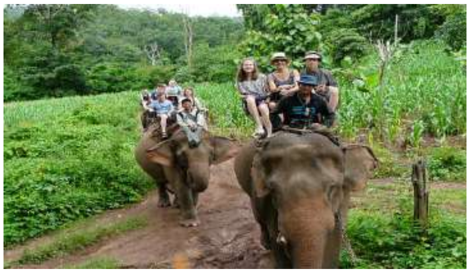
Supplemental Material: Customized Image Narrative Generation via Interactive Visual Question Generation and Answering

Andrew Shin[1] Yoshitaka Ushiku[1] Tatsuya Harada[1,2]
[1]The University of Tokyo, [2]RIKEN
{andrew,ushiku,harada}@mi.t.u-tokyo.ac.jp

Table 1: Examples of captions and questions for the same image. While captions essentially describe the same contents, questions widely vary in terms of the topics.

Image

COCO Captions
• A group of people sitting on the back • Several people are taking a ride on elephants • Some people are riding elephants in the jungle • The people are riding on the two elephants • People riding on elephants in the jungle
VQA Questions
• Is this standard transportation in the United States? • Are they on a paved roadway? • How many people are riding elephants?

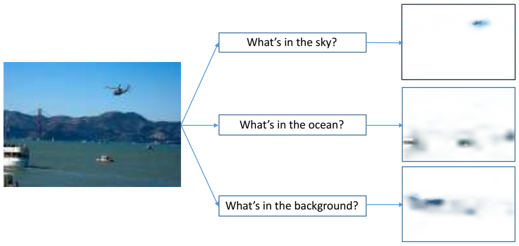


Figure 1: Viewer's attention varies depending on the context provided.

0.1. Why generate quesetions?

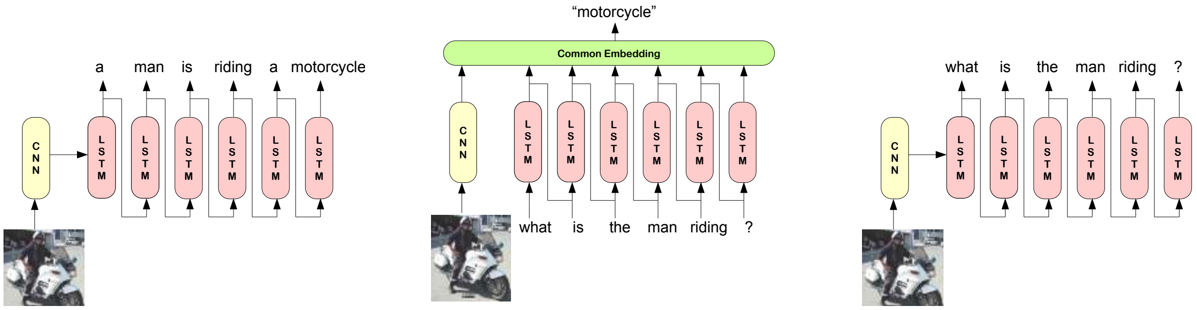
A question may arise as to why not to simply ask the users to select the region or part of the image that stands out the most to them. In such case, there would be no need to *generate* the questions for each image, as the question '*what stands out the most?*' would suffice for all images. This, however, would be equivalent to a simple saliency annotation task, and would not allow for any meaningful customization or optimization per user. Thus, as discussed above, generating a question for each image is intended to provide a context in which each user can apply their own specific interest. Figure 1 shows how providing context via questions can diversify people's attention. Apart from simply generating diverse image narratives based on the user input, many potential applications can be conceived of. For example, in cases where thorough description of an entire scene results in a redundant amount of information both quality and quantity-wise, application of our model can be applied to describe just the aspect that meets the user's interest that was learned.

Table 2: Statistics from the crowd-sourcing task on collecting answers to non-visual questions.

# of answers collected	48,090
# of unique answers	15,469
# of workers participated	187
max. # assignments by worker	1609
avg. # assignments per worker	51.43
rewards per assignment	$.10
10 most common answers	'*yes*','*no*','*tom*', '*london*','*mine*', '*downtown*','*john*','*me*' '*halloween*','*new york*'

0.2. Clarification of DIANE

Few works tackled the task of narrative evaluation, hardly taking visual information into consideration. Al-



(a) Image captioning. (b) Visual question answering. (c) Visual question generation.

