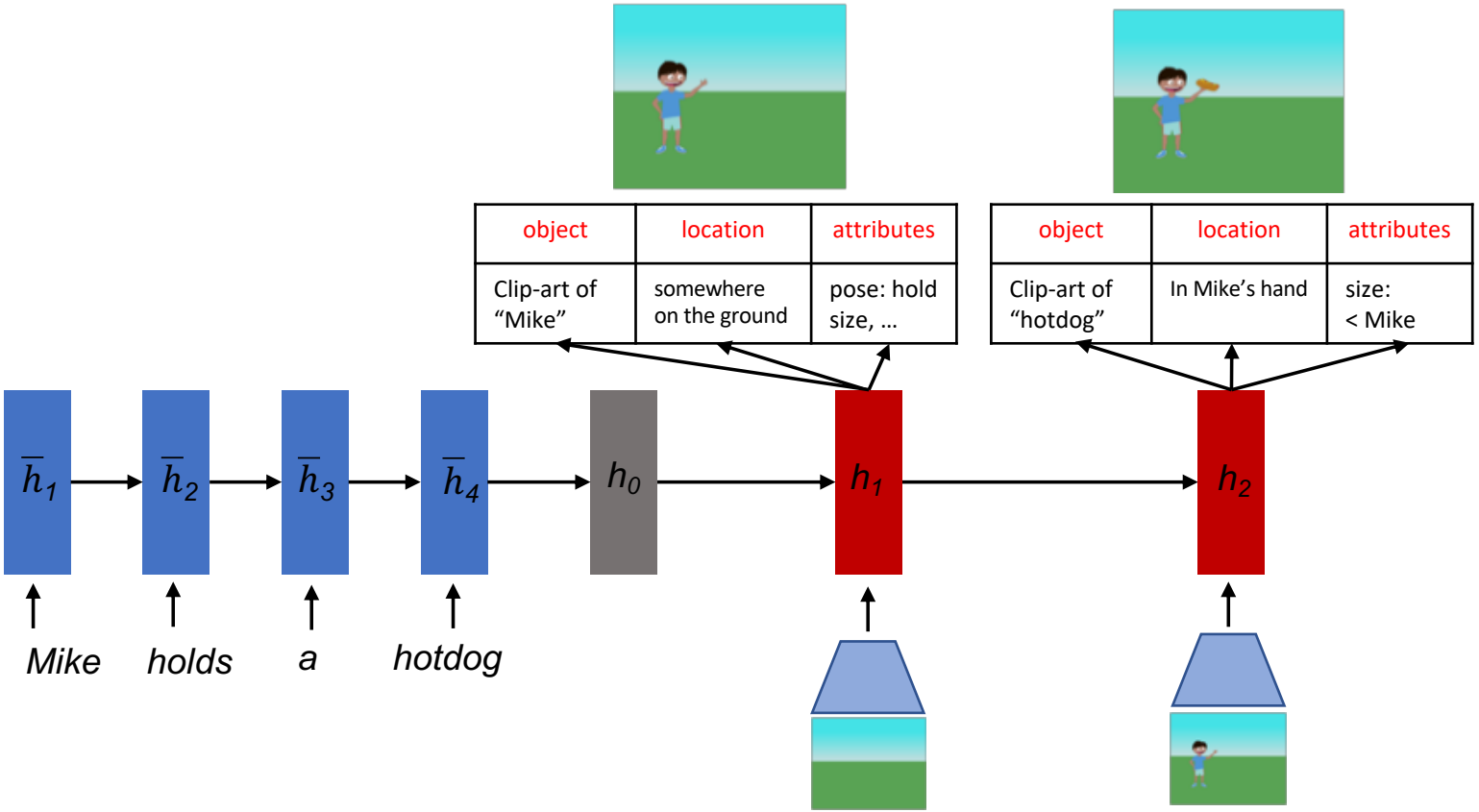
*"A person is **holding** a surfboard"*

object category: person

location: somewhere in the 2D world

attributes: size, pose, expression, …

Learning the distributions of
categories, locations, attributes
from the training samples


object
location
attributes
Clip-art of "Mike"
somewhere on the ground
pose: hold size, …
object
location
attributes
Clip-art of "hotdog"
In Mike's hand
size: < Mike
$\bar{h}_1$
$\bar{h}_2$
$\bar{h}_3$
$\bar{h}_4$
h_0
h_1
h_2
Mike
holds
a
hotdog

Task 1: Abstract Scene Generation

"Mike is surprised at the duck. The duck is standing on the grill. Jenny is running towards Mike and the duck."

