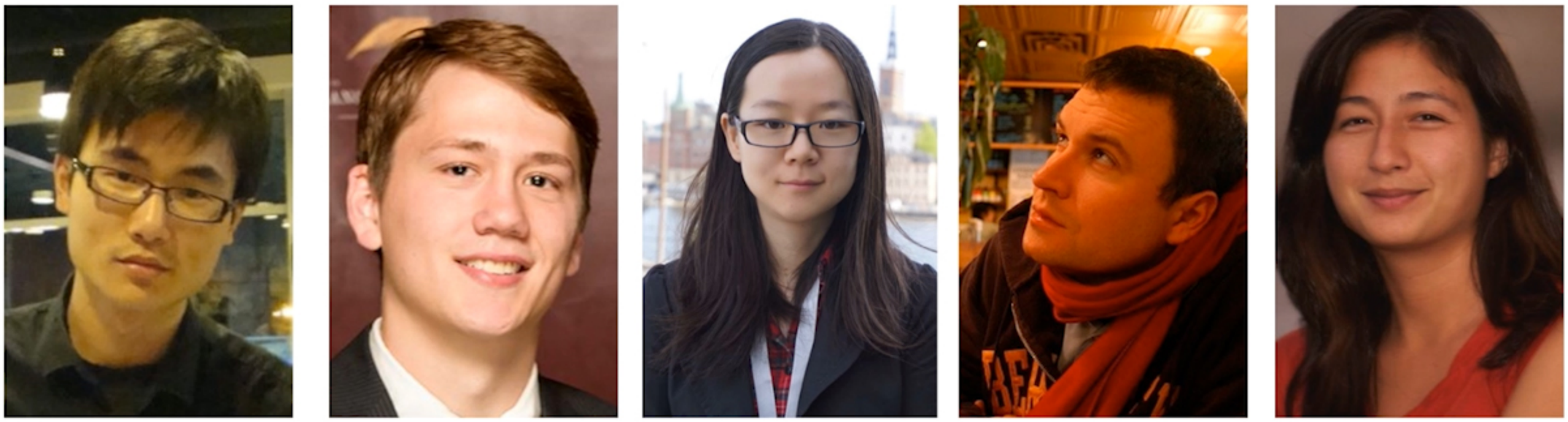
Modeling Context in Referring Expression

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, Tamara L. Berg



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Referring Expression

Referring Expression Game

woman washing dishes

ReferIt Game, EMNLP2014

Referring Expression Game

woman washing dishes

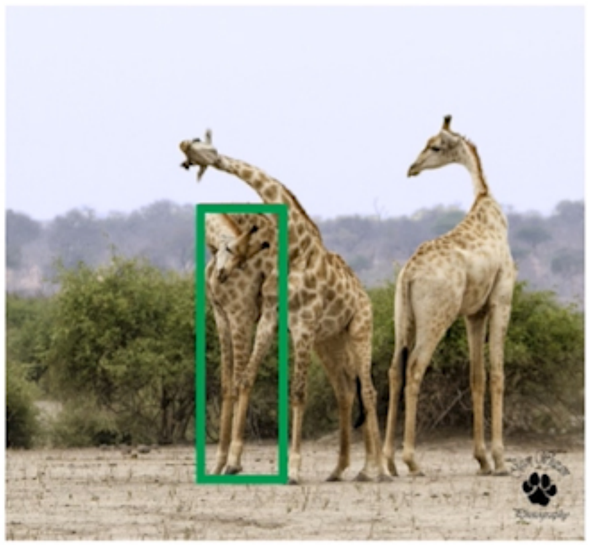
ReferIt Game, EMNLP2014

Datasets

- **RefCOCO** UNC
  - 142210 referring expressions for 50000 objects in 19994 images
- **RefCOCO+** UNC
  - 141564 referring expressions for 49856 objects in 19992 images
- **RefCOCOg** Google
  - 104560 referring expressions for 54822 objects in 26711 images