Figure 2: Illustration of the overall workflow for each task.

Table 3: Examples of answers collected on VQG.

Question	Answer
‘*What is the name of the man?*’	‘*Tom*’
‘*What is the score in the game?*’	‘*0-0*’
‘*What kind of record is being played?*’	‘*rap records*’
‘*How long until the bathroom is fixed?*’	‘*1 week*’
‘*Why is he making weird face?*’	‘*he’s drunk*’
‘*What’s the cat’s name?*’	‘*Moni*’
‘*How much did that cost?*’	‘*10 dollars*’
‘*What destroyed this town?*’	‘*bomb*’
‘*Why are the trees lit up?*’	‘*It’s Christmas time*’
‘*What are the ingredients?*’	‘*fish,bread,broccoli*’

Table 4: Examples of questions generated using non-visual questions in VQG dataset.

Image	Generated Questions
	• Is this a hotel room? • Is that a picture of your house? • Where did you get the pillows? • Is this new tile? • Was that clean there? • How big is that room?
	• Is the woman drunk? • Is this a church? • Is this structure in a museum? • What city was this in? • Are they protesting?
	• What kind of pizza is that? • Is it for dinner? • What kind of topping is this on the pizza? • What does the plate say?
	• What is this bird staring at? • How long will it be there? • Is that a real bird? • What sort of bird is that? • What kind of flower is that?

Table 5: Examples of human-written image narratives collected on Amazon Mechanical Turk.

Image	Human-written Narrative
	The food truck looks good. I bet they have good food. Does everyone in a food truck have a beard? I am so done with the beard thing. Hope his beard does not get into the food.
	Car is in very good shape for the age. This is a prefect car for California. I think i see that this is in Huntington beach. This would attract a lot of attention. Great way to pick up girls or guys.
	A dad and his daughter are sitting on the couch. They have just woken up. They each have a cup of juice. They use cups with lids so they don’t spill on the couch.
	Tom is playing with a frisbee. He is practicing new moves. He jumped up in the air. He is trying to catch it between his legs. He was successful in his attempt.

though we could not find an authoritative work on the topic of narrative evaluation, this was our best attempt at not only reflecting precision/recall, but various aspects contributing to the integrity of the image narrative. *Diversity* deals with the coverage of diction and contents in the narrative, roughly corresponding to *recall*. *Interestingness* measures the extent to which the contents of the narrative grasp the user’s attention. *Accuracy* measures the degree to which the

Table 6: Statistics for human-written image narratives collected on Amazon Mechanical Turk.

# of answers collected	13,221
rewards per assignment	$.20
minimum length of image narrative	10
maximum length of image narrative	83
average length of image narrative	31.629

Table 7: Examples of image narratives generated by training with human-written image narratives.

Image	**Generated Narrative**
	a man is sitting on a chair he is wearing a white shirt he seems to
	a man is holding a hot dog he is wearing a white shirt he seems to

Table 8: Examples of generated questions for user interaction using our proposed model and VQG only respectively.

Image	**Generated Questions**
	Ours
	What is the dog doing?
	VQG
	What is the color of the couch?
	Ours
	What is the color of the car?
	VQG
	What is the weather like?

description is relevant to the image, corresponding to *precision*. Contents that are not visually verifiable are considered accurate only if they are compatible with salient parts of the image. *Naturalness* refers to the narrative's overall resemblance to human-written text or human-spoken dialogue. *Expressivity* deals with the range of syntax and tones in the narrative.

0.3. Additional Experiments

We also performed an experiment in which we generate image narratives by following conventional image captioning procedure with human-written image narratives collected on Amazon Mechanical Turk. In other words, we trained LSTM with CNN features of images and human-written image narratives as ground truth captions. If such setting turns out to be successful, our model would not have much comparative merit.