Task 1: Abstract Scene Generation

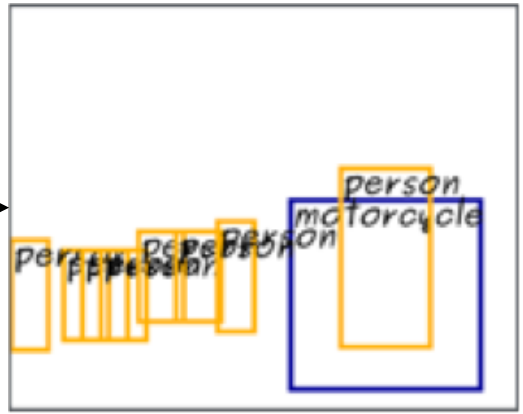
Object category → 58 clip-art objects

Location → 28 x 28 grid

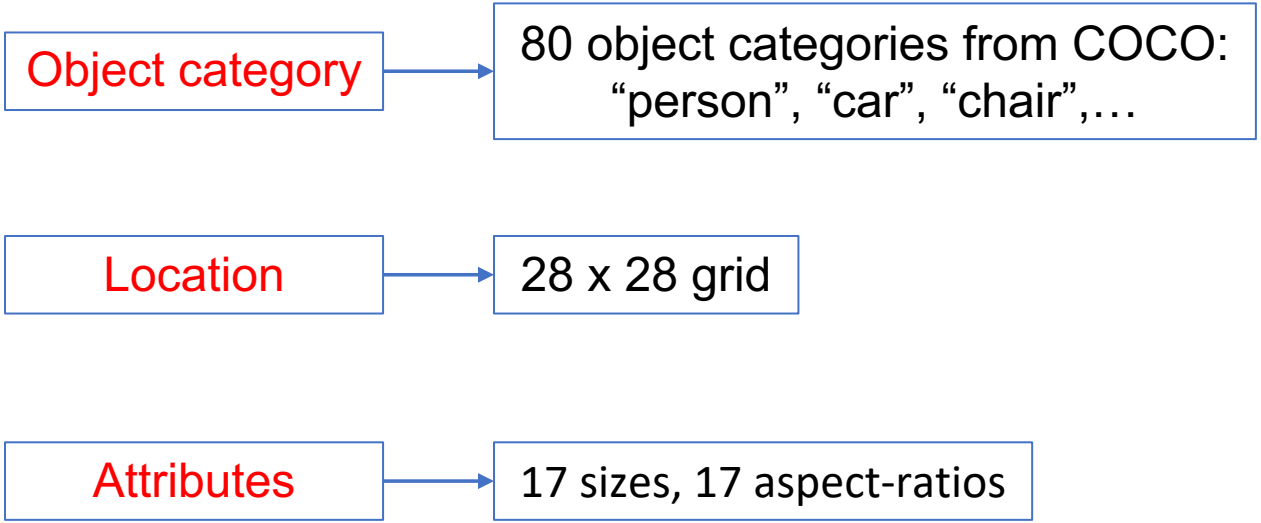
Attributes → 3 sizes, 2 orientations, 7 poses and 5 expressions for “Mike” and “Jenny”.

Task 2: Scene Layout Generation

"A guy on a motorcycle with some people watching."



Task 2: Scene Layout Generation



Task 3: Composite Image Generation

"Several elephants walking together in a line near water."