UNC's datasets are available at https://github.com/lichengunc/refer

Difference between three datasets



RefCOCO: giraffe on the left

RefCOCO+: giraffe with lowered head down

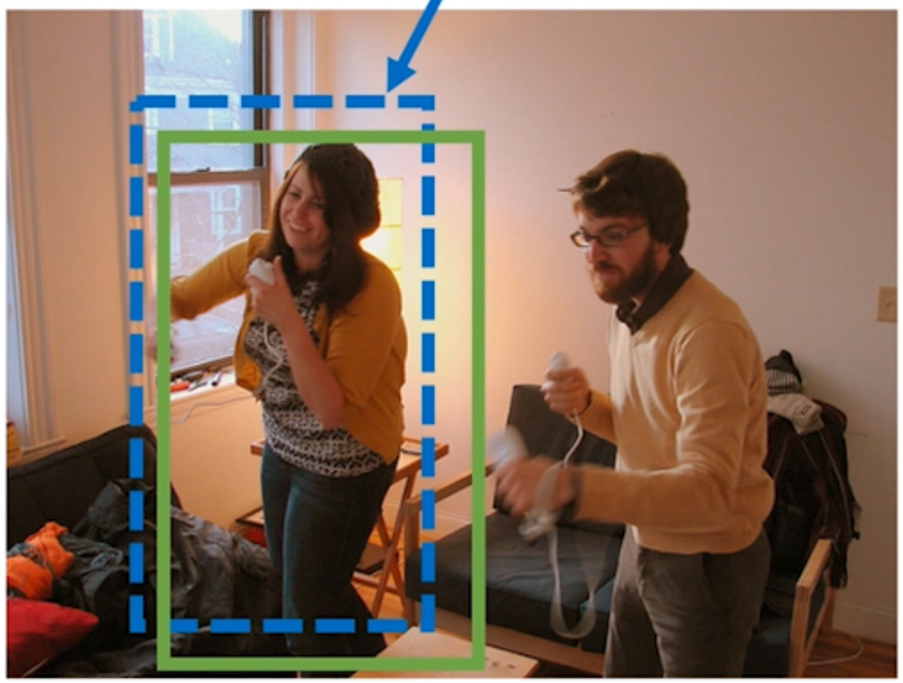
RefCOCOg: an adult giraffe scratching its back with its horn

Dataset	Collection way	Expression Style	Allow location words
RefCOCO (UNC)	Interactive Game	Free style	yes
RefCOCO+ (UNC)	Interactive Game	Free style	no
RefCOCOg (Google)	Non-interactive	COCO-caption style	yes

Two tasks

Task 1: comprehension

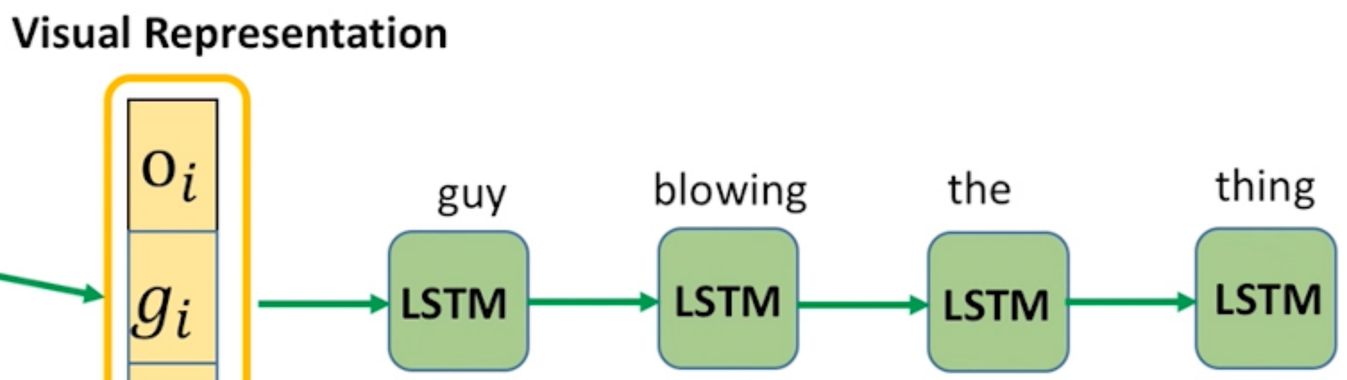
Which object is "**Girl on the left**" indicating?



Task 2: Expression Generation

Generate referring expression for this target person.

