

# Generating Images from Captions with Attention

**Elman Mansimov**

Emilio Parisotto

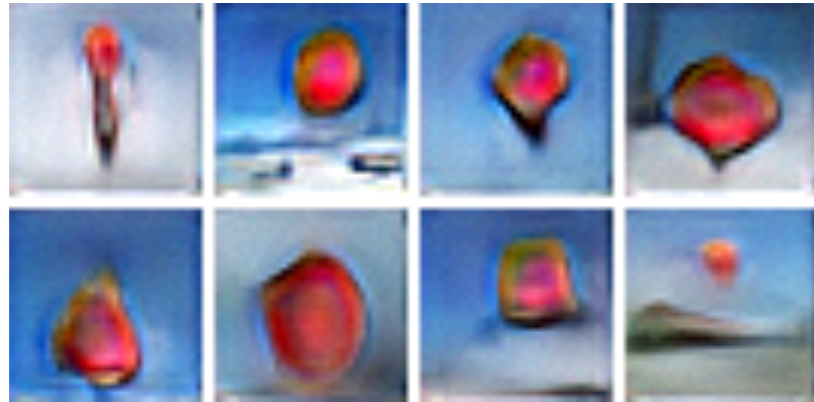
Jimmy Ba

Ruslan Salakhutdinov

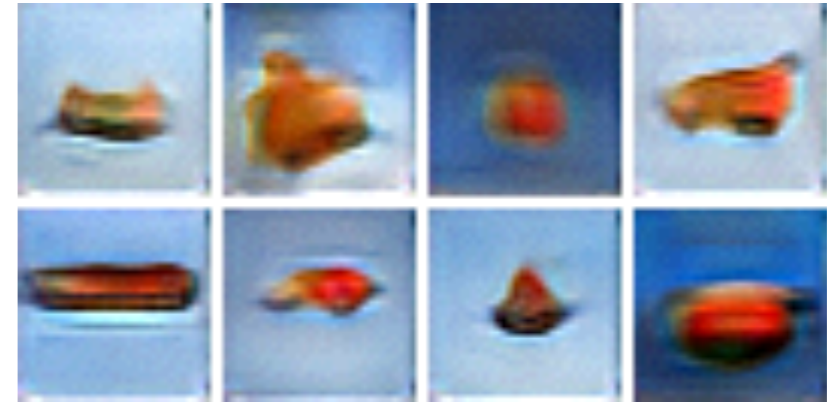
# Motivation

- To simplify the image modelling task
  - Captions contain more information about the image.
  - Although you need to learn language model.
- To better understand model generalization
  - Create textual descriptions of completely new scenes not seen at training time.

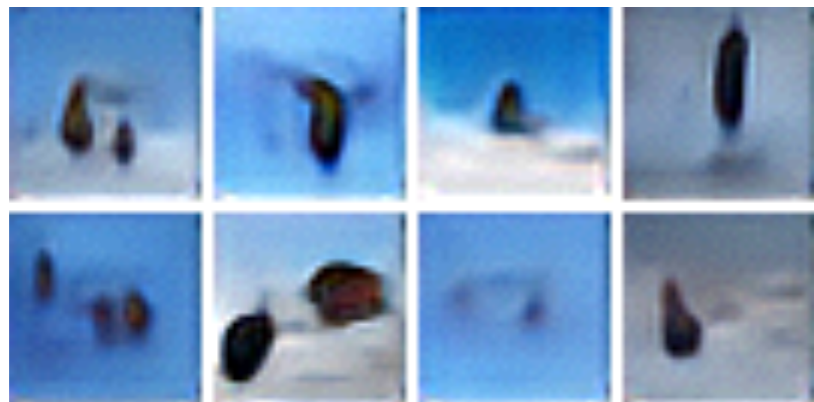
# Novel Compositions



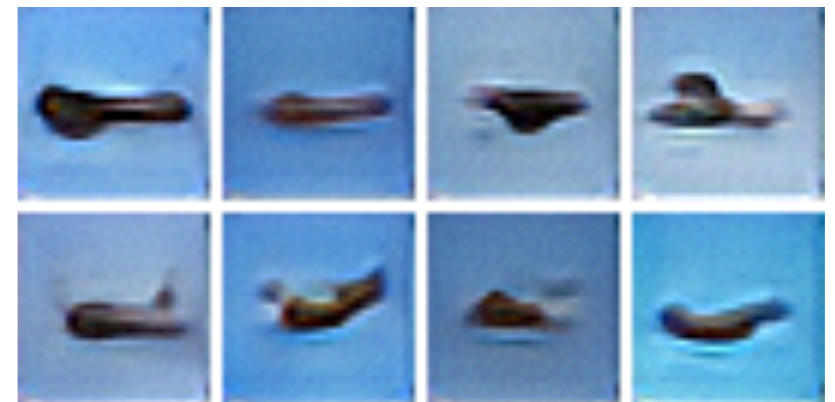
A **stop sign** is flying in blue skies.



A **pale yellow school bus** is flying in blue skies.

