Boosted Attention: Leveraging Human Attention for Image Captioning

Shi Chen Qi Zhao
University of Minnesota
chen4595@umn.edu qzhao@cs.umn.edu

Introduction

Motivation. Visual attention has shown usefulness in image captioning, with the goal of enabling a caption model to selectively focus on regions of interest. Existing models typically rely on top-down language information and learn top-down model attention implicitly by optimizing the captioning objectives. While somewhat effective, the learned top-down attention can fail to focus on correct regions of interest without direct supervision of attention (see Figure 1).

Methodology. Inspired by the human visual system which is driven by not only the task-specific top-down signals but also the visual stimuli, we in this work proposed a Boosted Attention module that combines both types of attention for image captioning. With proposed fusion method, we show that two attention play complementary roles on attending regions of interest.

Contributions:

- Boosted Attention model that combines human stimulus-based attention and top-down model attention for image captioning.
- Attention fusion enabling different attention to be complementary
- Analysis on the corporation between two types of attention.

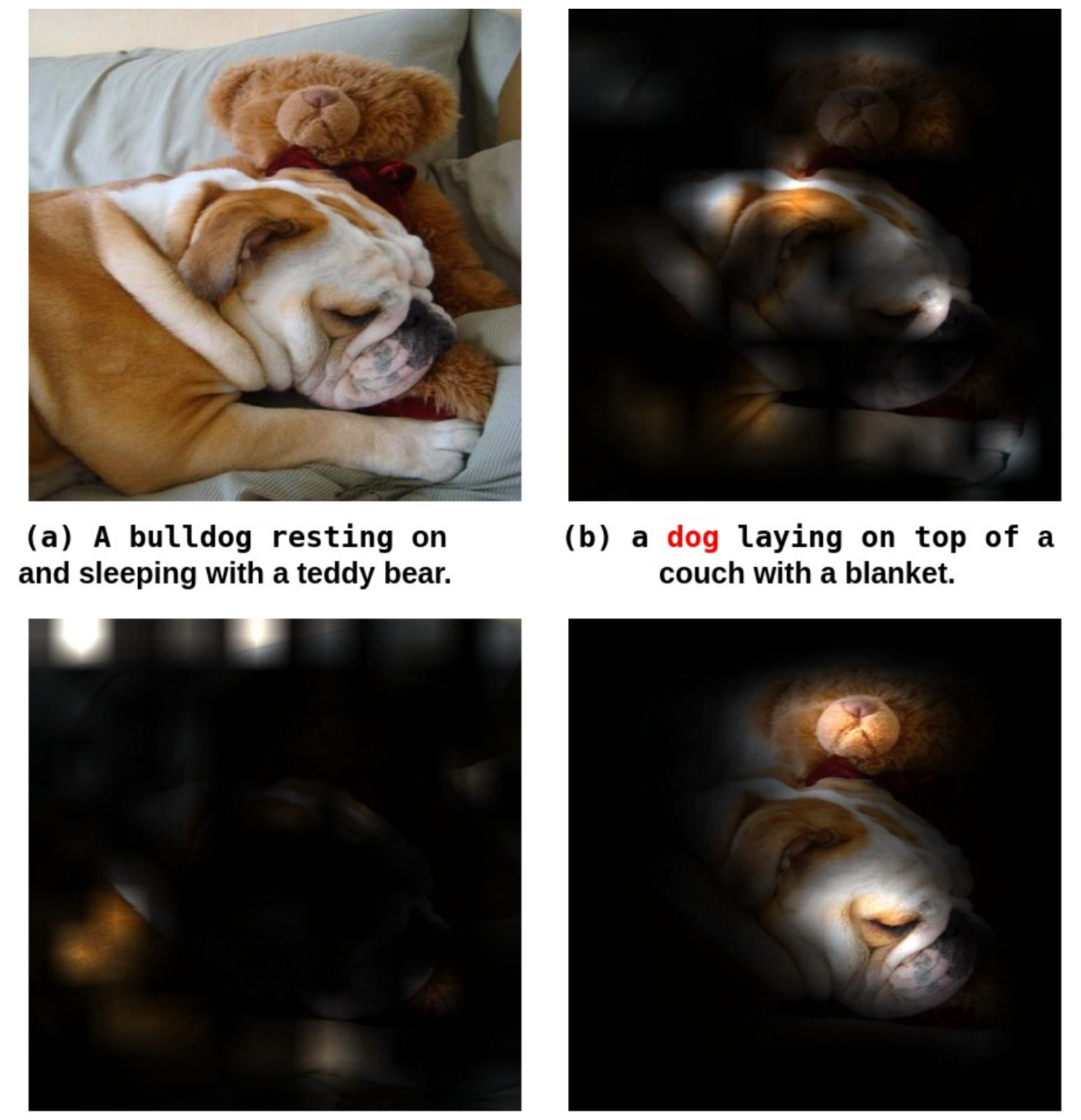
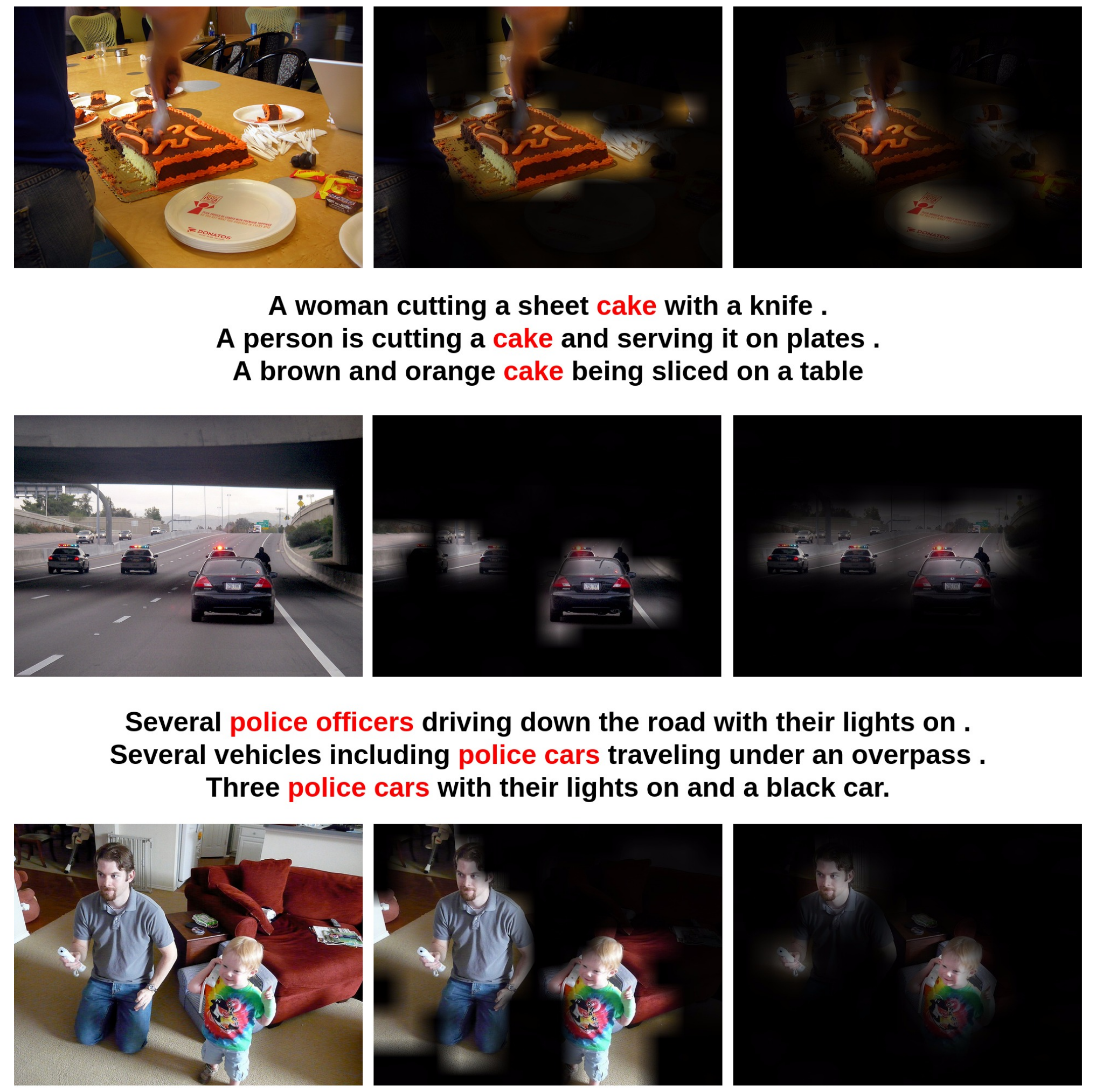