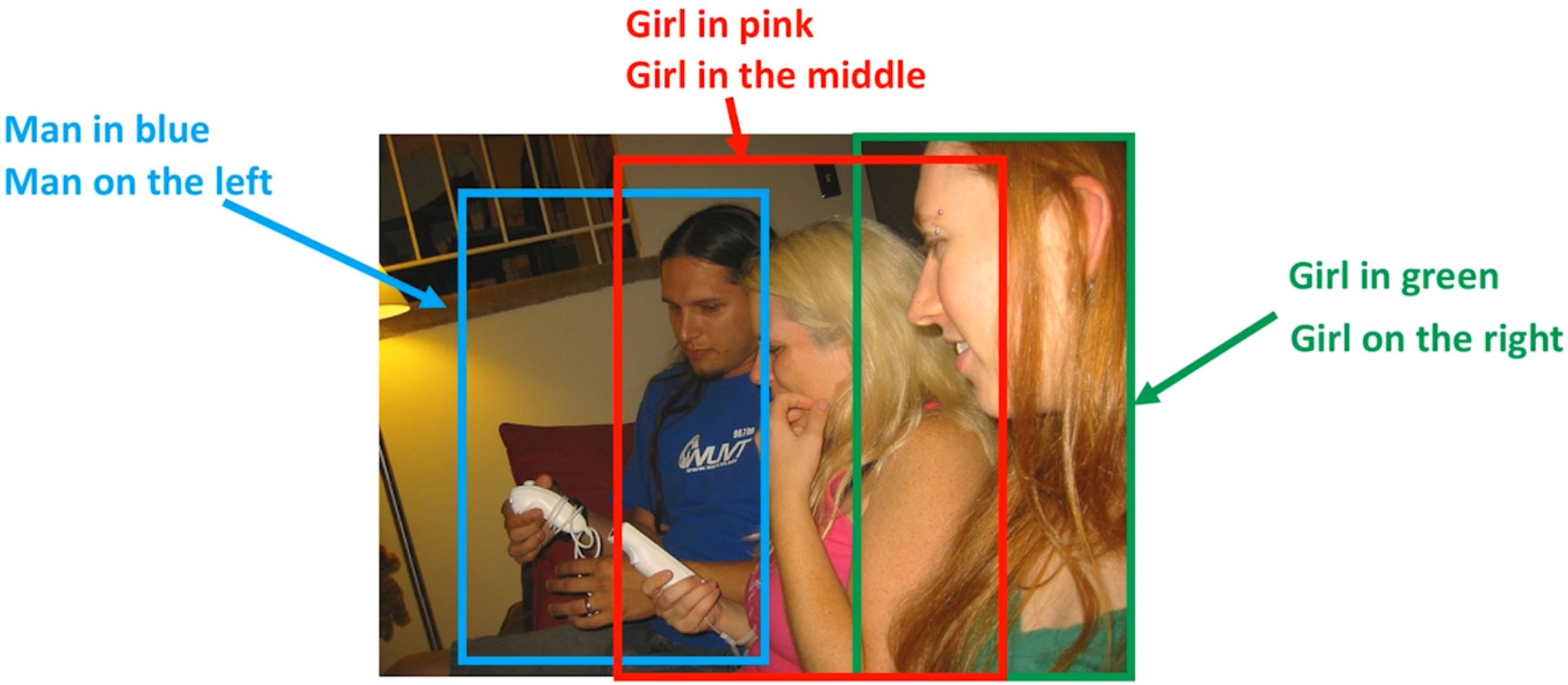
Baseline model



$$(o_i, g_i, l_i) = \left(\text{CNN}, \text{CNN}, [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{w}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}] \right)$$