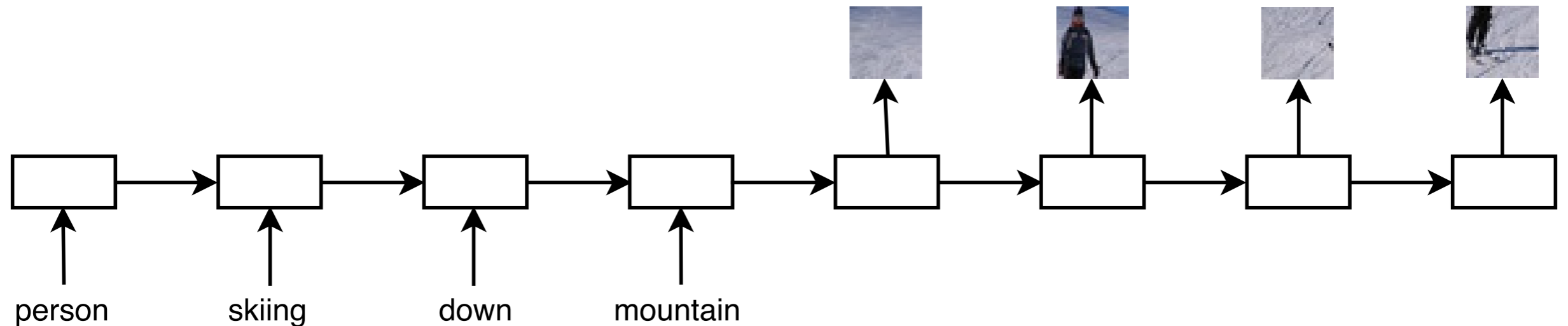


A **herd of elephants** flying in blue skies.



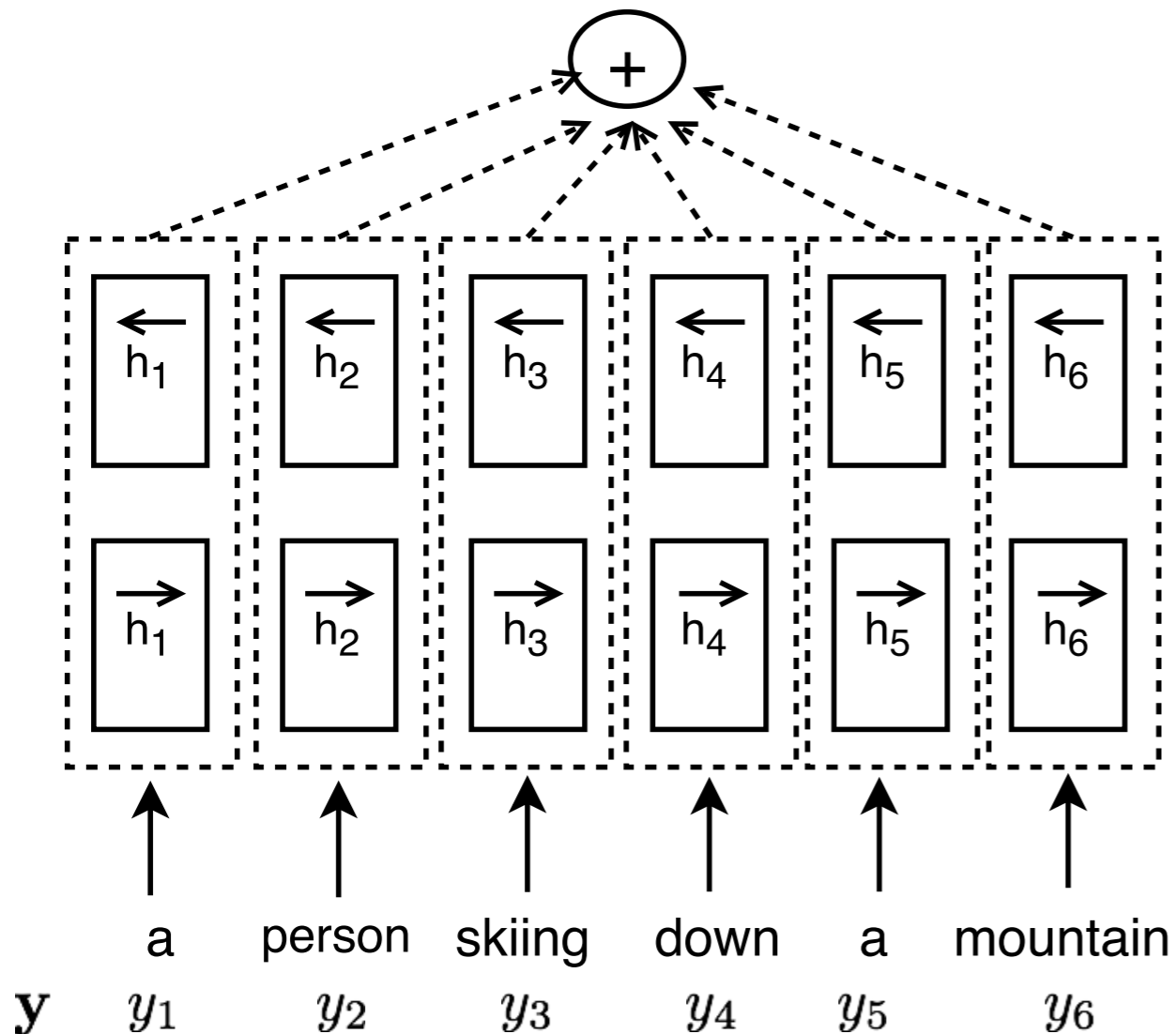
A **large commercial airplane** flying in blue skies.