(a) A bulldog resting on and sleeping with a teddy bear.
(b) a dog laying on top of a couch with a blanket.
(c) a dog laying on top of a couch with a blanket.
(d)

Figure 1. (a) Input image with ground truth caption, (b-d) top-down attention maps with model generated captions (word associated with attention highlighted in red), (d) saliency map for human stimulus-based attention.

Role of Stimulus-based Human Attention in Image Captioning


A woman cutting a sheet cake with a knife .
A person is cutting a cake and serving it on plates .
A brown and orange cake being sliced on a table

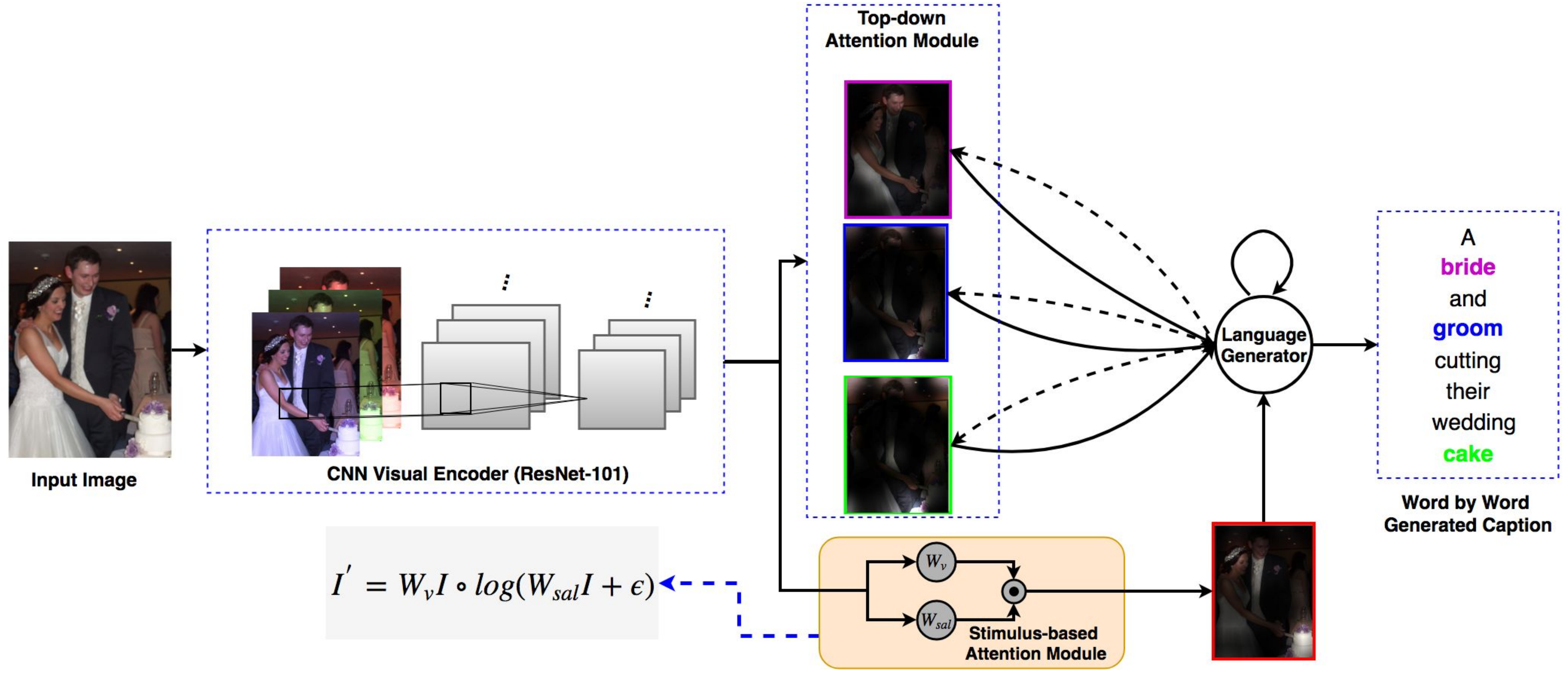
Several police officers driving down the road with their lights on .
Several vehicles including police cars traveling under an overpass .
Three police cars with their lights on and a black car.