We trained an LSTM with collected image-narratives for training split of MS COCO. We retained the experimental conditions identically as previous experiments, and trained for 50 epochs. Table 7 shows example narratives generated. Not only does it utterly fail to learn the structure of image narratives, but it hardly generates text over one sentence, and even so, its descriptive accuracy is very poor. Since LSTM now has to adjust its memory cells' dependency on much longer text, it struggles to even form a complete sentence, not to mention inaccurate description. This tells us that simply training with human-written image narratives does not result in reliable outcomes.

With reference human-written image narratives, we further performed CIDEr [3] evaluation as shown in Table 13.

0.4. Discussion

It was shown via the experiments above that there exists a certain consistency over the choices made by the same user, and that it is thus beneficial to train with the choices made by the same users. Yet, we also need to investigate whether such consistency exists across different categories of images. We ran Fast-RCNN [2] on the images used in our experiment, and assigned the classes with probability over 0.7 as the labels for each image. We then define any two images to be in the same category if any of the assigned labels overlaps. Of 3,000 pairs of images used in the experiment, 952 pairs had images with at least one label overlapping. Our proposed model had average human evaluation score of 4.35 for pairs with overlapping labels and 2.98 for pairs without overlapping labels. Baseline model with image features only had 2.57 for pairs with overlapping labels and 2.10 for pairs without overlapping labels. Thus, it is shown that a large portion of the superior performance of our model comes from the user's consistency for the images of the same category, which is an intuitively correct conclusion.

However, our model also has superiority over baseline model for pairs without overlapping labels. This may seem more difficult to explain intuitively, as it is hard to see any explicit correlation between, for example, a car and an apple, other than saying that it is somebody's preference. We manually examined a set of such examples, and frequently found a pattern in which the color of the objects of choices was identical; for example, a red car and an apple. It is difficult to attribute it to a specific cause, but it is likely that there exists some degree of consistency in user choices over different categories, although to a lesser extent than for images in the same category. Also, it is once again confirmed that it is better to train with actual user choices made on specific questions, rather than simply with most conspicu-

Table 9: Conversion rules for transforming question and answer pairs to declarative sentences.

Type	Rule (Q→A)	Question	Ans.	Converted Ans.
yes/no	VB1+NP+VB2/JJ?	–	–	–
	→NP+*conjug*	-	-	
	(VB2/JJ,*tense*(VB1))	*Did he get hurt?*	*yes*	*He got hurt.*
	or, NP	-		
	+*negate*(*conjug*	-	-	
	(VB2/JJ,*tense*(VB1)))	*Is she happy?*	*no*	*She is not happy.*
	MD+ NP+VB?	–	–	–
	→NP+MD+VB **or**,	*Will the boy fall asleep?*	*yes*	*The boy will fall asleep.*
	NP+*negate*(MD)+VB	*May he cross the road?*	*no*	*He may not cross the road.*
number	"*How many*"+NP+	-	-	
	/is/are+EX?	-	-	
	→EX+*is/are*+*ans*+NP	*How many pens are there?*	*2*	*There are 2 pens.*
	"*How many*"+NP1(+MD)	-	-	
	+VB(+NP2)?	–	–	–
	→*ans*(+MD)+VB(+NP2)	*How many people are walking?*	*3*	*3 people are walking.*
	"*How many*"+NP1+	-	-	
	VB1/MD+NP2+VB2?	–	–	–
	→NP2	-	-	
	+(MD+VB2)/*conjug*	-	-	
	(VB2,*tense*(VB1))	-	-	
	+*ans*+NP1	*How many pens does he have?*	*4*	*He has 4 pens.*
others	WP/WRB/WDT+	-	-	
	"*is/are*"+NP?	-	-	
	→ NP+"*is/are*"+*ans.*	*Who are they?*	*students*	*They are students.*
	WP+NP+VP? → *ans.*+VP	*What food is on the table?*	*apple*	*Apple is on the table.*
	WDT+NP+VP(+NP2)?	-	-	
	→*ans.*(+NP)+VP(+NP2)	*Which hand is holding it?*	*left*	*Left hand is holding it.*
	WP/WDT+MD+VB?	-		
	→*ans.*+MD+VB	*Who would like this?*	*dog*	*Dog would like this.*
	WP/WDT+MD+NP+VB?	-	-	
	→NP+MD+VB+*ans.*	*What would the man eat?*	*apple*	*The man would eat apple.*
	WP/WDT+VP(+NP)?	-	-	
	→*ans.*+VP(+NP)	*Who threw the ball?*	*pitcher*	*Pitcher threw the ball.*
	WP/WDT+VB1+NP+VB2?	–	–	–
	→NP+*conjug*	-	-	
	(VB2,*tense*(VB1))+*ans.*	*What is the man eating?*	*apple*	*The man is eating apple.*

ous objects.