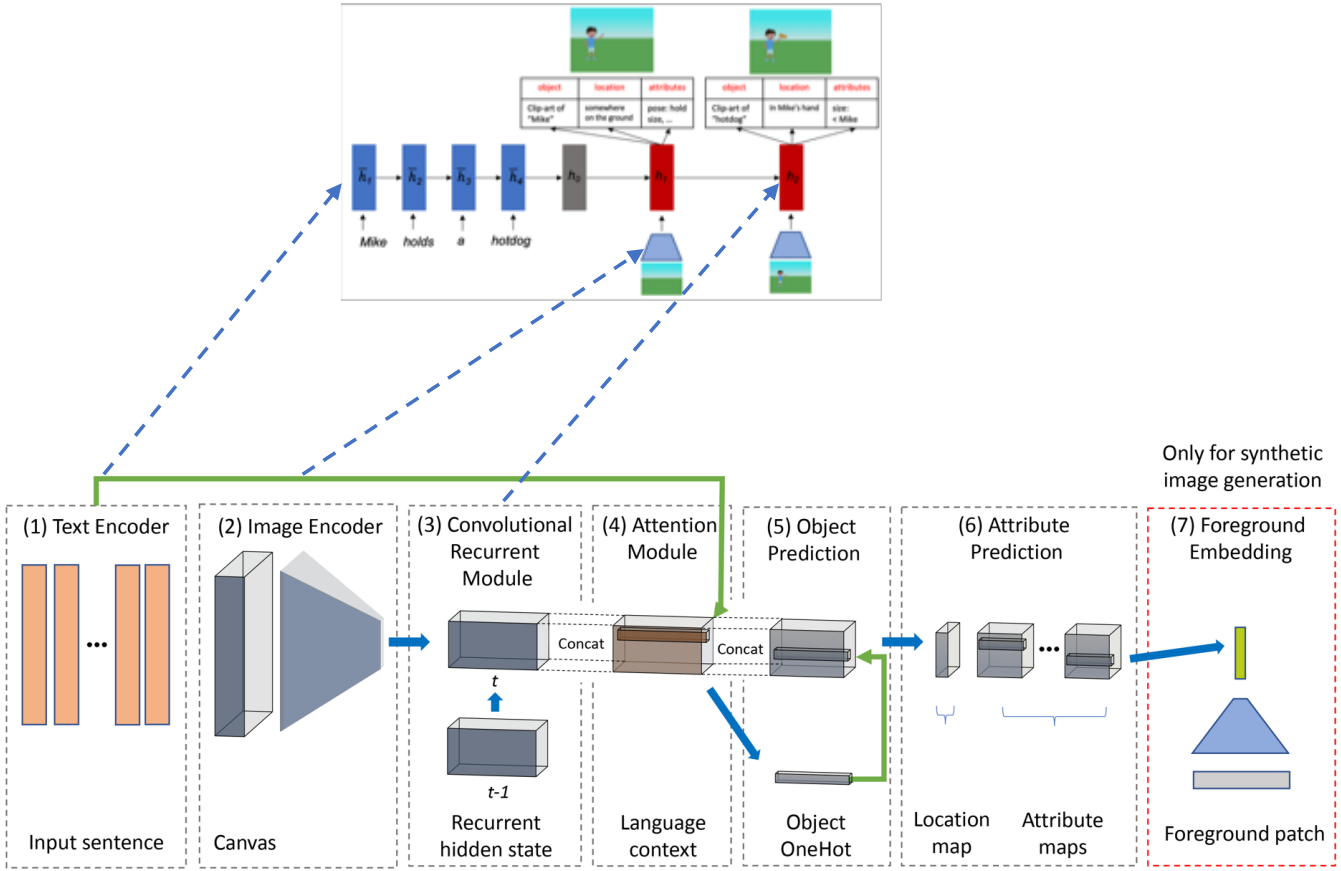
Task 3: Composite Image Generation

Object category → 95 object & stuff categories from COCO: “person”, “grass”, “sky”,…