# General Idea



- Part of the sequence-to-sequence framework. (Sutskever et al. 2014; Cho et al. 2014; Srivastava et al. 2015)
- Caption is represented as a sequence of consecutive words.
- Image is represented as a sequence of patches drawn on canvas.
- Also need to figure out where to put generated patches on canvas.

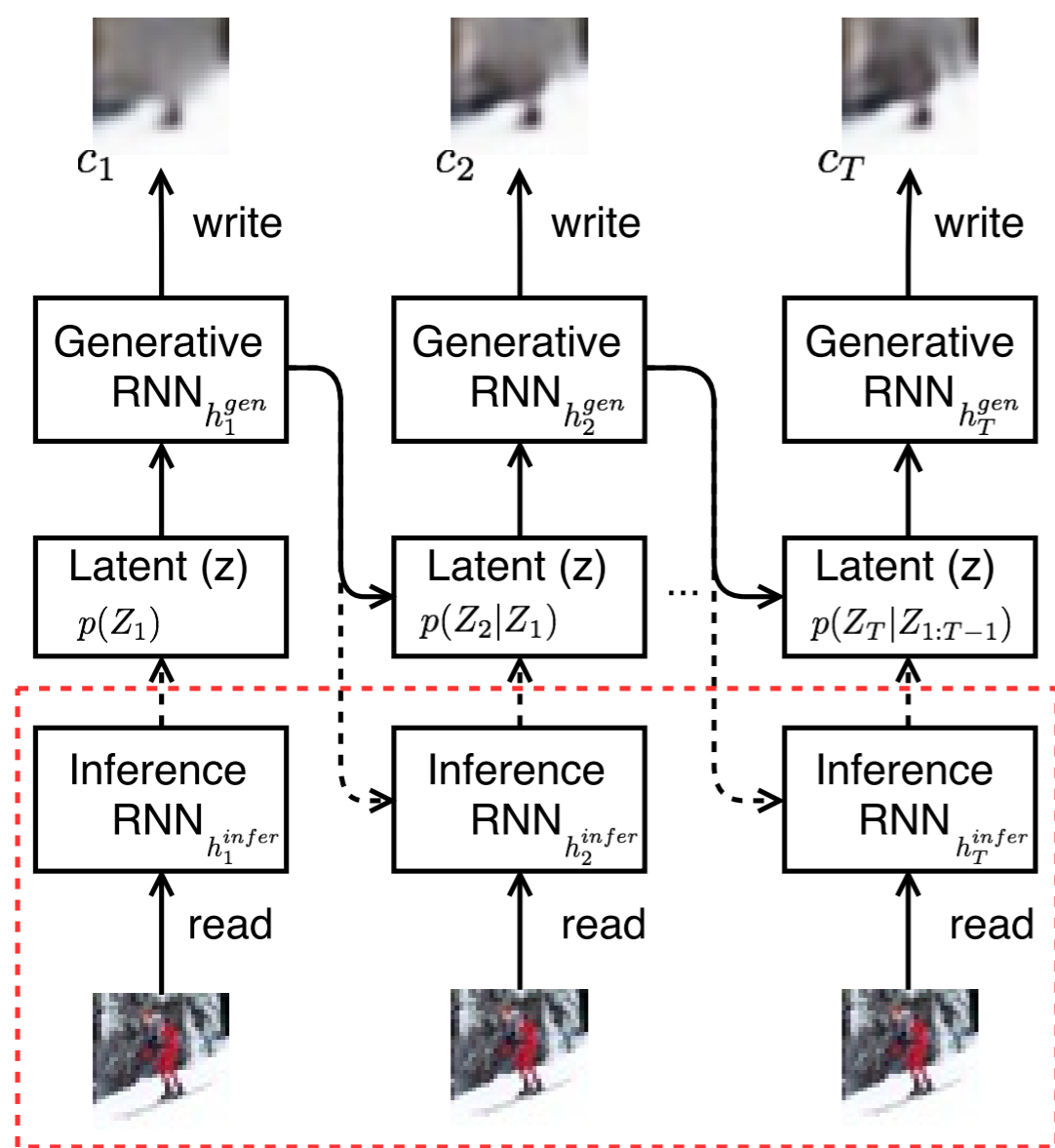
# Language Model (Bidirectional RNN)



- Forward LSTM reads sentence from left to right
- Backward LSTM reads sentence from right to left
- Sentence representation is average of hidden states

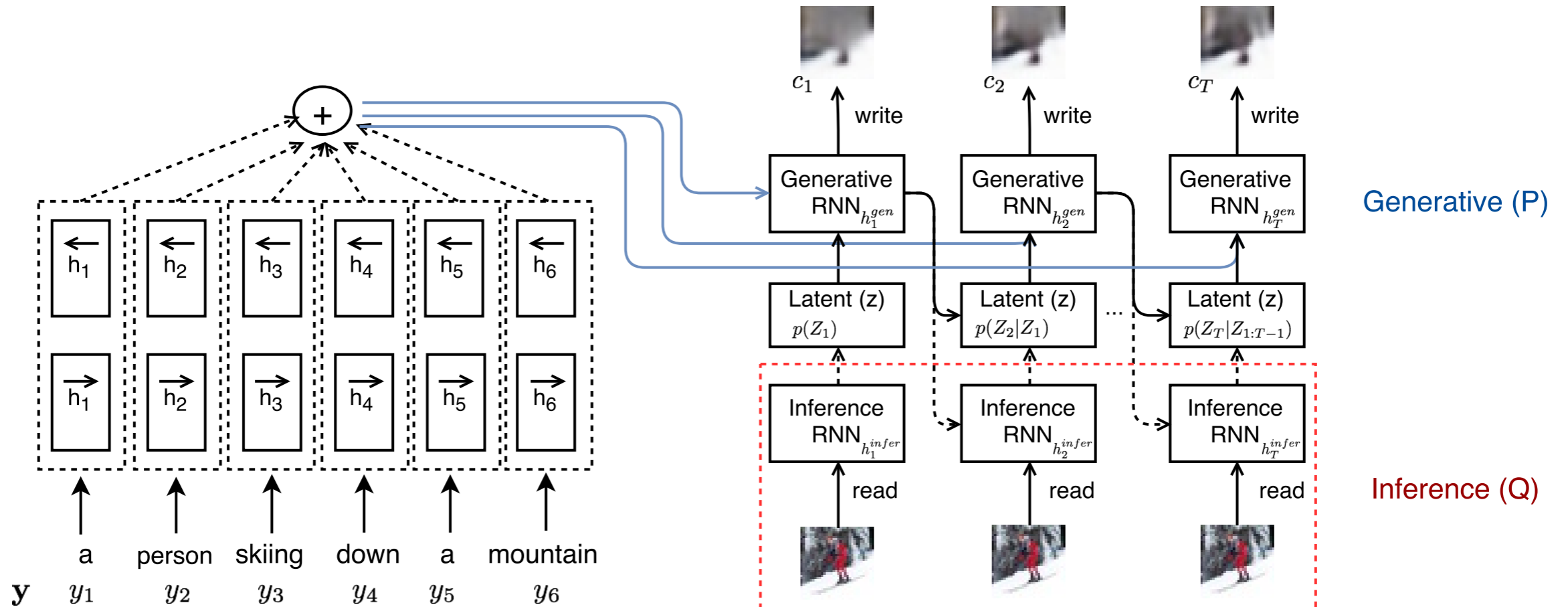
# Image Model

(DRAW: Variational Recurrent Auto-encoder with Visual Attention)



- At each step model produces  $p \times p$  patch.
- It gets transformed into  $h \times w$  canvas using two arrays of 1D filter banks ( $h \times p$  and  $w \times p$  respectively).
- Mean and variance of latent variables depend on the previous hidden states of generative RNN.