0.5. Additional Figures & Tables

Table 1 shows the contrast between semantic diversity of captions and questions. Figure 2 shows overall architecture each of image captioning, visual question answering, and visual question generation task. Table 2 shows statistics for crowd-sourcing task on collecting answers to non-visual questions in VQG dataset. Table 3 shows examples of answers to VQG questions collected on crowd-sourcing. Table 4 shows examples of generated questions using VQG dataset. Table 5 shows examples of human-written image narratives. Table 6 shows statistics for human-written image narratives collection. Table 9 shows conversion rules for natural language processing stage for narrative generation process as used in Section 3. Table 10 to Table 12 show more examples of image narratives. Table 8 shows examples of questions for user interaction that were generated using our proposed model of combining VQG and VQA, and the baseline of using VQG only. Table 14 shows another example of customized image narratives generated depending on the choices made by user upon the question.

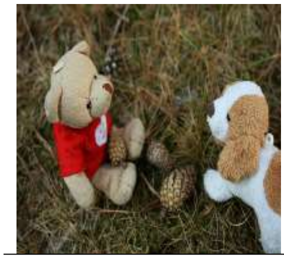
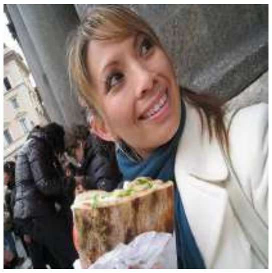
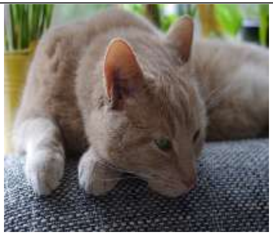
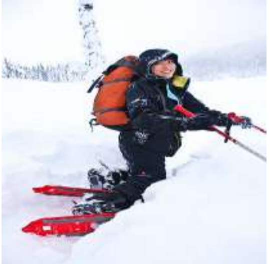
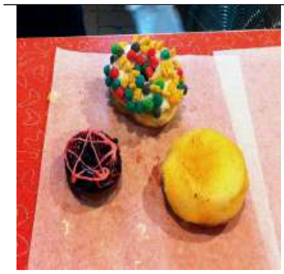
Image	COCO	SIND	DenseCap	Ours
	A teddy bear sitting on a wooden bench. A teddy bear sitting on top of a tree. A teddy bear is sitting on the ground. A train traveling down tracks next to a forest. A teddy bear is sitting on a tree branch.	The kids had a great time. The dog was very happy to see me. I had a great time. The view was amazing. We had a great time.	A teddy bear sitting on a wooden bench. A teddy bear in a red hat. Red teddy bear. A white teddy bear. The nose of a sheep. The teddy bear is s itting on the ground.	These are stuffed animals. That teddy bear can be scary when you see it at night. The animals are there for fun. The bear is sleeping. This is not a real bear.
	A man is eating a hot dog in a restaurant. A man holding a hot dog in his hand. A man holding a hot dog in a bun. A man in a suit and tie standing in front of a building. A man in a hat is holding a hot dog.	We had a great time.	A man is eating a hot dog in a restaurant. Woman holding a sandwich. Woman has brown hair. Woman in black jacket. A sandwich on a white plate. A brown wooden wall.	The girl is eating sandwich. Her name is mary. She is hungry. She eats a lot. She is smiling.
	A cat laying on top of a bed next to a remote control. A cat laying on top of a bed next to a laptop. A cat laying on top of a bed next to a window. A person is holding a piece of broccoli. A large building with a clock on it.	The dog was very happy to see me. I had a great time. We saw a lot of old buildings. The view from the top was amazing.	A cat laying on top of a bed next to a remote control. A cat laying on a bed. The head of a cat. Ear of a cat. The cat is brown. The ear of a cat.	There is 1 cat. This cat looks lonely. The cat is not sleeping. The weather is sunny.
	A man riding skis down a snow covered slope. A man in a red jacket is snowboarding. A man wearing a hat and a tie. A street light with a building in the background. A group of people standing on a beach with a kite.	We had a great time. I was so happy to see me. We went to the city to see the sights. I was so excited to see my friends. I went to the city last weekend.	A man riding skis down a snow covered slope. Man in red jacket. Snow covered mountain. The woman is wearing a helmet. The man is wearing black pants. Black and white jacket.	The person is skiing. This is alps. This person is having fun. He won the competition. The person is holding ski poles.
	A plate with a piece of cake on it. A close up of a pair of scissors on a table. A plate with a sandwich and a salad on it. A piece of cake on a plate with a fork. A close up of a plate of food on a table.	The food was delicious. The flowers were so beautiful. I had a great time.	A plate with a piece of cake on it. A yellow piece of donut. A red basket on the table. A box of donuts. A table with a wooden table. A donut with sprinkles.	There are 3 different types of food. This is not a healthy breakfast. This cake looks so fun. The orange color is bread.