A adult and a child with remotes in a room .
A man kneeling on a floor next to a little boy .
A man plays Wii with a young boy in a living room .

Figure 2. Comparison between captioning attention maps (2nd column) and corresponding saliency maps (3rd column) for input images (1st column). Three human generated captions are shown for each pair with frequently mentioned objects highlighted in red.

- **Goal**: Analyzing correlations between attention necessary for caption generation (captioning attention) and human stimulus-based attention.
- **Data**: SALICON [11] dataset with 10000 (training) + 5000 (validation) images from MSCOCO and corresponding ground truth saliency maps. Captioning attention maps generated based on ground truth captions and object bounding box.
- **Observations**: Human attention is able to focus on correct regions of interest (P(described | fixated)= 0.465), but typically covers only part of them (Coefficient Correlation = 0.222, Similarity = 0.353, Spearman Correlation = 0.324).

Boosted Attention Module (BAM)


Top-down Attention Module
Input Image
CNN Visual Encoder (ResNet-101)
Language Generator
A
bride
and
groom
cutting
their
wedding
cake
Word by Word Generated Caption
W_v
W_{sal}
Stimulus-based Attention Module

$$I' = W_v I \circ log(W_{sal} I + \epsilon)$$

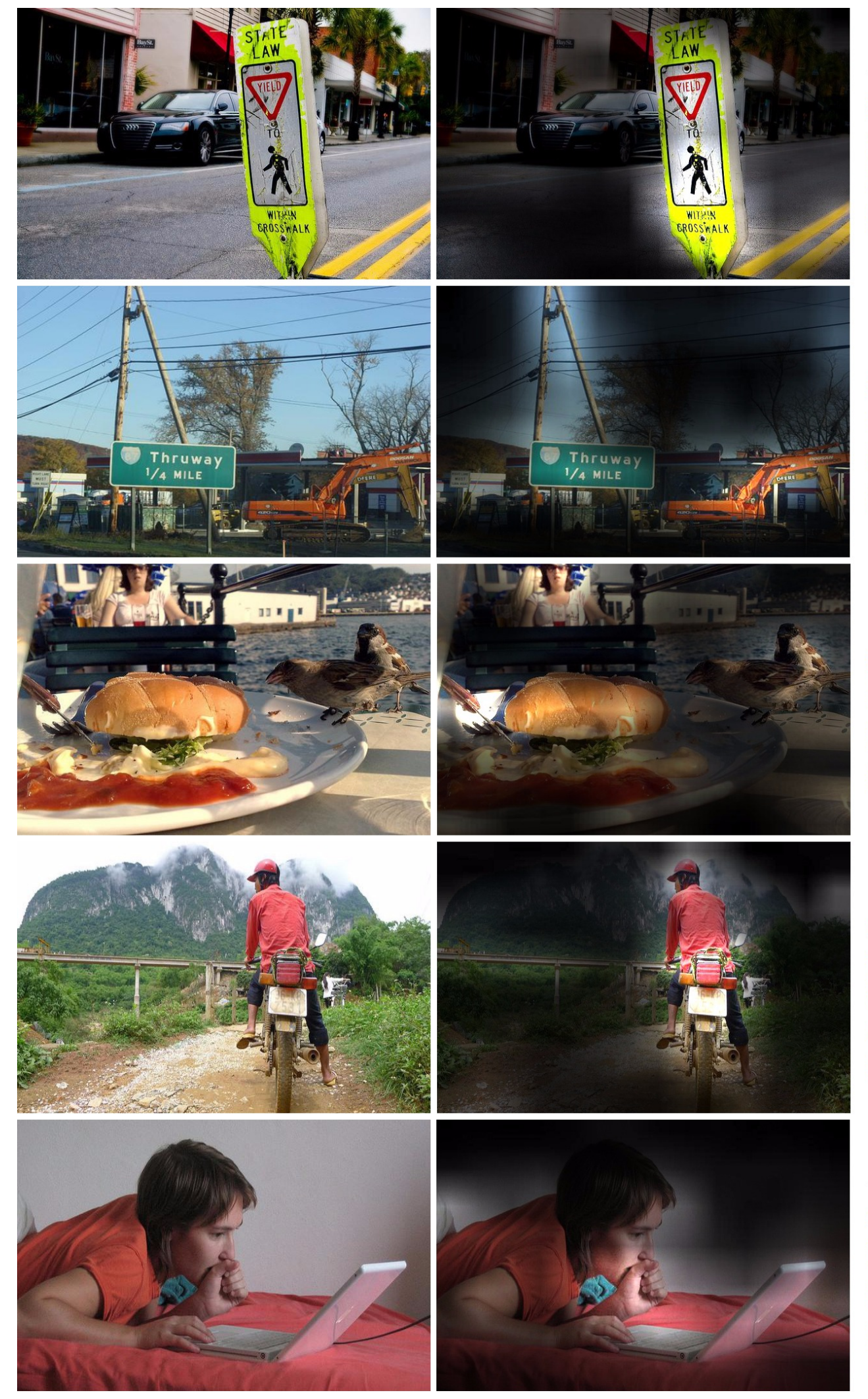
Figure 3. Overview of proposed Boosted Attention model. Top-down model attention maps are colored in purple, blue and green based on the associated word, while human stimulus-based attention is highlighted in red. The mechanism of our attention fusion method is shown in the gray box via equation.

Results

Quantitative Results

Model	Flickr30K				MSCOCO			
	B@4	MT	RG	CD	B@4	MT	RG	CD
Soft Attention [32]	0.191	0.185	-	-	0.243	0.239	-	-
ATT [34]	0.230	0.189	-	-	0.304	0.243	-	-
SCA-CNN [2]	0.223	0.195	0.449	0.447	0.311	0.250	0.531	0.952
SCN-LSTM [5]	0.265	0.218	-	-	0.330	0.257	-	1.012
RLE [25]	-	-	-	-	0.304	0.251	0.525	0.937
AdaATT [21]	0.251	0.204	0.467	0.531	0.332	0.266	0.549	1.085
Att2all [26]	-	-	-	-	0.342	0.267	0.557	1.140
ours-Baseline	0.267	0.197	0.471	0.523	0.335	0.258	0.551	1.062
Baseline-BAM	0.274	0.208	0.482	0.586	0.354	0.265	0.562	1.122
Improvement (%)	2.6%	5.6%	2.3%	12.0%	5.7%	2.7%	2.0%	5.6%
Att2in [26]	-	-	-	-	0.333	0.263	0.553	1.114
Att2in-BAM	-	-	-	-	0.360	0.269	0.565	1.142
Improvement (%)	-	-	-	-	8.1%	2.3%	2.2%	2.5%