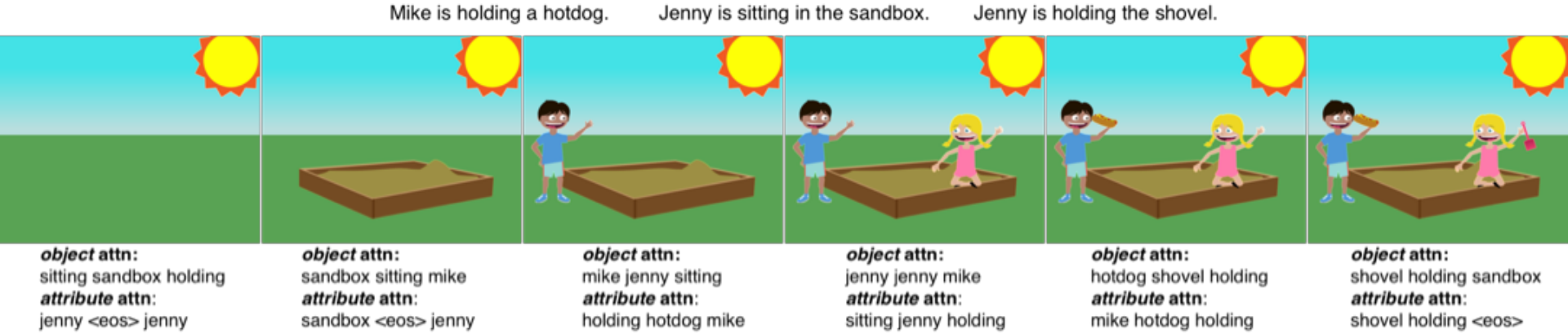
Location → 32 x 32 grid

Attributes → 17 sizes, 17 aspect-ratios

Attributes → a feature vector for patch retrieval


object
location
attributes
Clip-art of "Mike"
somewhere on the ground
pose: hold size, ...
object
location
attributes
Clip-art of "hotdog"
in Mike's hand
size: < Mike
Mike
holds
a
hotdog
Only for synthetic image generation
(1) Text Encoder
Input sentence
(2) Image Encoder
Canvas
(3) Convolutional Recurrent Module
Concat
t
t-1
Recurrent hidden state
(4) Attention Module
Concat
Language context
(5) Object Prediction
Object OneHot
(6) Attribute Prediction
Location map
Attribute maps
(7) Foreground Embedding
Foreground patch

Step-by-step generation of Abstract Scene



Step-by-step generation of composite image

Inputs: a room with TV and some different types of couches.

couches, room, TV

room, couches, <eos>

<eos>, room, types

TV, types, different

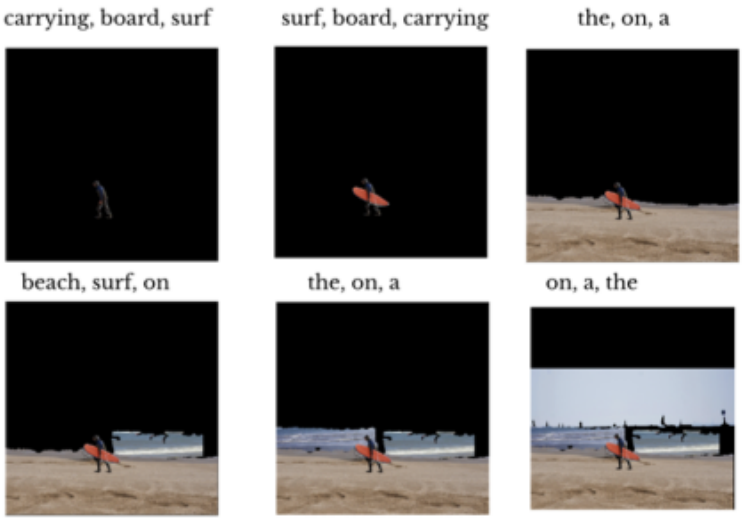
types, of, and

Generated sequences and the top-3 attended words in each step.

Reference image

Step-by-step generation of composite image

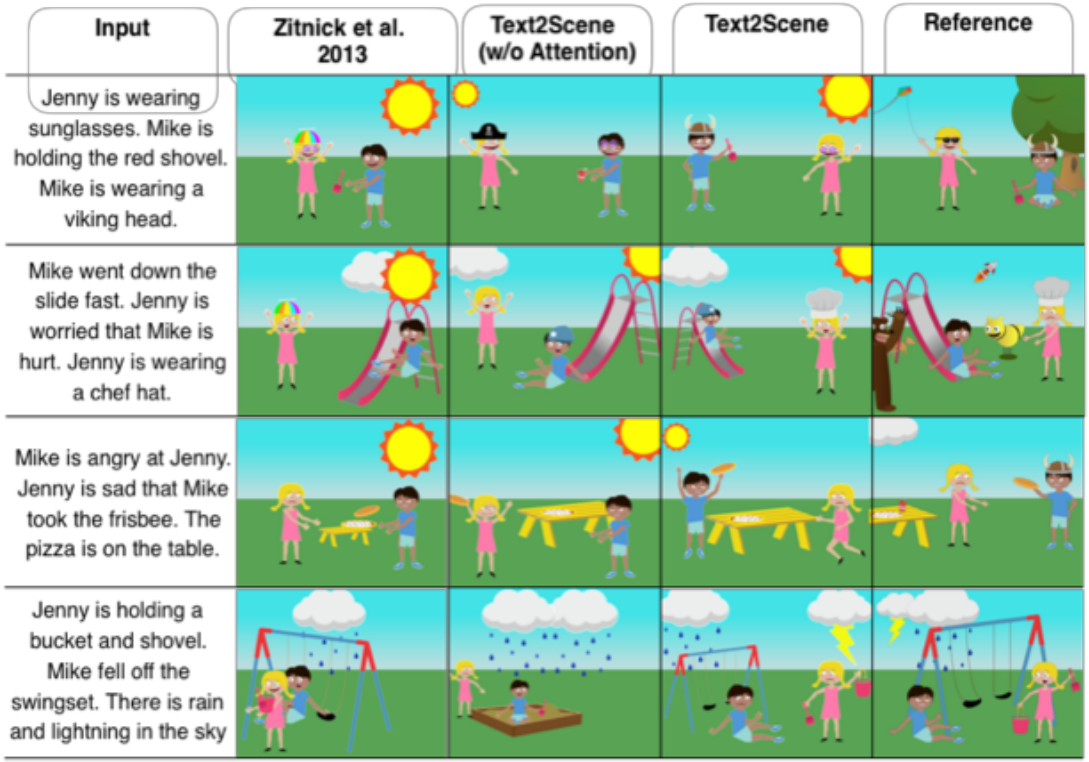
Inputs: a person walks on the beach, carrying a surf board.



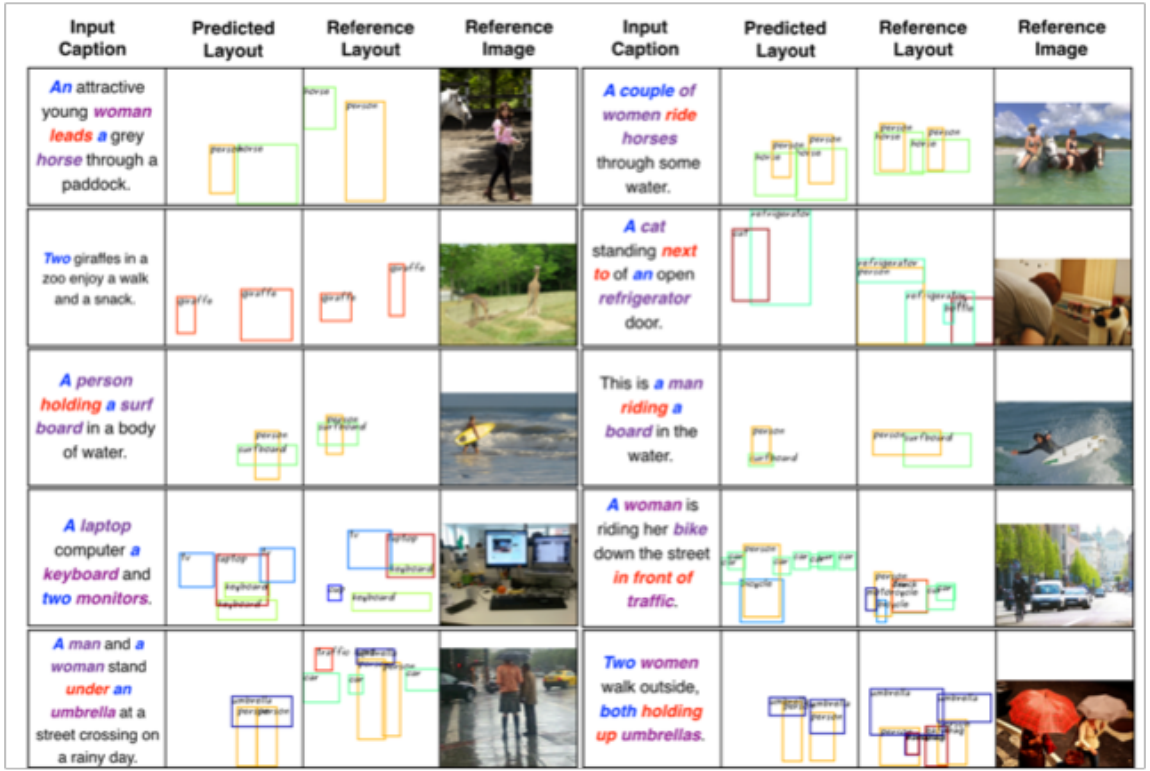
Generated sequences and the top-3 attended words in each step.