Generation and Comprehension of Unambiguous Object Descriptions, CVPR2016

Our model: visual comparison

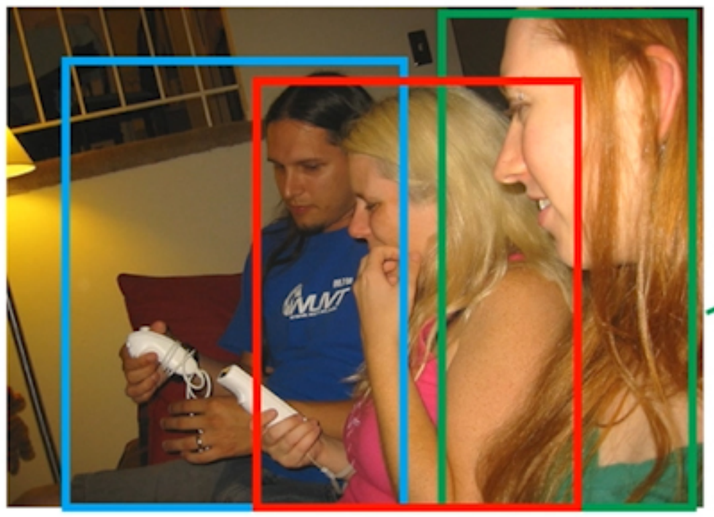


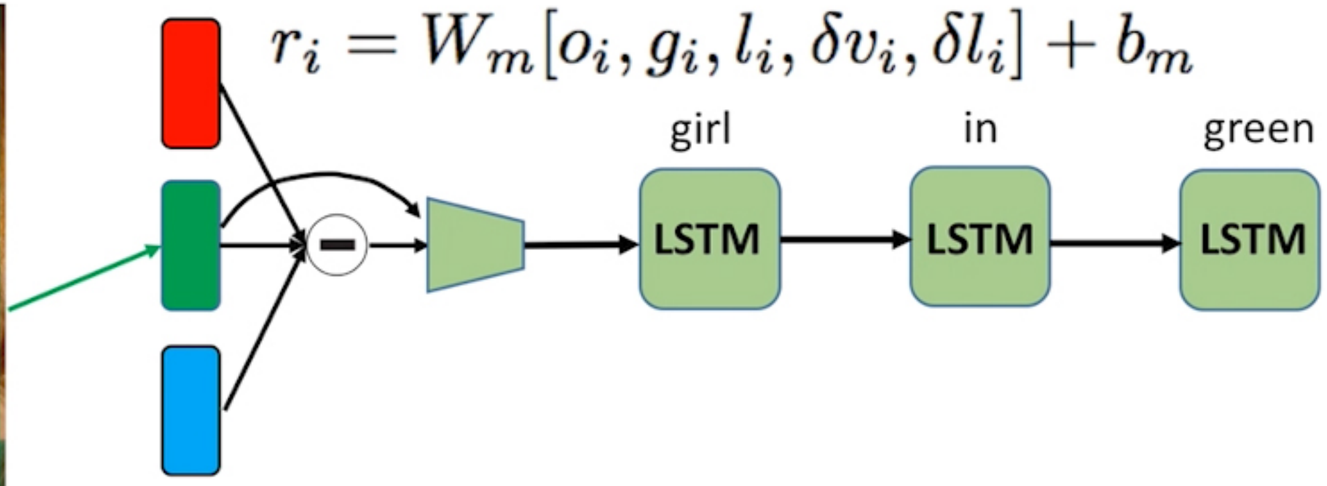
Visual Comparison

- Object comparisons are critical for producing unambiguous referring expression

$$\delta v_i = \frac{1}{n}\sum_{j\neq i}\frac{o_i - o_j}{\|o_i - o_j\|}$$

$$\delta l_{ij} = [\frac{[\triangle x_{tl}]_{ij}}{w_i}, \frac{[\triangle y_{tl}]_{ij}}{h_i}, \frac{[\triangle x_{br}]_{ij}}{w_i}, \frac{[\triangle y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}]$$