Qualitative Results


a parking meter on a street with palm trees.
a street sign on the side of the road.
1. A close up of a crosswalk sign in the middle of the road.
2. A tatter street sign sits in the crosswalk.
3. The yield to pedestrians sign is all scratched up.

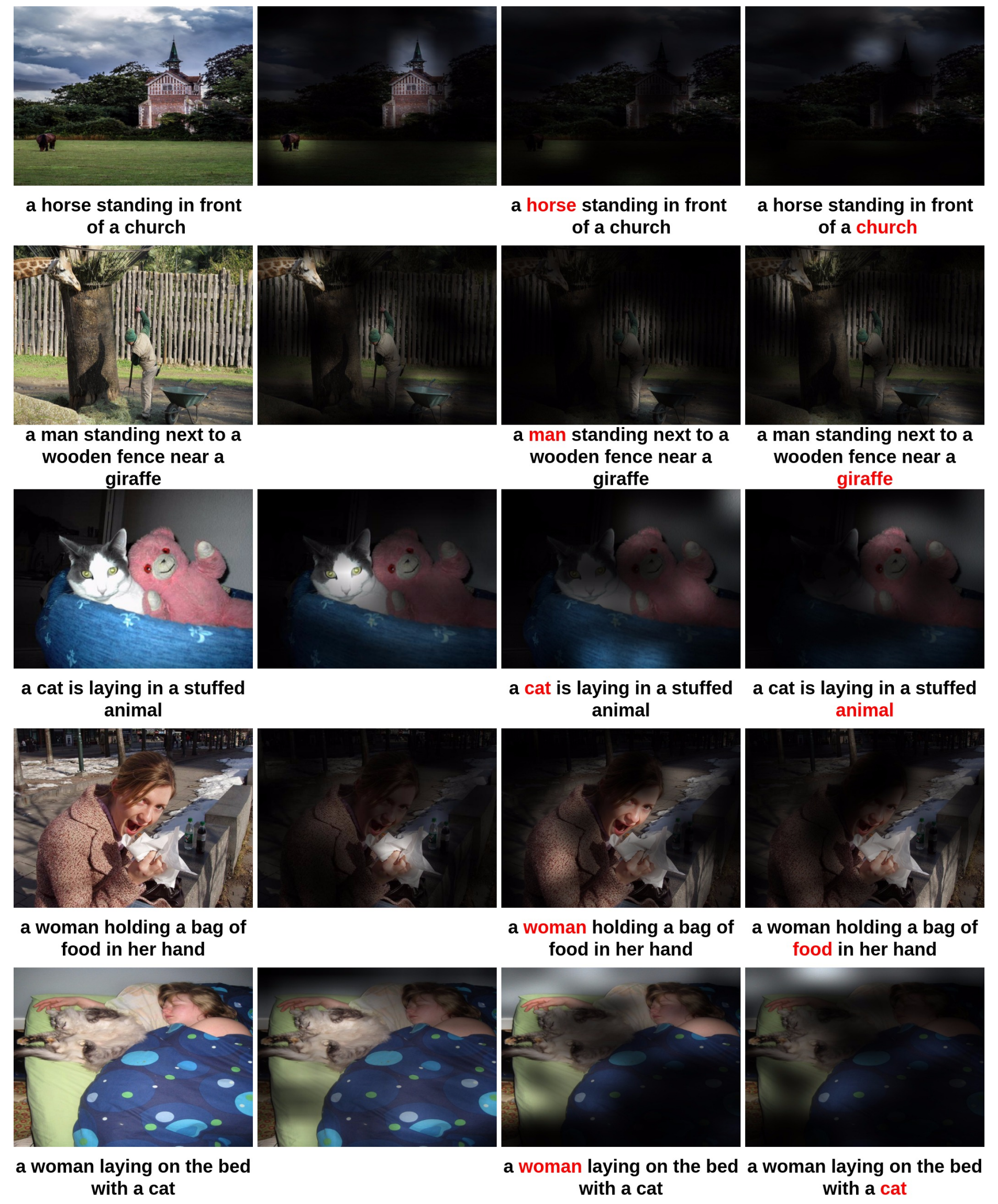
a green and yellow bus parked next to a street sign.
a sign is on the side of a road
1. A green sign says Thruway one fourth mile.
2. A road sign stands next to the road.
3. a street sign below a bunch of power lines