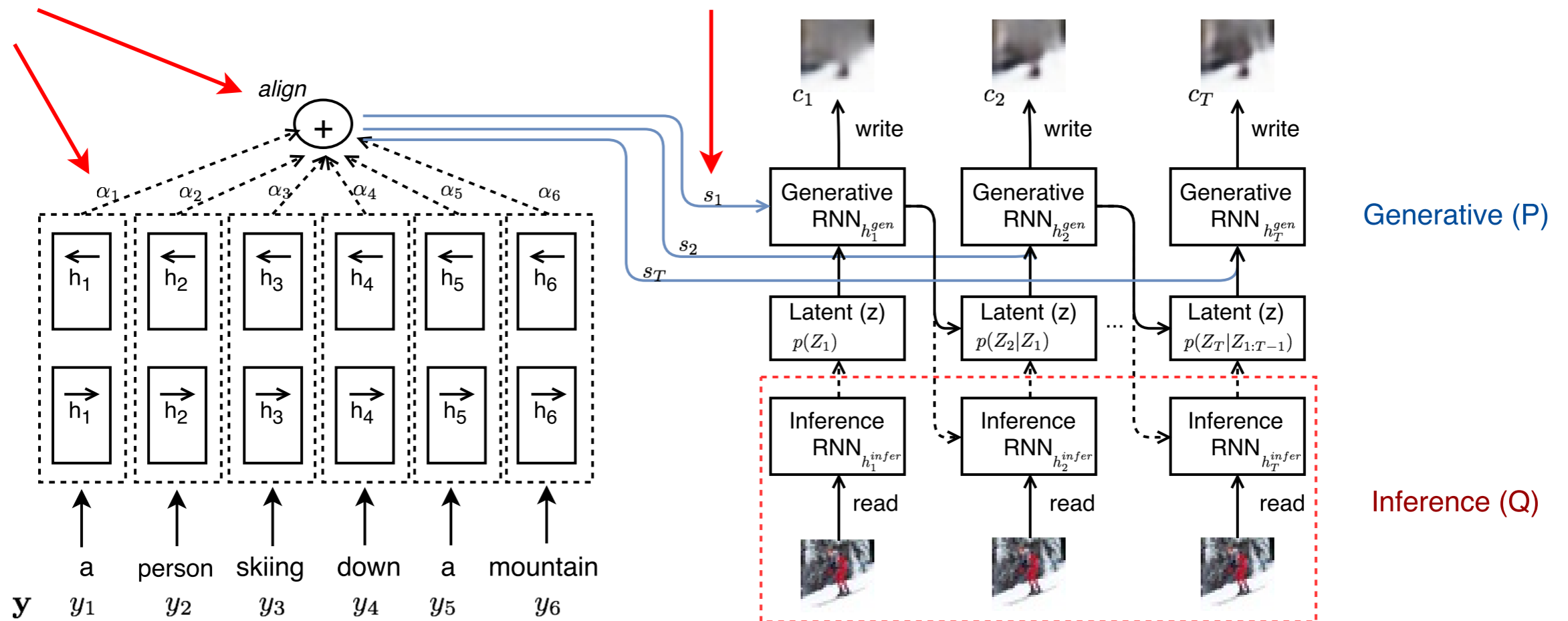
# Model



Model is trained to maximize variational lower bound

$$\mathcal{L} = \mathbb{E}_{Q(Z_{1:T} | \mathbf{y}, \mathbf{x})} \left[ \log p(\mathbf{x} | \mathbf{y}, Z_{1:T}) - \sum_{t=2}^T D_{\text{KL}} (Q(Z_t | Z_{1:t-1}, \mathbf{y}, \mathbf{x}) || P(Z_t | Z_{1:t-1}, \mathbf{y})) \right] - D_{\text{KL}} (Q(Z_1 | \mathbf{x}) || P(Z_1))$$