Reference image

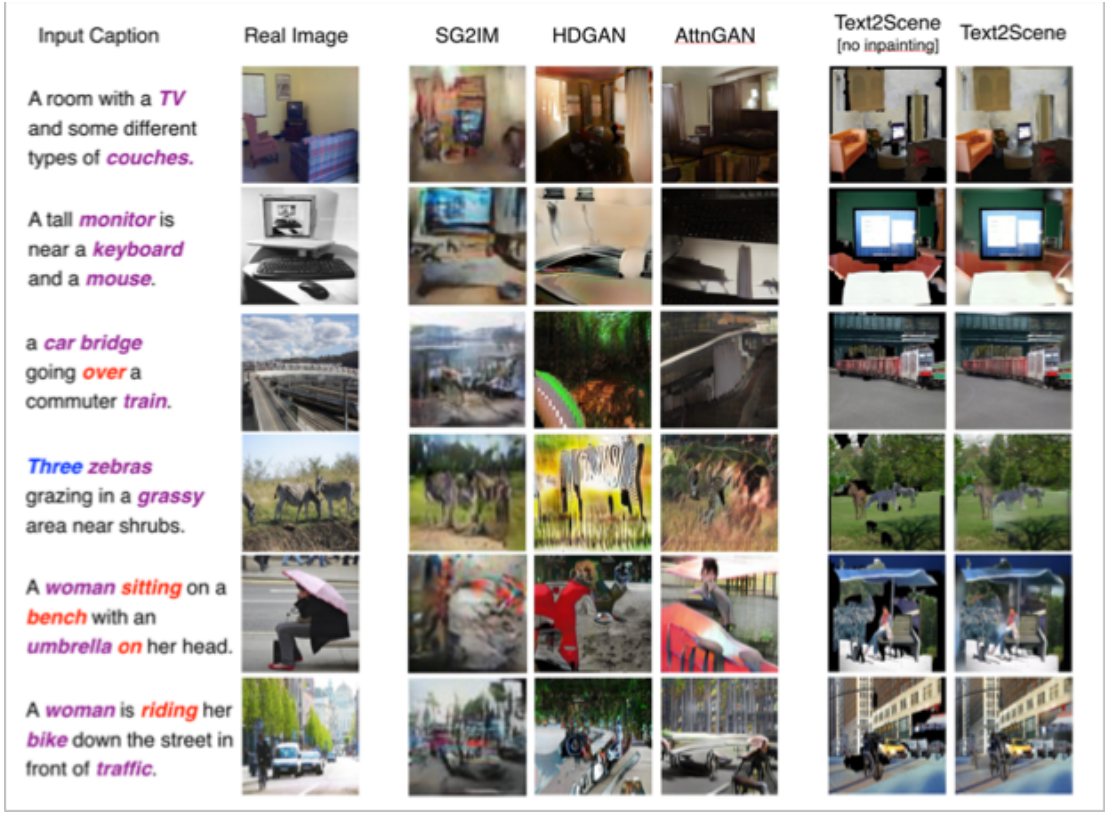
More examples



More examples



More examples



More examples

Questions?