Table 10: More examples of image narratives.

Table 15 shows examples of how the choices made by user upon the question were reflected in new images.

0.6. Additional Clarifications

Why were yes/no questions excluded? Yes/no questions are less likely to induce multiple answers. The number of possible choices is limited to 2 in most cases, and rarely

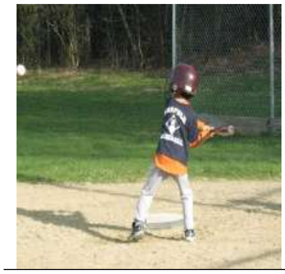
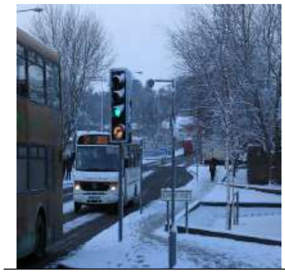
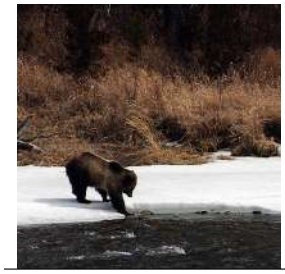
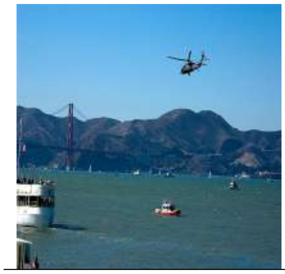
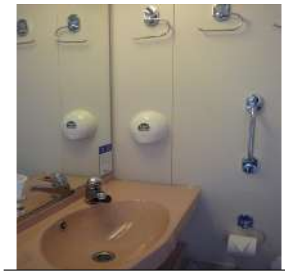
Image	COCO	SIND	DenseCap	Ours
	A baseball player swinging a bat at a ball. A man is standing in the grass with a kite. A bench sitting on top of a lush green hillside. A small boat in a body of water. A park bench with a tree in the background.	The man is giving a speech. We went to the park to see the sights. The view from the top was amazing. The building was very tall. We saw a lot of old buildings	A baseball player swinging a bat. Dirt on the ground. The jersey is blue. Player holding a bat. The helmet is black. Green grass on the field.	The boy is playing baseball. The score of the match is tied at zero. The name of that who is playing is john. The weather is like sunny. The bat is black. Trees are in the background.
	A city street filled with lots of traffic. A red double decker bus driving down a street. A bus is parked on the side of the road. A bus that is driving down the street. A bathroom with a toilet and a sink.	We had a great time. The car was covered in snow. The inside of the building was very tall. The view from the top was amazing.	A traffic light. Trees with no leaves. A tree with no leaves. The front wheel of a bus. A street light. A traffic light on a pole.	That bus looks loud together. The traffic light is green. The man is walking. The color of the car is white. The weather is like cold. This is not a busy road.
	A black bear walking across a lush green field. A polar bear walking in the snow near a tree. A black bear is walking in the water. A black and white photo of a bird on a tree branch.	The dog was so excited to see the water. The dog was very tired. The dog was very happy to see me. We had a great time. We saw a lot of old buildings. We went to the lake to see the sights.	Bear walking on the ground. A bear in the water. A large tree. A bear in the snow. Trees covered with snow. The bear is brown.	The bear is walking. This bear does not look silly. The bear is not eating. The bear is in the wild. It is cold outside. The weather is like cold.
	A group of people flying kites on a beach. A large boat floating on top of a body of water. A bathroom with a sink and a mirror. A person on a snowboard in the snow. A plane flying in the sky over mountains.	The view from the top was amazing. We had a great time. The view was amazing. I went to the museum today. The beach was beautiful.	Plane in the sky. A blue sky with no clouds. Photo taken during the day. Red and white boat in water. The water is blue. The sky is blue.	The water is calm. This is daytona beach. The weather is like cloudy. The plane is flying. The water is blue. It is day.
	A bathroom with a toilet sink and mirror. A bathroom with a toilet and a sink. A pair of scissors sitting on top of a wooden table. A street sign on a pole in front of a building. A pair of scissors sitting on top of a table.	I had a great time. I went to the park to see the sights. I went to the museum today.	A white bathroom sink. A silver metal towel rack. Silver metal faucet. A silver faucet. Toilet paper holder on wall. A white toilet paper.	This bathroom is in a hotel. The bathroom is clean. This room is very good. The color of the wall is white. There is a reflection in the mirror. The light is on.