# Alignment

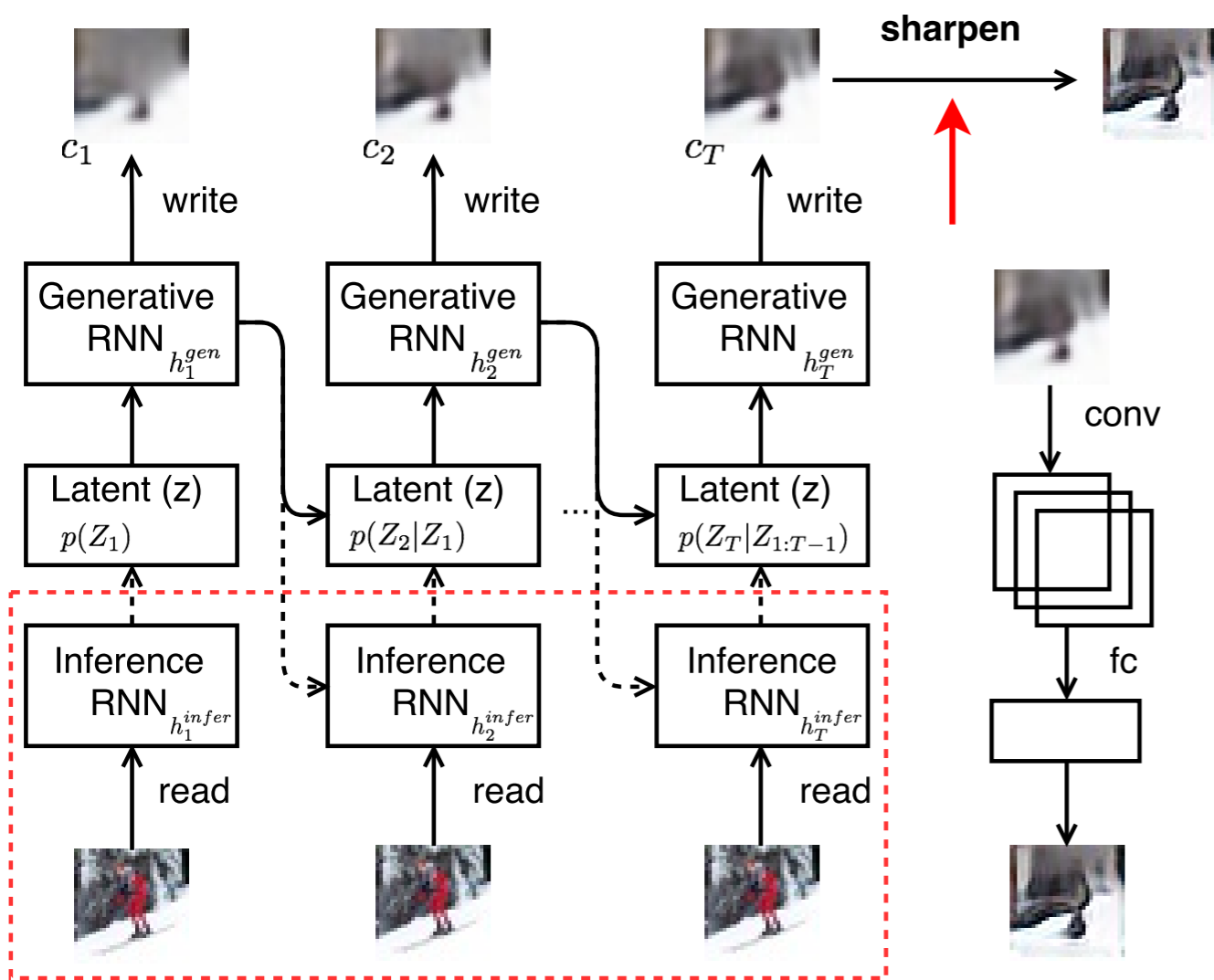


Compute alignment between words and generated patches

$$e_j^t = v^\top \tanh(Uh_j^{lang} + Wh_{t-1}^{gen} + b) \quad \alpha_j^t = \frac{\exp(e_j^t)}{\sum_{j=1}^N \exp(e_j^t)}$$

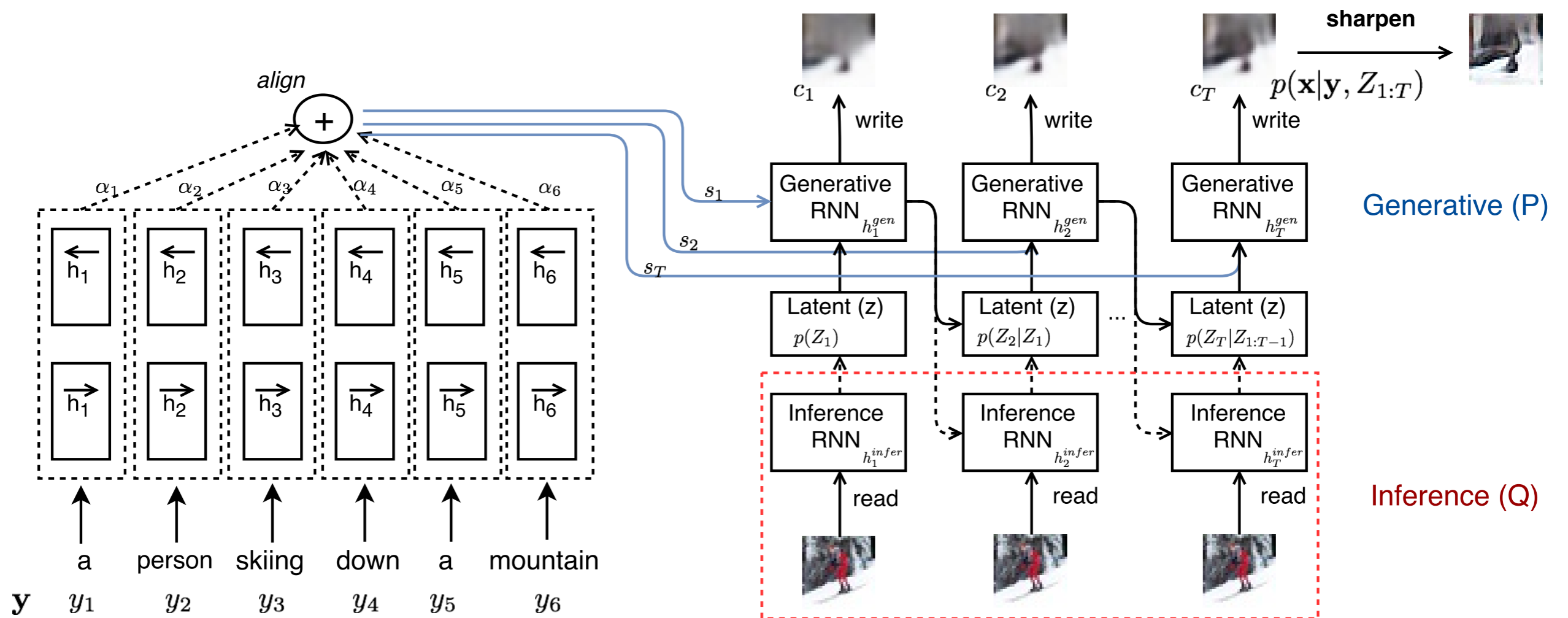


# Sharpening

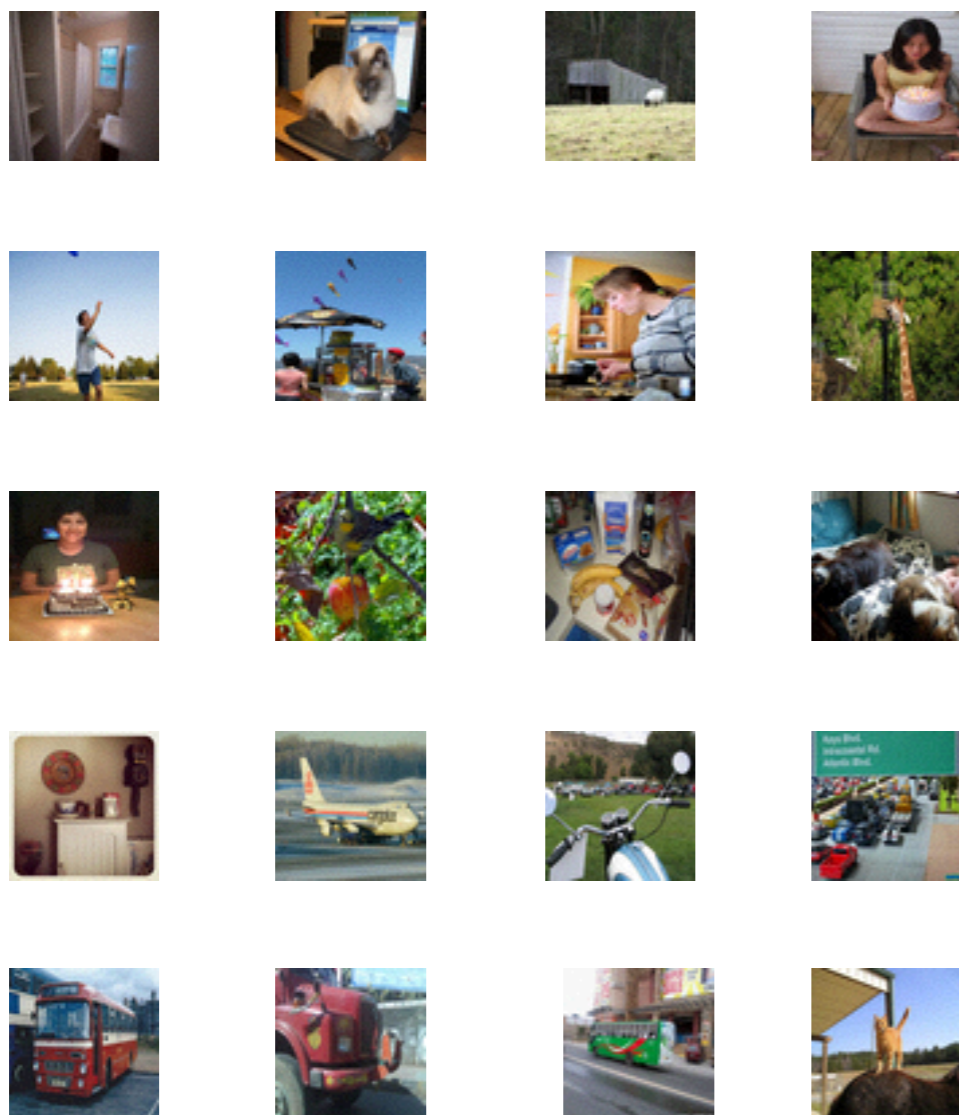


- Another network trained to generate edges sharpens the generated samples.
- Instead is trained to fool separate network that discriminates between real and fake samples.
- Doesn't have reconstruction cost and gets sharp edges.

# Complete Model



# Main Dataset (Microsoft COCO)



- Contains ~83k images
- Each image has 5 captions
- Standard benchmark dataset for recent image captioning systems

# Flipping Colors



**A yellow school bus**  
parked in a parking lot.



**A red school bus**  
parked in a parking lot.

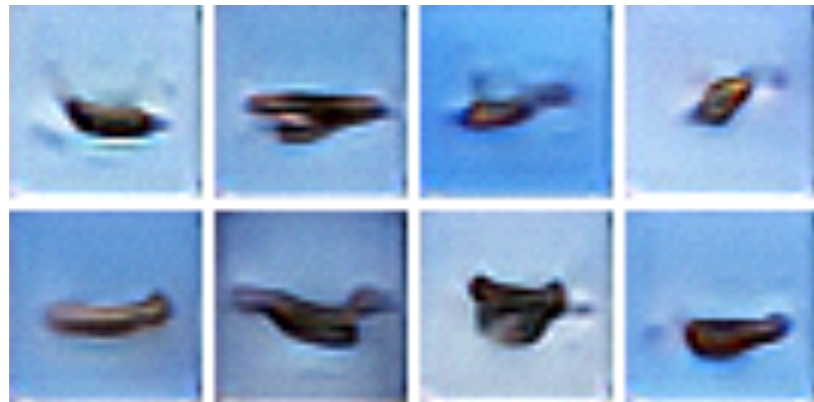


**A green school bus**  
parked in a parking lot.



**A blue school bus**  
parked in a parking lot.

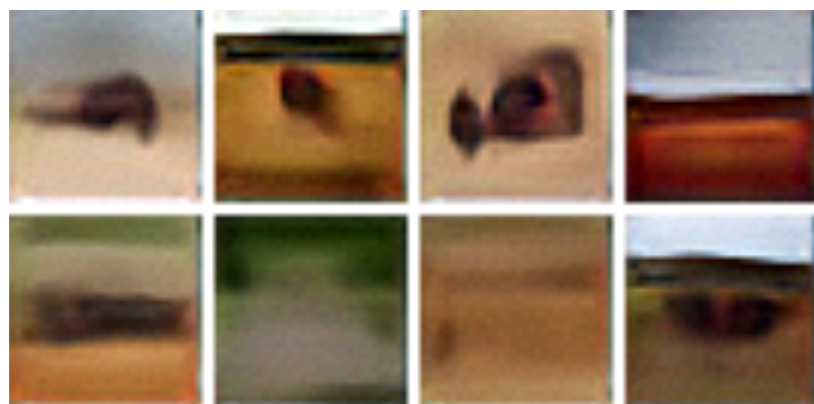
# Flipping Backgrounds



A very large commercial plane flying **in clear skies**.



A very large commercial plane flying **in rainy skies**.



A herd of elephants walking across a **dry grass field**.



A herd of elephants walking across a **green grass field**.

# Flipping Objects



**The decadent chocolate dessert** is on the table.



**A bowl of bananas** is on the table.

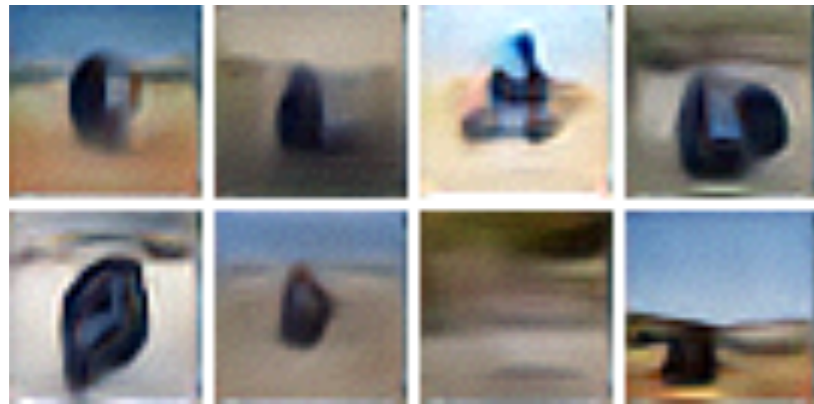


A vintage photo of **a cat**.



A vintage photo of **a dog**.

# Examples of Alignment



A rider on the blue motorcycle in the desert.



A rider on the blue motorcycle in the forest.



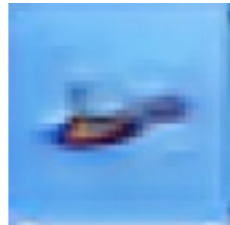
A surfer, a woman, and a child walk on the beach.



A surfer, a woman, and a child walk on the sun.

# text2image <-> image2text

A very large commercial plane flying in clear skies.



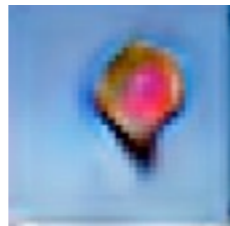
A large airplane flying through a blue sky.



machine generated caption



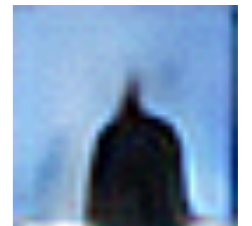
A stop sign is flying in blue skies.



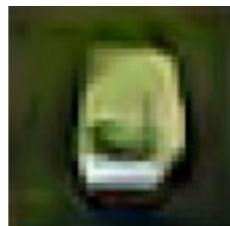
A picture of a building with a blue sky.



machine generated caption



A toilet seat sits open in the grass field.



A window that is in front of a mirror.



machine generated caption





# More Results

## (Image Retrieval and Image Similarity)

Model	R@1	R@5	R@10	R@50	Med r	SSI
LAPGAN	-	-	-	-	-	0.08
Fully-Conn VAE	1.0	6.6	12.0	53.4	47	0.156
Conv-Deconv VAE	1.0	6.5	12.0	52.9	48	0.164
skipthoughtDRAW	2.0	11.2	18.9	63.3	36	0.157
noalignDRAW	2.8	14.1	23.1	68.0	31	0.155
alignDRAW	3.0	14.0	22.9	68.5	31	0.156

# Lower Bound of Log-Likelihood in Nats

Model	Train	Test	Test (after sharpening)
skipthoughtDRAW	-1794.29	-1791.37	-2045.84
noalignDRAW	-1792.14	-1791.15	-2051.07
alignDRAW	-1792.15	-1791.53	-2042.31

# Qualitative Comparison



Our Model



LAPGAN



Conv-Deconv VAE



Fully-Connected VAE

*A group of people walk on a beach with surf boards*

# Conclusions

- Samples from our generative model are okay; but aren't great.
- Mostly because the model is underfitting.
- The model generalizes to captions describing novel scenarios that are not seen in the dataset.
- Key factor, treat image generation as computer graphics. Learn what to generate and where to place it.

# Thank You!

Paper: <http://arxiv.org/abs/1511.02793>

Code: <https://github.com/emansim/text2image>