a plate of food that is sitting on a table.
a bird sitting on top of a plate of food.
1. A plate topped with bread, greens and pasta and a bird.
2. there are two birds standing on the plate of food.
3. A bird attempting to bite a piece of sandwich bread.

a man is standing next to a motorcycle.
a man riding a motorcycle with a mountain in the background.
1. A man in a red shirt and a red hat is on a motorcycle on a hill side.
2. A man riding on the back of a motorcycle.
3. Man riding a motor bike on a dirt road on the countryside.

a woman sitting on a bed with a red shirt.
a woman sitting on a bed with a laptop.
1. there is a woman laying in a bed using a lap top.
2. A girl on a bed studying something on her laptop.
3. a woman using a white laptop on the bed.

Figure 4. From left to right: Input images, stimulus-based attention maps computed by proposed method, captions generated by model without stimulus-based attention (black), captions generated by model with our Boosted Attention Module (red) and ground truth human-generated captions (blue).


a horse standing in front of a church | a horse standing in front of a church | a horse standing in front of a church

a man standing next to a wooden fence near a giraffe | a man standing next to a wooden fence near a giraffe | a man standing next to a wooden fence near a giraffe

a cat is laying in a stuffed animal | a cat is laying in a stuffed animal | a cat is laying in a stuffed animal

a woman holding a bag of food in her hand | a woman holding a bag of food in her hand | a woman holding a bag of food in her hand

a woman laying on the bed with a cat | a woman laying on the bed with a cat | a woman laying on the bed with a cat

Figure 5. Corporation between attention. From left to right: input images, stimulus-based attention maps, two sets of model attention maps. The generated words associated with specific model attention maps are highlighted in red.

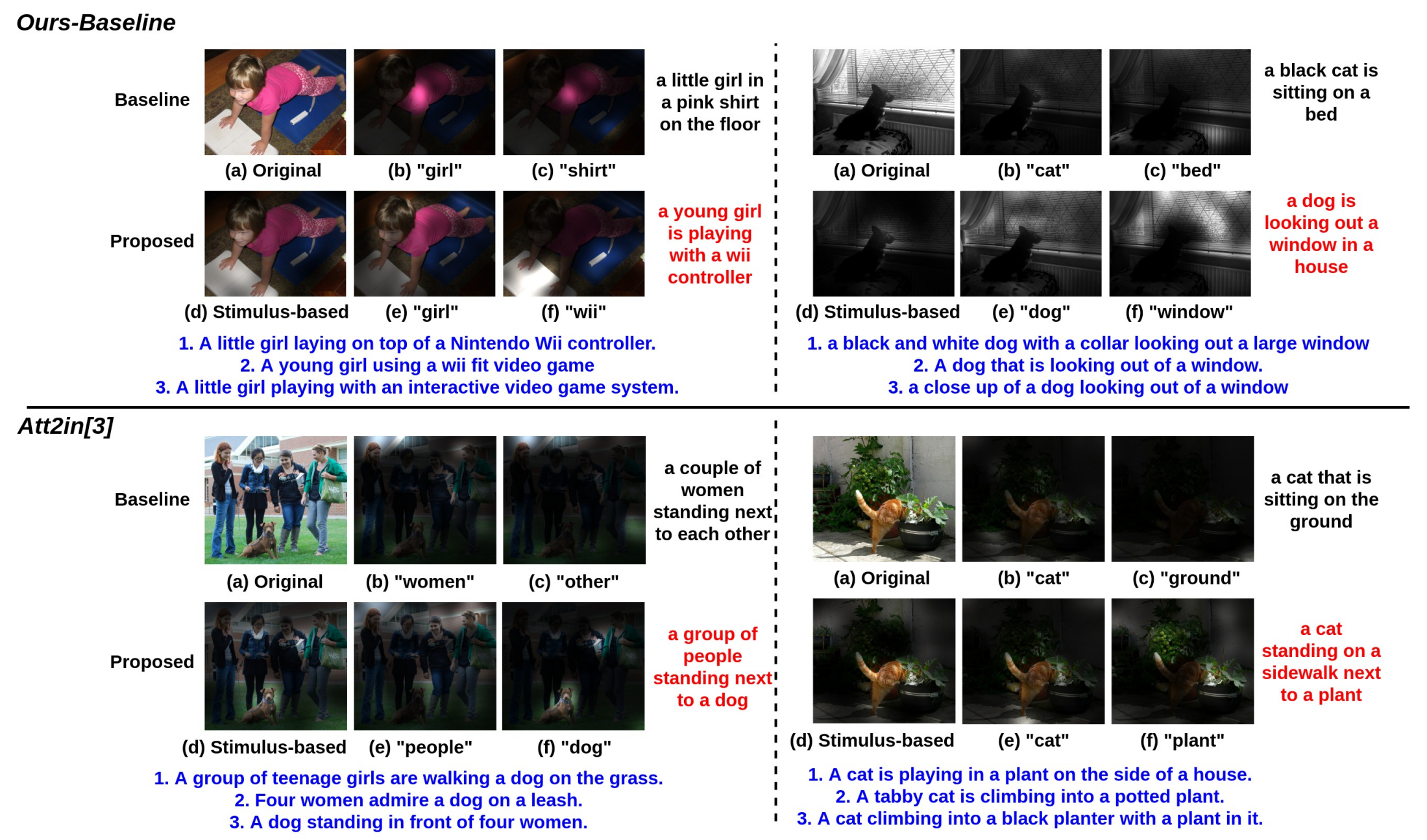
Scenarios for Attention Corporation

Scenario I: Human attention successfully captures all regions of interest. In this case, model attention plays a minor role on discriminating salient regions, having no clear focus (1st row, Figure 5) or attending to same regions as human attention (2nd row).

Scenario II: Human attention partially covers regions of interest (3rd and 4th rows). Under this situation, model attention will focus on the missing regions and complement to human attention.

Scenario III: Human Attention fails to distinguish salient object (5th row), model attention plays a major role on concentrating the regions of interest.

How Human Attention Affects Model Attention?



1. A little girl laying on top of a Nintendo Wii controller.
2. A young girl using a wii fit video game
3. A little girl playing with an interactive video game system.

1. a black and white dog with a collar looking out a large window
2. A dog that is looking out of a window.
3. a close up of a dog looking out of a window

1. A group of teenage girls are walking a dog on the grass.
2. Four women admire a dog on a leash.
3. A dog standing in front of four women.

1. A cat is playing in a plant on the side of a house.
2. A tabby cat is climbing into a potted plant.
3. A cat climbing into a black planter with a plant in it.

Figure 6. Comparison of top-down model attention maps between models without (b-c) and with (e-f) stimulus-based attention (d). The corresponding captions generated by models are colored in black and red, while ground truth captions are shown in blue color.