Table 11: More examples of image narratives.

correspond well to particular regions.

Failure cases for rule-based conversion: Since both questions and answers are human-written, our conversion rule frequently fails with typos, abridgments, words with multiple POS tags, and grammatically incorrect questions. We either manually modified them or left them as they are.

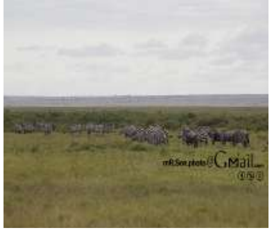
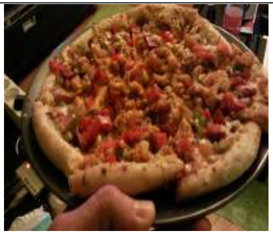
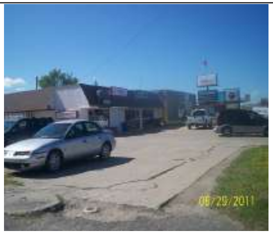
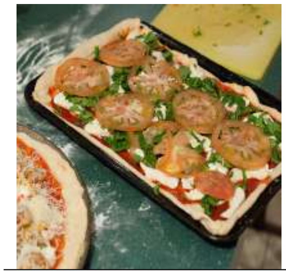
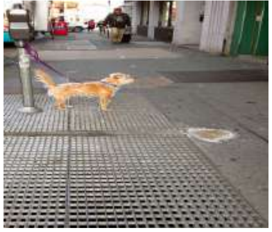
Image	COCO	SIND	DenseCap	Ours
	A herd of zebras grazing in a field. A herd of zebra standing on top of a lush green field. A bird flying over a building with a clock. A man standing on a sidewalk next to a street sign. A group of zebras are standing in a field.	We had a great time. I went to the museum today. We saw a lot of interesting things. We went to the city to see the sights. We saw many different types of animals. We went to the museum.	A herd of zebras grazing in a field. A field of grass. Two zebras in a field. The photo was taken in the daytime. White clouds in blue sky. The grass is tall.	The zebras like each other. These animals are related. The zebras are not in a zoo. The animal is grazing.
	A close up of a pizza on a plate. A close up of a sandwich on a plate. A cat sitting on top of a window sill. A bathroom with a toilet and a sink. A plate of food with a sandwich and french fries. A person holding a hot dog in a bun.	The food was delicious. We had a great time. I went to the museum today.	A close up of a pizza on a plate. Pizza on a plate. Pizza on a table. The hand of a person. A cup of coffee. The pizza has red sauce.	500 calories are in the meal. This is a pizza. This is not a healthy meal. This is not for vegetarian.
	A street with cars parked on the side of it. A car parked in front of a parking meter. A street sign on a pole on a street. A car parked on the side of a road. A street sign that is on a pole.	We went to the city to see the sights. The car was covered in snow. I went to the museum today. We went to the museum. We had a great time. We went to the location.	A street with cars parked on the side of it. A silver car parked on the street. A black car parked on the street. A white truck. Blue sky with no clouds. A black truck.	The car is gray. The car is parked illegally. Where the car is is inappropriate. That is pine tree behind.
	A plate of food with a fork and knife. A pizza with a lot of toppings on it. A plate with a sandwich and a salad. A close up of a plate of food with broccoli.	The food was delicious.	A plate of food with a fork and knife. Pizza on a table. A pizza on a plate. A slice of pizza. The pizza has red sauce. A slice of tomato.	This is a vegetarian pizza. This is not a cheese pizza. The green vegetable is spinach. This is a healthy meal.
	A dog that is sitting on a bench. A street sign on a pole in front of a building. A man riding a skateboard down a street. A dog is running with a frisbee in its mouth. A large building with a clock on it.	The dog was very happy to see me. We had a great time. The house was very nice. I went to the museum today.	A dog that is sitting on a bench. A brown dog. A brick sidewalk. Man walking on sidewalk. Dog walking on sidewalk. A white line on the ground.	There is a dog. The dog is sad. The color of the wall is white. The color of the fire hydrant is gray.

Table 12: More examples of image narratives.

Model	COCO	SIND	DenseCap	Ours
CIDEr	18.0	9.9	**28.0**	**27.7**

Table 13: Each model's performance on CIDEr with human-written image narratives as ground truths.

Experiments with different VQA models. Most of well-known VQA models' performances are currently in a relatively tight range. In fact, we tried [1], SOTA at the time of experiment, but did not see any noticeable improvement.

Is attention network retrained to handle sentences? No, but we found that attention network trained for questions works surprisingly well for sentences, which makes sense since key words that provide attention-wise clue are likely limited, and hardly inquisitive words.

Why not train with "I dont know?" We were concerned that answers like "I don't know" would likely overfit.

Table 14: Examples of image narratives generated depending on the user choices.

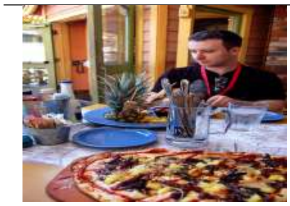
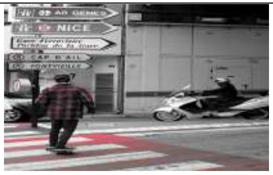
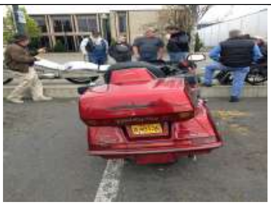
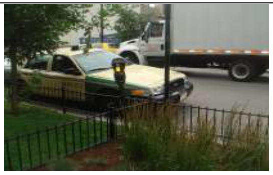
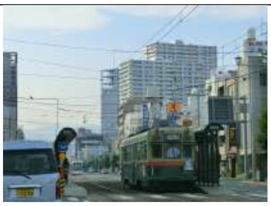
Image	**Answers, Regions and Narratives**		
	Pizza	Pine Apple	Plate
Generated Question	Pizza is on the table. The man is eating pizza. The pizza is thin crust.	Pine apple is on the table. The man is vegetarian.	Plate is on the table. The man is eating. The man is eating more than one person would.
What is on the table?			

Table 15: Examples of image narratives generated on new images, depending on the choices made.

Image & Question	**Choice**	**New Image**	**Image Narrative**
 What is the man riding?	skateboard		No one is riding bicycle. The man is standing.
	motorcycle		The motorcycle is red. No one is riding motorcycle.
	car		This is not a modern building. The image is not in black and white.
 What color is the car?	white		The white object is bus. The car is white.
	green		The bus is green. The train is headed to Washington.
	yellow		The train is yellow. The image is not in black and white.

It would also undermine creative aspect of image narrative, without adding much to functional aspect.

References

[1] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Emnlp. 2016. 7

[2] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 3

[3] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 3