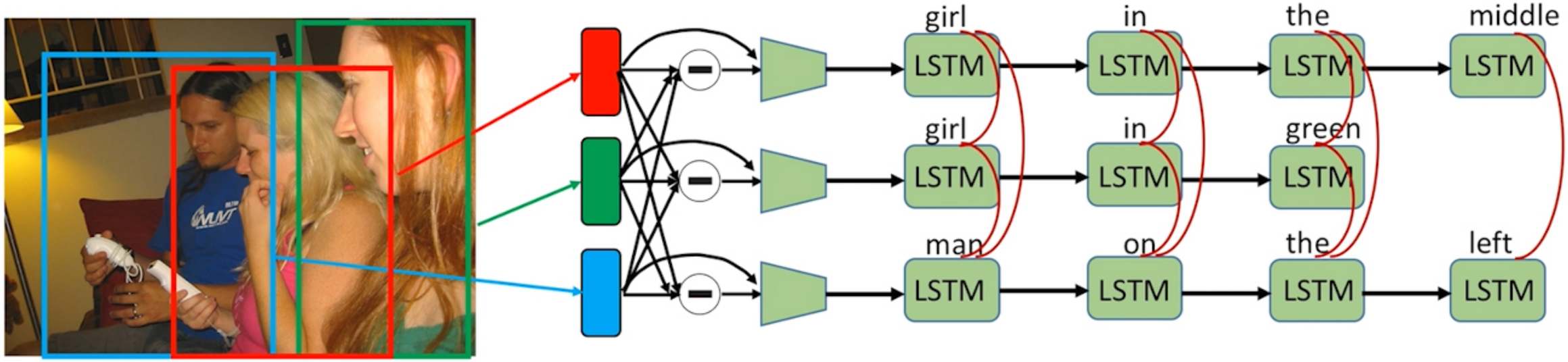


Language Tying

- Tying language generation together by adding the connections between expressions.

$$h_{dif_{i_t}} = \frac{1}{n} \sum_{j \neq i} \frac{h_{i_t} - h_{j_t}}{\|h_{i_t} - h_{j_t}\|}$$

$$P(w_{i_t} | w_{i_{t-1}}, ..., w_{i_1}, v_i, \{h_{j_t, j \neq i}\}) = \text{softmax}(W_h[h_{i_t}, h_{dif_{i_t}}] + b_h)$$



Task1: Comprehension

	RefCOCO		RefCOCO+		RefCOCOg
	Test A	Test B	Test A	Test B	Validation
Baseline[22]	63.15%	64.21%	48.73%	42.13%	55.16%
visdif	67.57%	71.19%	52.44%	47.51%	59.25%
MMI[22]	71.72%	71.09%	58.42%	51.23%	62.14%
visdif+MMI	**73.98%**	**76.59%**	**59.17%**	**55.62%**	**64.02%**

Task2: Generation

RefCOCO

	Test A				Test B			
	Bleu 1	Bleu 2	Rouge	Meteor	Bleu 1	Bleu 2	Rouge	Meteor
Baseline [22]	0.477	0.290	0.413	0.173	0.553	0.343	0.499	0.228
MMI [22]	0.478	0.295	0.418	0.175	0.547	0.341	0.497	0.228
visdif	0.505	**0.322**	0.441	0.184	0.583	0.382	0.530	0.245
visdif+MMI	0.494	0.307	0.441	0.185	0.578	0.375	0.531	0.247
Baseline+tie	0.490	0.308	0.431	0.181	0.561	0.352	0.505	0.234
visdif+tie	**0.510**	0.318	**0.446**	**0.189**	**0.593**	**0.386**	**0.533**	**0.249**
visdif+MMI+tie	0.506	0.312	0.445	0.188	0.579	0.370	0.525	0.246

RefCOCO+

	Test A				Test B			
	Bleu 1	Bleu 2	Rouge	Meteor	Bleu 1	Bleu 2	Rouge	Meteor
Baseline [22]	0.391	0.218	0.356	0.140	0.331	0.174	0.322	0.135
MMI [22]	0.370	0.203	0.346	0.136	0.324	0.167	0.320	0.133
visdif	0.407	0.235	0.363	0.145	0.339	0.177	0.325	0.145
visdif+MMI	0.386	0.221	0.360	0.142	0.327	0.172	0.325	0.135
Baseline+tie	0.392	0.219	0.361	0.143	0.336	0.177	0.325	0.140
visdif+tie	**0.409**	**0.232**	**0.372**	**0.150**	**0.340**	**0.178**	**0.328**	**0.143**
visdif+MMI+tie	0.393	0.220	0.360	0.142	0.327	0.175	0.321	0.137

RefCOCO

Ground-truth
Prediction

RefCOCO+

guy in white on far right

blurry person
with sleeveless and sitting

RefCOCO

head on left

woman in middle

person on right

RefCOCO+

closest sheep

sheep behind other sheep

Thank you!

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL